STATGRAPHICS® Plus 5

A Manugistics Product

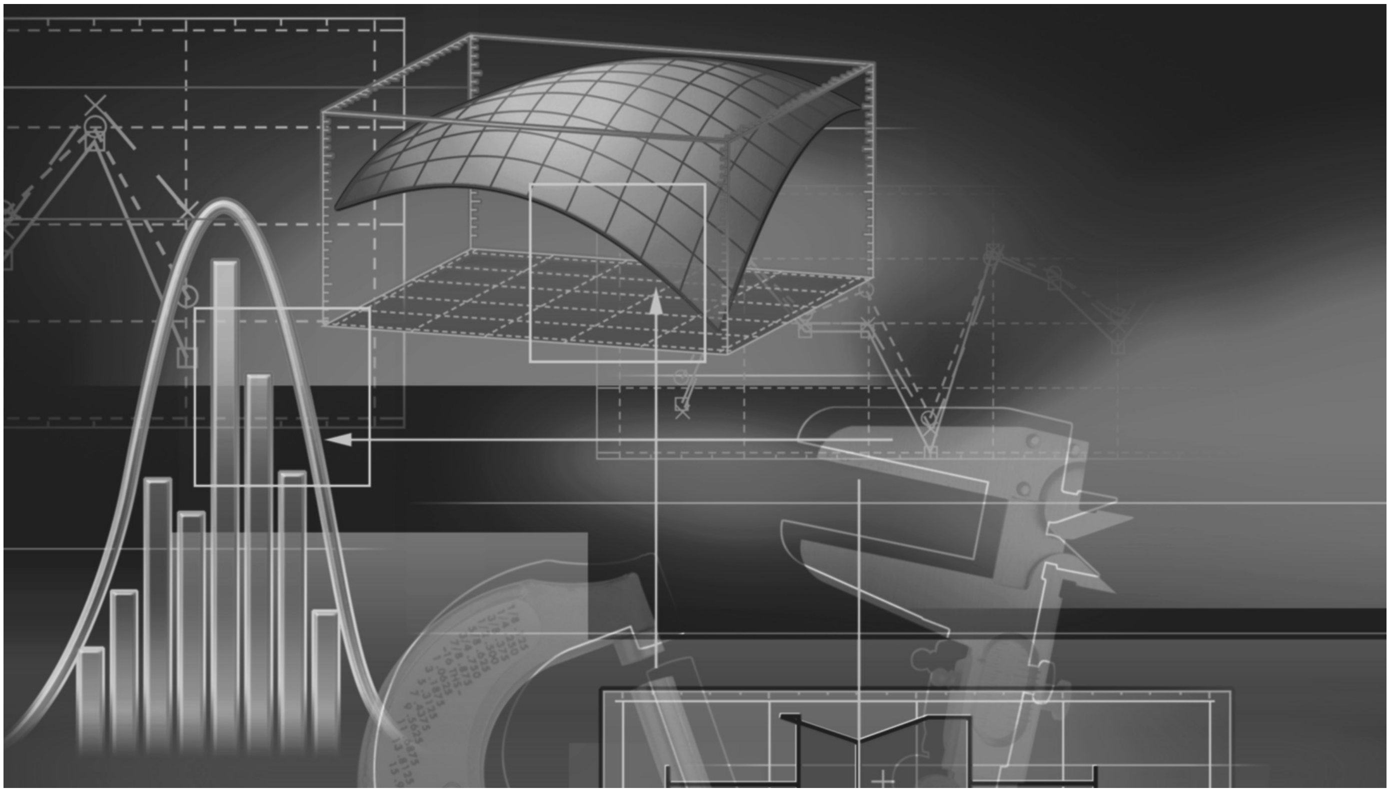
www.statgraphics.com

www.statgraphics.com

www.statgraphics.com

www.statgraphics.com

www.statgraphics.com



This document was prepared to help users of software products from Manugistics, Inc. You cannot use the contents of this document for any other purpose without written permission. We supply the material in this document without representation or warranty of any kind. Manugistics, Inc. therefore assumes no responsibility and shall have no liability of any kind arising from the supply or use of this document or the material in it.

10 9 8 7 6 5 4 3 2 1

Throughout this document we may use trademarked names. We do not use a trademark symbol in every occurrence of the trademarked name. Rather, we only use the trademarked names in an editorial fashion and to the benefit of the trademark owner. Infringement of the trademark is not intended.

Table of Contents

Before You Begin

Welcome to STATGRAPHICS *Plus*, powerful Windows products that include all the tools you need to master statistics. Four different products are offered:

- **STATGRAPHICS *Plus* Enterprise Edition**
  Includes all features of STATGRAPHICS *Plus* Professional in a multi-user network configuration.

- **STATGRAPHICS *Plus* Professional**
  Includes STATGRAPHICS *Plus* Standard Edition; STATGRAPHICS *Plus* Quality and Design; as well as analyses for time series, multivariate methods, and advanced regression.

- **STATGRAPHICS *Plus* Quality and Design**
  Includes STATGRAPHICS *Plus* Standard Edition, and analyses for quality control, and design of experiments.

- **STATGRAPHICS *Plus* Standard Edition**
  Includes standard analyses for plotting, describing, comparing, and relating data.

The four product configurations provide the tools you need for statistical analysis, interactive graphics, and presentation-quality reports and graphs.

Eight unique features add to the ease-of-use quality of STATGRAPHICS *Plus*: SnapStats, StatAdvisor, StatFolios, StatGallery, StatLink, StatPublish, StatReporter, and StatWizard.

- **SnapStats**
  Analyses designed to provide one-page summaries for commonly encountered data analysis problems.

- **StatAdvisor**
  Provides an easy-to-understand explanation of the analysis you are performing.

- **StatFolios**
  A macro-like feature that captures an entire set of analyses and saves them for future use.

- **StatGallery**
  Archives up to 100 graphics panes and an unlimited number of text panes, which you can edit and save alone or with a StatFolio.

- **StatLink**
  Provides the capability to tie StatFolios directly to data sources from spreadsheets, databases, and measuring devices such as digital micrometers via linking software.

- **StatPublish**
  Provides the capability for publishing reports in a HTML format.

- **StatReporter**
  Provides a capability for publishing reports directly from STATGRAPHICS *Plus* by saving and printing reports that contain text and graphics from the StatGallery or from analyses you create.

- **StatWizard**
  Provides a guide for new or casual users that helps determine which analysis is appropriate for use with certain types of data.

Registering For and Receiving Product Support

To receive support services for any STATGRAPHICS *Plus* product, complete and return the Registration Card attached to the Warranty and License Agreement.

If you have a technical question about any STATGRAPHICS *Plus* product, first look in the printed manuals, or look in Online Help. If you cannot find the answer, contact the STATGRAPHICS *Plus* Technical Support HelpLine.

Users outside the United States should contact the dealer from whom they purchased their software.

Using the Technical Support HelpLine

Support services are available to help you gain the most benefit from your STATGRAPHICS *Plus* product. Every user receives Basic Support at no cost for ninety (90) days starting at the date of purchase.

You must be using the current version of STATGRAPHICS *Plus*. The STATGRAPHICS *Plus* "Software Support Plan" brochure included with your software package outlines the available services.

Manugistics also offers HelpLine *Plus*, an annual support plan that includes all the benefits of Basic Support but provides even more support services, such as priority telephone service and free software patches. This fee-for-service plan ensures fast technical assistance. The Software Support Plan brochure also outlines the services available when you sign up for HelpLine *Plus*. This service is available only in the United States.

In addition to explaining how to get support, the "How to Get Help" card included with your package explains how you order replacement items, lists the information we need when you contact us, and provides instructions for using our 24-hour Information Line.

Requesting Technical Assistance

To request technical assistance, you can:

- Call **301-984-5489**
- Send a fax to **301-255-8406**
- Send an e-mail to **answer@manu.com**
- Send a letter to
  Manugistics, Inc.,
  Attn.: STATGRAPHICS *Plus* HelpLine,
  2115 East Jefferson Street,
  Rockville, MD 20852

You can call our Technical Support HelpLine from 10:00 a.m. to 5:30 p.m., Eastern time, Monday through Friday (excluding holidays). From you, we will need the following information:

- serial number of your copy of the software
- your full name
- your company's name

- your telephone and/or fax number
- your e-mail address, if you have one
- name of the software product
- version number of the software
- complete description of your problem or question.

Requesting Replacement Items

If you need replacement items, please call us at 301-984-5489. If you believe that the software media you received is defective, check the media on another computer, or try to read the media using another method. If you continue to get error messages, call the Technical Support HelpLine.

Using the 24-Hour Information Line

You can also call our 24-hour Information Line at 301-984-5463. This recorded message provides answers to frequently asked technical questions. You must have a touch-tone telephone to use this menu-driven service.

Using the Web Site

The STATGRAPHICS *Plus* web site is available to enhance your productivity. The goal of the web site is to provide you with additional information and a variety of related sources that will allow you to maximize your use of the product.

To reach our site, use: **http://www.statgraphics.com**

Information on the site includes information about our training courses, technical specifications that list the recommendations for running the current version of the software, software patches, sample StatFolios, tutorials from our manuals and from other web sites, as well as statistics-related resources.

Registering for Product Training

Improving productivity is only one benefit you receive when you sign up for product training. You also learn to use the product and manipulate data, run statistical analyses, generate and customize graphs, and save and document results.

Most users who complete training courses have a high degree of appreciation for the power of STATGRAPHICS *Plus*. Many believe that the courses provide skills they can apply to their job on a frequent basis.

Public and on-site courses are available. The STATGRAPHICS *Plus* Training schedule is available on our website..

To Register for a Course or Receive Training Information

If you need more information about a course, or you want to register for one, you can:

- Visit our website for the most recent training information.
- Send us a fax at **301-255-8406**
- Call us at **800-592-0050**

Reviewing New Features in Version 5

New Analyses and Enhancements

Version 5 of STATGRAPHICS *Plus* contains a wealth of new analyses and enhancements. They are listed under the menus in which they appear.

Plot Menu

- **Matrix Plot (Scatterplots)**
  New analysis which creates a matrix of two-variable plots for a set of numeric variables, with box-and-whisker plots on the diagonal.

Describe Menu

- **One Variable Analysis (Numeric Data)**
  Revised *Confidence Intervals* tabular option allows calculation of one-sided bounds or two-sided intervals. Also calculates bootstrap intervals for the mean, median, and standard deviation.

- **Multiple-Variable Analysis (Numeric Data)**
  Now calculates Kenfdall’s tau, as well as Spearman’s ρ.

- **Outlier Identification (Numeric Data)**
  New analysis identifies outliers in a sample of numeric values using Studentized values, modified Z-scores, Grubb’s test, and Dixon’s test. Also computes resistant estimates of the location and scale parameters.

Compare Menu

- **Multiple Sample Comparison (Multiple Samples)**
  Levene’s test has been added to the *Variance Check* pane, Friedman’s test has been added for balanced data, and group ranges may now be saved.

- **Comparison of Rates (Multiple Samples)**
  Old “Comparison of Counts” analysis has been renamed and a new likelihood ratio test has been added to compare two or more Poisson rates.

- **One-Way ANOVA (Analysis of Variance)**
  Levene’s test has been added to the *Variance Check* pane, and group ranges may now be saved.

- **Multifactor ANOVA (Analysis of Variance)**
  A column with standard errors for the least squares means has been added to the *Multiple Range Tests* output.

- **Factor Means Plot**
  New analysis creates a matrix plot for data from designed experiment. It consists of main effects plots on the diagonal and interaction plots off the diagonal.

Relate Menu

- **Simple Regression**
  Adjusted R-Squared, MAE, and Durbin-Watson statistic have been added to the *Analysis Summary*.

- **Polynomial Regression**
  Approximate P-value for the Durbin-Watson statistic has been added to the *Analysis Summary*.

- **Box-Cox Transformations**
  A new Skewness and Kurtosis plot has been added to the list of graphical options. Approximate P-value for the Durbin-Watson statistic has been added to the *Analysis Summary*

- **Multiple Regression**
  Approximate P-value for the Durbin-Watson statistic has been added to the *Analysis Summary*.

SnapStats! Menu

- **One Sample Analysis**
  New analysis produces a one-page summary of a single sample of numeric data.

- **Two Sample Comparison**
  New analysis produces a one-page summary to compare two samples of numeric data.

- **Paired Sample Comparison**
  New analysis produces a one-page summary to compare two paired samples of numeric data.

- **Multiple Sample Comparison**
  New analysis produces a one-page summary to compare several samples of numeric data.

- **Curve-Fitting**
  New analysis produces a one-page summary of a simple regression relating Y and X.

- **Capability Assessment (Individuals)**
  New analysis produces a one-page summary of a sample of individual measurements, including control charts and capability indices.

- **Capability Assessment (Grouped Data)**
  New analysis produces a one-page summary of grouped measurements, including control charts and capability indices.

- **Gage R&R**
  New analysis produces a one-age summary of a gage repeatability and reproducibility study.

- **Automatic Forecasting**
  New analysis produces a one-page summary of automatic model fitting for a sample of time series data.

Experimental Design Menu

- **Analyze Design**
  Approximate P-value for the Durbin-Watson statistic has been added to the *Analysis Summary*.

Time Series Analysis Menu

- **Automatic Forecasting**
  New analysis automatically selects a forecasting model for a time series. Uses the general class of ARMA (p,p-1) models.

Advanced Regression Menu

- **General Linear Models**
  Now fits mixed models with both fixed and random factors. Automatically computes expected mean squares, variance components, and F-tests with proper denominators.

- **Calibration Models**
  Adjusted R-Squared, MAE, and Durbin-Watson statistic have been added to the *Analysis Summary*.

- **Logistic Regression**
  New features include estimation of percentiles, standard error bounds on fitted model, survival function plot, probability plot, and calculation of Pearson or deviance residuals.

- **Probit Analysis**
  New analysis fits probit regression models to dependent variables consisting of proportions.

- **Poisson Regression**
  New analysis fits Poisson regression models to dependent variables consisting of counts or rates.

- **Life Data Regression**
  New analysis fits regression models to dependent variables which are failure times. Data may be censored or uncensored. Linear and loglinear models may be fit with several error distributions.

New Features

- **StatPublish**
  Publishes a StatFolio to a web site for access by web browsers. Output from all analyses, the comments window, the StatGallery, and the Stat Reporter is saved in HTML format on a local disk or Internet server. Tables and graphs are saved as JPEG files, PNG files, or interactive Java applets. Sites may be configured to update automatically at periodic intervals.

- **DataSheet**
  Up to 1000 columns may now be added to a datasheet; cut and paste operations include column names; double-clicking on a column name goes directly to the Modify Column dialog box; and files may be sorted in random order, if desired.

- **XML**
  Now reads and writes data files in XML format. An ODBC driver is included with the program to access data in XML files.

- **3D Graphics Effects**
  Optional 3D effects have been added to many plots.

- **Graphics Image Files**
  Graphs may now be saved as Windows metafiles, or in JPEG, PNG, Windows BMP, or encapsulated PostScript format.

- **Help Menu**
  Automatic links have been added to the STATGRAPHICS web site.

- **User Monitoring**
  System now monitors the number of concurrent users to ensure that license is not violated.

- **Automatic Saving of Results**
  A new "Autosave" checkbox has been added to all Save Results dialog boxes so that results are automatically saved whenever an analysis is rerun.

- **Start-Up Scripts**
  Whenever a StatFolio is loaded, an optional start-up script is run. Start-up scripts may perform analyses, assign values to data variables, execute Windows commands, print results, publish results to a web site, and automatically end program execution. This allows batch-type processing.

- **Variable Names**
  The set of allowable characters in variable names has been extended to permit the use of many symbols and non-English characters.

- **International Fonts**
  All output now fully supports the use of non-English character fonts.

- **New Exclude Function**
  Simplifies the process of excluding individual observations from analysis.

Using the Documentation Set

The documentation set supports STATGRAPHICS *Plus* running in a 32-bit Windows 95, Windows 98, or Windows NT environment, and contains several types of documentation: a printed and bound manual, an electronic online manual, and online help. This section discusses each of these documentation methods and explains the conventions used in the documentation.

Using the Printed and Bound Manual

The *User Manual* provides information and instructions for using STATGRAPHICS *Plus*, such as how to learn the product; how to access and use files; how to work with data and DataSheets; how to work with graphs and graphics options; how to print; and how to use special unique features, such as the StatAdvisor, StatLink, or the StatReporter. The manual is intended for users of all experience levels, but especially for new or novice users.

Using the Online Manual

Beginning with Version 4, a new form of manual joined our documentation set — *Online Manuals* — electronic online publications. Initially, only the *Quality and Design* manual were available. In Version 5 all manuals are available online and can be installed with the product to your hard drive. Access to the manuals is through manuals.pdf which is launched as an option under the STATGRAPHICS help menu. Online manuals are viewed using Adobe Acrobat Reader+Search. A copy has been placed on the CD for your convenience. This product must be installed to use the online manuals.

Using the Online Help Program

If you run into a problem or if you simply need more information about a topic, in addition to other information, you can use the Online Help program that provides specific context-sensitive information about the dialog boxes, what you enter into them, and the defaults. In addition, information available from the Help Contents window provides concepts, reference and technical support information, and step-by-step directions for accomplishing tasks such as using Graphics options.

Note: The specific information about the descriptions of the items in a dialog box and the defaults now reside ***only*** in the Online Help program instead of appearing in both the printed manuals and Online Help.

You can access Online Help using a variety of methods.

- **Use F1**
  Pressing F1 displays *context-sensitive* help that is related to the area of the window you are in when you press F1. For example, if you press F1 when

you are on a Box-and-Whisker Plot, a help topic about the Box-and-Whisker Plot displays.

- **Use the Question Mark on the Application Toolbar**
  Clicking something in a pane, then pressing the question mark on the Application toolbar, also accesses *context-sensitive* help. The Help window that appears is related to the pane you clicked.

- **Use the Help Menu Button**
  Clicking the Help button on the Application Menu bar provides help topics that you access by selecting topics from the Contents window.

- **Use the Help Dialog Box Button**
  Clicking the Help button on any dialog box displays a Help window for the dialog box you are currently using. Continue clicking on topics until you reach the level of help you are seeking.

- **Use the Online Help Index**
  The Help window that displays after you enter Online Help, provides an index that lets you search for a topic in an alphabetical list. The Index contains every entry in the entire Help program. When you click the Index button, a tabbed window with two tabs displays: Index and Find. Use the Index tab to search the alphabetical list; use the Find tab to search for specific words and phrases.

Hotspot topics occur throughout Online Help. Hotspots are hypertext links that allow you to jump from one topic to another by clicking the hotspot, which is shown in color and underscored. There are two types of hotspots in the Help program for STATGRAPHICS *Plus*: Jump hotspots and Pop-up hotspots.

- Jump hotspots display the link's destination topic and allows you to jump from one topic to another, without regard to physical or hierarchical placement.

- Pop-up hotspots display the destination topic in a temporary pop-up window that overlaps the main window. This type of hotspot lets you read the definition of a topic without detracting too much from the primary topic you are reading.

Conventions

This section describes general as well as symbolic and stylistic conventions. The manuals use special symbols and stylistic conventions; however, in some cases, they are used only in the printed manual, which are noted.

General Conventions

- Like other Windows applications, there are many methods you can use to accomplish the same task. Throughout the step-by-step instructions in the manuals, an attempt was made to consistently use only one method.
- STATGRAPHICS *Plus* uses the following extensions to save StatFolios, data files, and StatGalleries: *.sgp, *sf3, and *.sgg, respectively. It is important to note that .sfx and .sgp files are not backwards compatible. Data files you create and save in Version 5 of STATGRAPHICS *Plus* are backwards compatible.

Symbolic Conventions

n — Indicates the first-order item in an outline or series; used in printed manual.

Stylistic Conventions

Italic — Indicates information you type; for example:

> Type *a:setup* or *b:setup*, as appropriate, and click OK.

SMALL CAPS — Indicates the names of menu items; used in printed manual. For example:

> Choose EDIT... CHANGE TEXT FONT....

Bold Keystrokes — Indicates the name of a variable and/or a file. A variable name appears in bold, lower-case type; a file name appears in bold, lower-case type, with initial capital letters. For example, the variable **mpg**; the file **Cardata**.

Initial Caps — Indicates the name of an object, such as a dialog box or a dialog-box component; for example, the One-Variable Analysis dialog box or the Not Equal option.

Click OK — Indicates you can either click OK using the mouse or by pressing the Enter key to carry out an action.

Bulleted Lists — Provides information, not procedural steps.

Numbered Lists — Indicates a process of sequential steps.

Return and Enter Keys — Pressing one of these keys performs the same action; pressing either key processes an action.

A Plus Sign between Names — Indicates you must press the keys sequentially. For example, **Alt+H** means that you must press the Alt key and hold it down while you press the H key.

Finding Information

Knowing what to expect when you open a manual is a time-saver. The next two sections provide information about the organization of the manuals generally, and the organization of the chapters in this manual, specifically.

How Manuals Are Organized

The STATGRAPHICS *Plus* manuals are not designed to be read from cover to cover; instead, they should provide quick access to the information you need by providing basic instructions for using the products.

How the Chapters are Organized

Most of the chapters are instructional; that is, they explain how you perform specific tasks or statistical analyses.

- **Chapter 1, Getting Started**
  Verifies the requirements for running STATGRAPHICS *Plus*, then provides instructions for installing a software protection device (international users), and installing the program on a single computer or on separate workstations on a network.

- **Chapter 2, Learning the Product**
  Helps you navigate STATGRAPHICS *Plus* by explaining basics such as the structure of the menus, and how to use commands, toolbars, taskbars, and different types of windows. It also explains how to change the

appearance of dialog boxes, set system-wide preferences, and use and access dialog boxes and options. The last section describes how you perform basic tasks that involve analyses.

- **Chapter 3, Accessing and Using Files**
  Explains the tasks of opening, saving, and closing various types of files, how you combine data files or StatFolios, and how you sort a file. Provides instructions for importing files from other applications, for using ODBC to query databases, and for linking and exchanging files with other applications.

- **Chapter 4, Working with Data and DataSheets**
  Includes information about working with variables; creating a DataSheet, which includes opening a blank form, setting up the form, entering the data, and knowing different ways of entering data. Sections on modifying data include changing a text font, changing the title of an analysis, and copying an analysis. The editing data section provides steps for undoing, cutting, and copying and pasting an entry; and inserting and deleting a column of data. A final section explains how to recode data.

- **Chapter 5, Working with Graphs and Graphics Options**
  Describes general graphics tasks such as accessing, opening, and saving graphs; explains how to perform basic graphics such as resizing and using the zoom feature; provides details for using the pages on the tabbed Graphics Options dialog box to perform specific tasks, including changing the grid, layout, lines, fills, and other attributes of a graph. The chapter then explains how you set system-wide graphics preferences, and set and save user preference profiles.

- **Chapter 6, Printing**
  Describes how you print the various types of windows in STATGRAPHICS *Plus* then how you print a StatGallery. The last sections describe how you use the various printing commands: Print Preview..., [Print] Setup..., and Page Setup....

- **Chapter 7, Using Special Features**
  Explains the basic concepts behind each of the unique features in STATGRAPHICS *Plus*, and provides instructions for using them.

- **Chapter 8, Using Basic Plots**
  Contains information about how you perform analyses using basic plots: Scatterplots, Exploratory Plots, Business Charts, Probability Distributions, Response Surfaces Plots, and Custom Charts.

- **Chapter 9, Describing Numeric Data**
  Contains information about the analyses and statistical methods you use to describe and summarize a single set of data: One-Variable Analysis, Multiple-Variable Analysis, Subset Analysis, Row-Wise Statistics, Power Transformations, and Statistical Tolerance Limits.

- **Chapter 10, Describing Categorical Data**
  Explains the analyzes you use with categorical data — data that contain numeric codes that represent discrete categories: Tabulation, Crosstabulation, and Contingency Tables.

- **Chapter 11, Working with Probability Distributions**
  Contains analyses for using probability distributions, creating a variety of probability plots, and using uncensored and censored data to perform distribution fitting.

- **Chapter 12, Performing Analyses Using Life Data**
  Consists of four analyses: life tables for intervals and times, and the Weibull and Arrhenius Plots analyses. Explains how to use uncensored and censored data to create life tables based on, respectively, counts of failures in intervals, and failure times; and how to use the Weibull Analysis to apply a distribution-modeling method to many product-failure mechanisms to create a wide variety of failure-rate curves. Also explains how to use the Arrhenius Plots Analysis to increase stress levels beyond normal operating conditions.

- **Chapter 13, Performing Hypothesis Tests and Determining Sample Size**
  Discusses the Hypothesis Tests and Sample Size Determination analyses you use with descriptive data. The chapter first explains how to perform hypothesis tests on a variety of parameters that involve a single sample. Then it explains how to select the proper sample size by calculating sampling sizes for four parameters.

- **Chapter 14, Comparing Two Data Samples**
  Presents analyses you use to calculate comparative-descriptive statistics for two or more samples — analyses you use to compare two samples —

Two-Sample Comparison, Paired-Sample Comparison, Hypothesis Tests, and Sample-Size Determination.

- **Chapter 15, Comparing Multiple Samples**
  Explains analyses you use to compare two or more means (Multiple-Sample Comparison); two or more proportions (Comparison of Proportions); and two or more Poisson counts (Comparison of Counts).

- **Chapter 16, Performing Analysis of Variance Tests**
  Focuses on Analysis of Variance tests and the analyses you use to perform the tests: One-Way ANOVA, Multifactor ANOVA, and Variance Components.

- **Chapter 17, Performing Regression Analysis**
  Explains the analyses available from the Relate menu: techniques for modeling the relationship between the dependent and independent variables: Simple Regression, Polynomial Regression, Box-Cox Transformations, and Multiple Regression.

- **Chapter 18, Using StapStats**
  Defines each of the nine one-page summaries for commonly-encountered data analysis problems. The summaries, called SnapStats, include a One Sample Analysis, Two Sample Comparison, Curve Fitting, and more.

This manual also includes various appendices and an index. The appendices are: Appendix A, Recognizing Icons and Buttons; Appendix B, Using Operators; Appendix C, Using Keyboard Equivalents; Appendix D, References; Appendix E, Calculations; and Appendix F, Glossary.

1 Getting Started

Verifying the Requirements

Important Note: This version of STATGRAPHICS *Plus* runs under Windows 95, Windows 98, Windows 2000 or Windows NT 4.0 and higher.

To run STATGRAPHICS *Plus*, you must have a Pentium, and:

- at least 32 megabytes of RAM (Random Access Memory)
- Microsoft's Windows 95, Windows 98, Windows 2000 or Windows NT, Version 4.0 or higher.
- a mouse
- a CD-ROM drive
- a hard-disk drive with at least 50 megabytes of available storage to hold the program files and the files for Online Help.

Installing the Software Protection Device

Many international users receive a copy-protected version of STATGRAPHICS *Plus*. If your STATGRAPHICS *Plus* package contains a software protection device, you must install it on your computer before you can run the program. The software protection device ensures that you use the software on only one machine at a time, in accordance with the Warranty and License Agreement.

Using the software protection device does not prevent you from using different machines or making as many backup copies of the software as you need.

To Install the Software Protection Device

1. Turn off your computer and all the peripheral devices — such as printers and monitors — that are attached to the printer.
2. Touch the case of your computer to discharge any static electricity.
3. Remove any peripheral devices attached to the parallel port you want to use.
4. Insert the pronged end of the software protection device into the port connector.

   If you have other software that also requires a software protection device, you can attach all these devices together in any order.

   Note: Every software protection device is different, so be sure to follow the instructions specific to that device.
5. Tighten the screws on the software protection device.
6. Reattach the cables for any peripheral devices you removed from the parallel port.

   When you start STATGRAPHICS *Plus*, the program checks to see if the software protection device is installed in the parallel port. If the device is installed, the session (the period of time between entering and leaving STATGRAPHICS *Plus*) will continue.

   In addition, the program checks for the software protection device at random intervals during your session. If the software protection device is not installed when STATGRAPHICS *Plus* checks for it, you will be prompted to attach it. When you install the software protection device, your session continues without interruption or loss of data.

   If STATGRAPHICS *Plus* prompts you to attach the software protection device even though it is already attached to your computer, try the following:

   - Make sure the software protection device is inserted securely into the parallel port. Try removing the device and reinserting it.
   - Be sure the software protection device is connected to the correct parallel port. If the device is installed in the wrong port, remove the device and attach it to the correct parallel printer port.

Installing the Software on a Single Computer or a Server

Note: STATGRAPHICS *Plus* is now distributed on a Compact Disc (CD).

When you install the program, the setup program:

- Prompts you for your name and the serial number of your STATGRAPHICS *Plus* software. The serial number appears on the jewel case for the CD.
- Creates the necessary subdirectories and copies the STATGRAPHICS *Plus* program files, sample datasets, Online Help file, and Online Manuals into those subdirectories.

To Install the software on a Single Computer or a Server

1. Insert the STATGRAPHICS *Plus* CD into the CD-ROM drive.
2. Start Windows, then select START... RUN... from the Taskbar.
3. Type ***n:setup***, where ***n*** is the letter of the CD-ROM drive, then click OK (or press Enter).

   As the installation progresses, the setup program prompts you to provide simple information: your name, your company name, the serial number of your STATGRAPHICS *Plus* program, and the path where you will install it. Read the instructions carefully and respond to the prompts.

   Note: If you are a network administrator, you should specify the installation path as the network path on the network server.

Installing the Software on a Network Workstation

To Install the Software on Each Workstation on a Network

Create the program item and program group (if desired). If you are not comfortable doing these tasks, refer to the Microsoft Windows *User's Guide* for detailed information.

When you create the program item, a dialog box prompts you to specify the command line and the working directory. In the Command Line text box, enter the path where you installed STATGRAPHICS *Plus* on the server, followed by the file name SGWIN.EXE.

For example, if the network server path is R:\SGNET\SGWIN, enter R:\SGNET\SGWIN\SGWIN.EXE. In the Working Directory text box, enter the path where you installed STATGRAPHICS *Plus* on the network server; do not include the file name SGWIN.EXE.

Note: Windows may display a warning that you specified a network path that may not be available during later Windows sessions; it will ask if you want to continue. Click Yes to continue.

Because there are a number of different configurations that can be used, in general, the network administrator will set up a link to the path where STATGRAPHICS *Plus* resides.

Note: Network versions of STATGRAPHICS *Plus* are licensed for use based on the number of simultaneous users. If you attempt to log on while the maximum number of users is already using the program, Version 5 will generate an error message and terminate.

Supplementary Files

A subdirectory called TESTDATA is included on the CD. It contains a large number of sample StatFolios and datasets. Although you will use your own data with the software during your work day, the samples are available if you want to try sample data before you use your own data.

A file called TESTDATA.DOC provides an index to the StatFolios. You can open this file using the Windows WordPad.

Starting and Exiting the Program

To Start the Program

Now that you have installed STATGRAPHICS *Plus*, you are ready to start it.

1. Start Windows.

2. Double-click the STATGRAPHICS *Plus* icon or run SGWIN.EXE.

   The program displays the Application window with various program buttons in the Taskbar. When you are ready to start working, you can enter new data or perform analyses using an existing file or StatFolio.

To Exit the Program

1. Choose FILE... EXIT STATGRAPHICS... from the Menu bar to end the current session and return to Windows.

2 Learning the Program

Learning the Basics

This chapter helps you navigate STATGRAPHICS *Plus* by explaining basics such as the structure of the menus, and how to use commands, toolbars, taskbars, and different types of windows. It also explains how to change the appearance of dialog boxes, set system-wide edit preferences, and use and access dialog boxes and options. The last section describes how you perform basic tasks that involve analyses.

Using Menus

Application programs in the Windows environment make extensive use of menus. The standards for the menu structure in STATGRAPHICS *Plus* are consistent with those of other Windows products; that is, the Menu bar provides access to most of the user commands. When you choose a selection on the Menu bar, a menu pops down.

STATGRAPHICS *Plus* provides a Menu bar across the top of the Application window, which lets you access files, perform file-management tasks, and access editing tools and statistical analyses. Figure 2-1 shows an Application window with some of its components labeled.

STATGRAPHICS *Plus* contains the following menus: File, Edit, Plot, Describe, Compare, Relate, Special, SnapStats!!, View, Window, and Help.

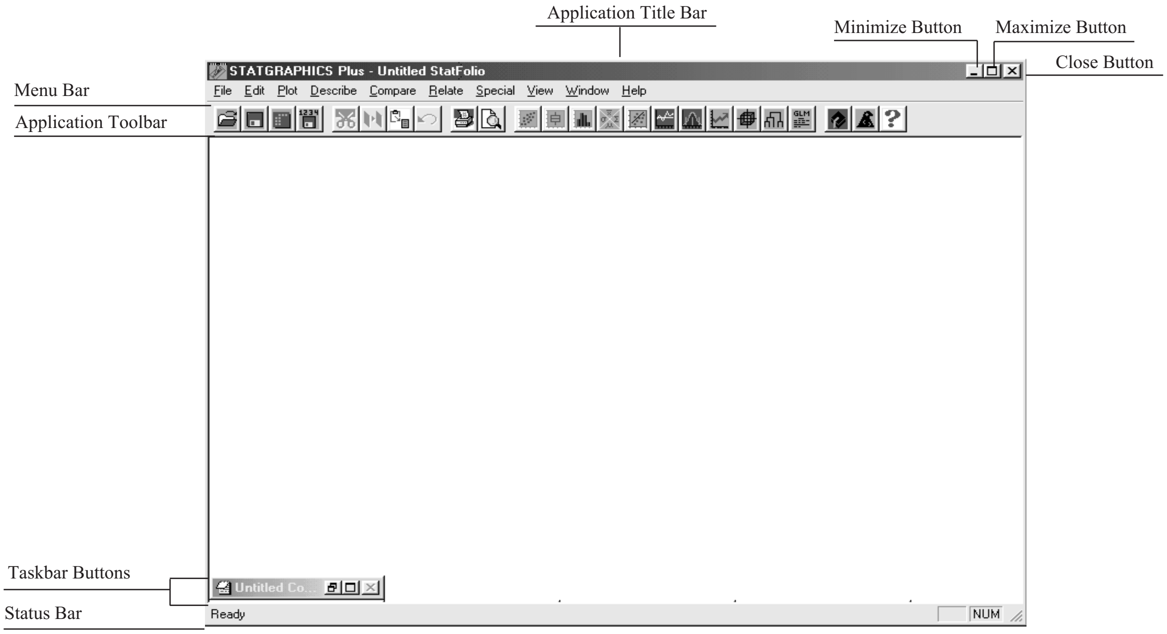


Figure 2-1. An Application Window with Some Components Labeled

Using the File Menu

The File Menu is shown in Figure 2-2. It contains commands to open and close files, print, open recently used files, and exit the program. Specifically, the File menu allows you to:

- Open, close, save and save StatFolios, a StatGallery, a StatReporter, and/or data files with a different name; query a database; and read a Clipboard directly from STATGRAPHICS *Plus*. *See Chapter 7, Using Special Features, for information about StatFolios, the StatGallery, and the StatReporter.*
- Use StatLink, a feature that can tie StatFolios directly to other data sources such as spreadsheets, databases, and measuring devices such as digital micrometers via linking software. StatLink queries the data source when a StatFolio is opened. Tasks you can perform include connecting to another data source, displaying status, starting and stopping polling, updating, and disconnecting a data source. *See Chapter 7, Using Special Features, for more information about these topics.*

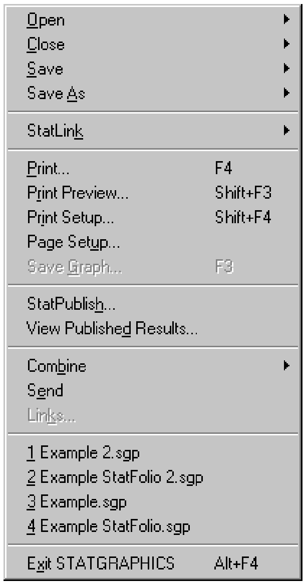


Figure 2-2. The File Menu

- Print files and preview a page before you print it; use Print Setup to select printers other than the default; and Page Setup to select values other than the defaults for page attributes; and use Save Graph to save a graph as a metafile. *See Chapter 6, Printing, Publishing and the StatFolio Start-Up Script, for more information about this topic.*
- Combine StatFolios and/or data files; use Send, an interface to e-mail programs that requires the use of Microsoft's Exchange program; and use Links to edit spreadsheet links. *For information about the Links command, see Chapter 4, Working with Data and DataSheets.*

The last section of the menu lists the files you most recently used and allows you to exit the program.

Using the Edit Menu

The Edit Menu, shown in Figure 2-3, contains commands that allow you to:

- Use editing commands such as Undo, Cut, Copy, Paste, and Paste Link.

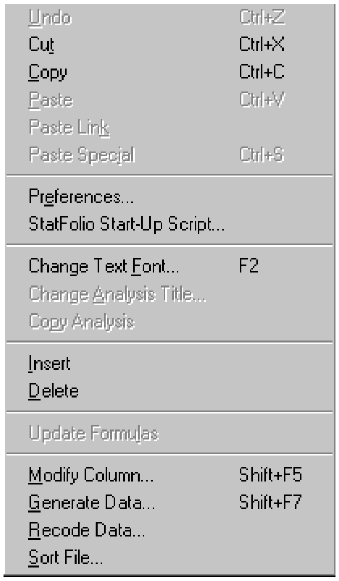


Figure 2-3. The Edit Menu

- Set system-wide edit preferences for system options, graphics, and link behavior.
- Change a text font and its attributes; change the title of an analysis; and create a duplicate of an analysis.
- Insert a new column into or delete an existing column from a DataSheet.
- Update a formula-type column in a DataSheet.
- Modify a column in a DataSheet; generate data values by using operators to transform the data; recode data into a STATGRAPHICS *Plus* format; and sort the data in a file.

Note: STATGRAPHICS *Plus* uses a “spreadsheet-like” worksheet — a DataSheet — within the DataEditor. The DataSheet is a document you use to store and manipulate data. *See Chapter 4, Working with Data and DataSheets*, for detailed information about how to use the DataEditor and a DataSheet.

Using the Plot Menu

The Plot Menu, shown in Figure 2-4, allows you to access basic analyses through the use of various common plots.

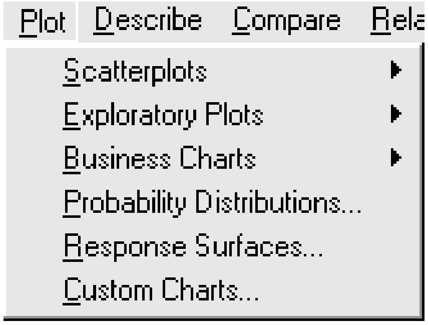


Figure 2-4. The Plot Menu

Scatterplots

Contains analyses you use to create basic types of scatterplots. The analyses are Univariate Plot, X-Y Plot, X-Y-Z Plot, Matrix Plot, Multiple X-Y Plot, Multiple X-Y-Z Plot, and Polar Coordinates Plot. *See Chapter 8, Using Basic Plots, for more information.*

Exploratory Plots

Contains analyses you use to create exploratory plots, which are useful for studying symmetry, checking distributional assumptions, and detecting outliers. The analyses are Box-and-Whisker Plot, Multiple Box-and-Whisker Plot, Normal Probability Plot, Frequency Histogram, Dot Diagram, Multiple Dot Diagram, Bubble Chart, and Radar/Spider Plot.

Business Charts

Contains analyses that produce business charts, which are useful for presenting relative quantities in an easily understood visual format. The analyses are: Barchart, Multiple Barchart, Piechart, Component Line Chart, and High-Low-Close Plot.

Probability Distributions

Contains 24 distributions that are useful for generating and saving random numbers, calculating probabilities, and plotting the probability and cumulative distributions.

Response Surfaces Plots

Contains an analysis that creates response plots and contour plots based on a user-defined function. These plots are helpful when you need to visualize the relationship between factors and a response variable.

Custom Charts

Contains an analysis for creating charts that are useful when circumstances require the use of a customized chart, usually in a manufacturing environment.

Using the Describe Menu

The Describe Menu, shown in Figure 2-5, allows you to access analyses used to investigate and summarize data.

Numeric Data

Contains analyses used to describe and summarize a set of data:
One-Variable Analysis, Multiple-Variable Analysis, Subset Analysis, Row-Wise Statistics, Power Transformations, Statistical Tolerance Limits and Outlier Identification.

Categorical Data

Contains analyses that work with data that contain numeric codes representing discrete categories: Tabulation, Crosstabulation, and Contingency Tables.

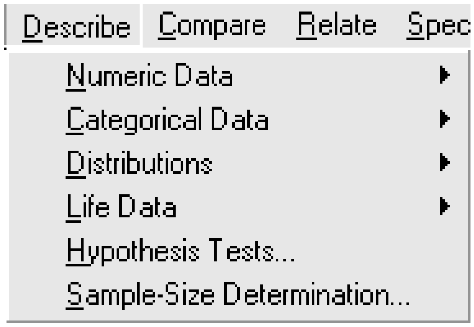


Figure 2-5. The Describe Menu

Distributions

Allows you to create probability distributions, produce probability plots to test distributional assumptions, or fit a specific distribution to a set of data. The analyses are: Probability Distributions, Probability Plots, and Distribution Fitting for Uncensored and Censored Data.

Life Data

Contains analyses that are useful for analyzing and summarizing lifetime data. The analyses include Life Tables (Intervals and Times), Weibull Analysis, and Arrhenius Plots.

Hypothesis Tests (Describe)

Performs hypothesis tests on a variety of parameters that involve a single sample.

Sample-Size Determination (Describe)

Determines the proper size for a sample.

Using the Compare Menu

The Compare Menu, shown in Figure 2-6, allows you to access analyses used to create and compare descriptive statistics.

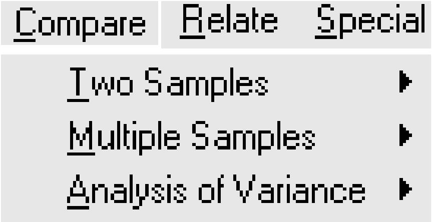


Figure 2-6. The Compare Menu

Two Samples

Performs analyses that compare two or more samples of data: Two-Sample Comparison, Paired-Sample Comparison, Hypothesis Tests, and Sample-Size Determination.

Multiple Samples

Performs analyses that compare several sets of data: Multiple-Sample Comparison, Comparison of Proportions, and Comparison of Rates.

Analysis of Variance

Performs one of three different analyses on data typically collected in a designed experiment. The purpose is to determine the effect of one or more experimental factors on the values for a response variable. The analyses are Factor Means Plot, One-Way ANOVA, Multifactor ANOVA, and Variance Components.

Using the Relate Menu

The Relate Menu, shown in Figure 2-7, allows you to access statistical analyses that model the relationship between dependent and independent variables.

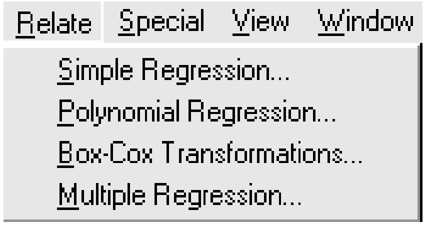


Figure 2-7. The Relate Menu

Simple Regression

Fits a model that relates one dependent variable to one independent variable by minimizing the sum of the squares of the residuals for the fitted line.

Polynomial Regression

Performs a special case of simple linear regression, whose models contain higher-order terms of the predictor variable.

Box-Cox Transformations

Determines an optimal transformation in the context of fitting a simple regression model. An approximate P-value for the Durbin-Watson statistic is now included in the Analysis Summary.

Multiple Regression

Analyzes the relationship between one dependent variable and one or more independent variables.

Using the Special Menu

The Special Menu, shown in Figure 2-8, allows you to access analyses in other STATGRAPHICS *Plus* products: STATGRAPHICS *Plus* Quality and Design and STATGRAPHICS *Plus* Professional.

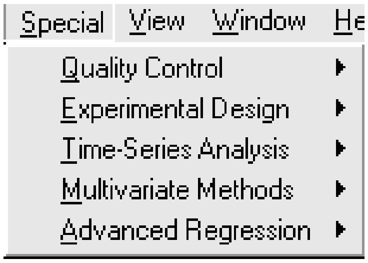


Figure 2-8. The Special Menu

Quality Control Analyses

Contains analyses that cover all aspects of quality control: determining statistical control of variable, attribute, and multivariate data; identifying defects; determining the capability of a process, including Gage R&R; using time-weighted and special-purpose control charts; optimizing processes, and acceptance sampling.

Experimental Design Analyses

Contains a built-in catalog of designs that include a wide range of choices, including screening, response surface, mixture, multilevel factorial, Taguchi, single and multifactor categorical, and variance component.

Time-Series Analyses

Contains a complete set of analyses for building models to handle data that change over time.

Multivariate Methods

Contains analyses that help to sort and group data, determine relationships between variables, and construct and test hypotheses.

Advanced Regression Analyses

Contains analyses that let you fully explore the data by formulating complex multiple regression models and validating methods, to choosing the best regression model.

Using the SnapStats!! Menu

The SnapStats!! Menu, shown in Figure 2-9, contains one-step analyses for many commonly encountered data analysis problems. The output is specially formatted to create one printed page containing both tabular and graphical output. These procedures are designed to produce quick, standardized output and, thus, have more limited options than the corresponding analyses found under other sections of the main menu.

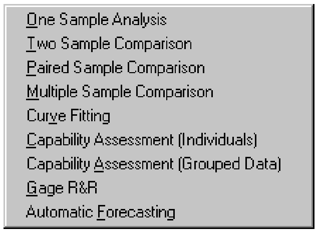


Figure 2-9. SnapStats!! Menu

One Sample Analysis

Produces a one-page summary of a single sample of numeric data.

Two Sample Comparison

Produces a one-page summary to compare two samples of numeric data.

Paired Sample Comparison

Produces a one-page summary to compare several samples of numeric data.

Multiple Sample Comparison

Produces a one-page summary to compare several samples of numeric data.

Curve Fitting

Produces a one-page summary of a simple regression relating Y and X.

Capability Assessment (Individuals)

Produces a one-page summary of a sample of individual measurements, including control charts and capability indices.

Capability Assessment (Grouped Data)

Produces a one-page summary of grouped measurements, including control charts and capability indices.

Gage R&R

Produces a one-page summary of a gage repeatability and reproducibility study.

Automatic Forecasting

Analyses that produce a one-page summary of automatic model fitting for a sample of time series data.

Using the View Menu

The View Menu, shown in Figure 2-10, controls the appearance of visible items in STATGRAPHICS *Plus*; that is, you use it to select and deselect the Toolbar, Status Bar, and StatAdvisor.

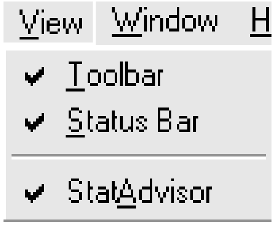


Figure 2-10. The View Menu

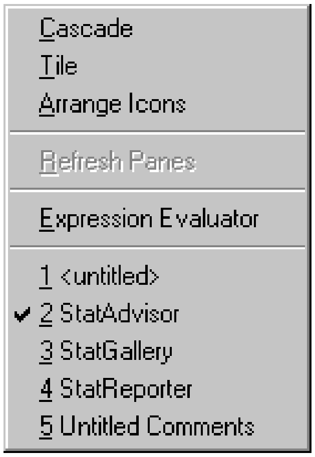


Figure 2-11. The Window Menu

Using the Window Menu

The Window Menu, shown in Figure 2-11, allows you to cascade, tile, and arrange icons; refresh panes; and move open items to the top viewing position.

Using the Help Menu

The Help Menu, shown in Figure 2-12, accesses the Contents window for the Online Help program; accesses Online manuals and notes; accesses the

STATGRAPHICS web site; accesses the StatWizard; and accesses the About window.

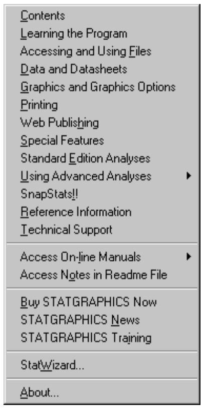


Figure 2-12. The Help Menu

Using Commands

As in all Windows applications, STATGRAPHICS *Plus* utilizes dialog boxes as the way to interact with the program. After you complete a dialog box, you use commands to advise the program of the action you want to perform; for example, you might click OK to process an action, or Cancel to void it.

Most dialog boxes in STATGRAPHICS *Plus* contain at least three command buttons: OK, Cancel, and Help, which are self-explanatory. However, while these commands are self-explanatory, some commands on the menus and submenus are not. In some instances, the same command might have the same name with a slightly varying meaning depending on the analysis you are using.

The Reference portion of Online Help contains an all-inclusive alphabetical list of the commands to provide you with a quick look-up when you have questions about a command's purpose.

Using Toolbars

Toolbars are a set of buttons on the Application window placed immediately below the Menu bar. The buttons correspond to a menu command. Clicking a button is the same as choosing a corresponding menu command. If you move the mouse pointer to a button on a toolbar and leave it there for a few seconds, a small tag displays a terse description of the button's purpose.

STATGRAPHICS *Plus* has two toolbars: the Application toolbar and the Analysis toolbar.

- **Application Toolbar**
  The Application toolbar displays the buttons you use to gain access to features you use to perform common tasks, such as opening and saving files, or performing editing tasks.

  This toolbar also accesses Analysis dialog boxes for basic plots, such as Scatterplots and Box-and-Whisker plots. Most functions that you can access using toolbar buttons you can also access using menus. For example, when you click the first button in the last set of buttons on the toolbar, the Analysis dialog box for the X-Y Plot Analysis displays; the pop-up identifier label says "Scatterplots" (see Figure 2-13). *See the Visual Glossary in Online Help, or Appendix A, Recognizing Icons and Buttons, for pictures and descriptions of icons and buttons.*

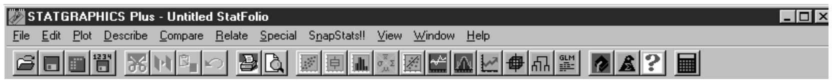

STATGRAPHICS Plus - Untitled StatFolio
File Edit Plot Describe Compare Relate Special SnapStats!! View Window Help

Figure 2-13.   The Application Toolbar

- **Analysis Toolbar**
  The Analysis toolbar appears below the Application toolbar and displays buttons you use to redisplay Analysis dialog boxes, access tabular and graphical options, save the results of certain statistical analyses, and when applicable, access additional graphics options (see Figure 2-14).

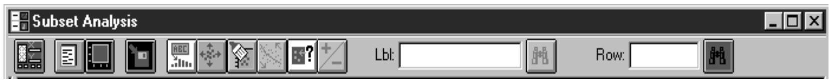


Figure 2-14.   The Analysis Toolbar

Using Taskbars

The long horizontal bar at the bottom of the desktop is a *taskbar.* In STATGRAPHICS *Plus*, your desktop will show two taskbars, one for the Application, and one for STATGRAPHICS *Plus* analyses. These are separated by the status bar (see Figure 2-15).

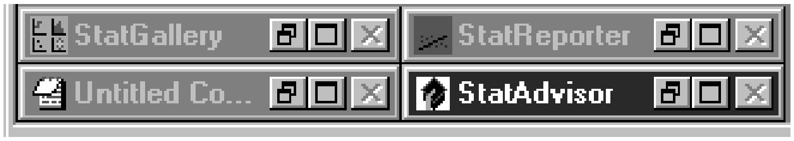


Figure 2-15. The Application and Analysis Taskbars Separated by the Status Bar

The Application taskbar contains three elements: the Start menu button; taskbar buttons for open applications, and the time of day indicator at the right end. The Analysis taskbar contains taskbar buttons that represent open windows in STATGRAPHICS *Plus*. For example, when you first open the program, five taskbar buttons appear: Untitled Comments, Untitled DataSheet, StatAdvisor, StatGallery, and StatReporter.

Using Different Types of Windows

A window can represent many objects. In STATGRAPHICS *Plus* there are eight distinct types of windows:

- Analysis window
- Application window
- Comments window
- DataSheet window
- Preview window
- StatAdvisor window
- StatGallery window
- StatReporter window.

Using the Analysis Window

The Analysis window appears after you select a plot or statistical analysis from the Menu bar and provide appropriate information in the Plot or Analysis dialog box.

An Analysis window contains these components: Analysis Icon title bar, Analysis toolbar, and text and graphics panes.

The Analysis Icon title bar displays the icon for the analysis and its title. The Analysis toolbar displays the buttons you use to redisplay dialog boxes; to access the tabular and graphical options; to save the results of certain statistical analyses; and, when applicable, to access additional graphics options. You can display both text and graphics in single panes — text in the left pane, graphics in the right — or each one individually in a maximized pane; or display both text and graphics in minimized panes.

Using the Application Window

The Application window appears when you first start STATGRAPHICS *Plus*. Figure 2-1 shows the Application window with some of the standard components labeled. At least some of the components appear on most, if not all, the windows.

When the Application window initially displays, it contains the Title bar and its icon, the Menu bar, the Application toolbar and its icons and buttons, and the Taskbar and its buttons. Initially, the taskbar buttons include: Untitled Comments button, Untitled Data button, StatAdvisor button, StatGallery button, and StatReporter button. Each of these buttons contains its own icon. The Status bar appears between the Analysis and Application taskbars.

Using the Comments Window

The Comments window appears when you click the Untitled Comments taskbar button, then click the Restore option. You can use this window to record important information about the StatFolios you create or you can use it like you would a notebook to make notations about information you need to remember.

Using the DataSheet Window

The DataSheet window appears in the Application window when you click the Untitled Data taskbar button, then click the Restore command. You use the window to create new DataSheets, to modify existing DataSheets, and to modify data using the DataEditor.

Using the Preview Window

The Preview window becomes active when you use the Print Preview... command from the File menu. When the Preview window displays, it contains the Application title bar and the Window toolbar with seven buttons: Print, Next Page, Prev Page, Two Page, Zoom In, Zoom Out, and Close. *The section, "To Use the Preview Buttons on the Preview Window Toolbar," in Chapter 6, Printing, Publishing and the StatFolio Start-Up Script, describes each of these buttons.*

Using the StatAdvisor Window

If a statistical interpretation is available for either an analysis or a graph, it will appear in this window. If an interpretation is not available, a message to that effect appears in the window instead. *For detailed information about the StatAdvisor, see Chapter 7, Using Special Features.*

Using the StatGallery Window

The StatGallery window appears when you click the StatGallery taskbar button, then click Restore on the pop-up menu. Use the window to arrange text and graphics that you will either view or print. *For detailed information about the StatGallery, see Chapter 7, Using Special Features.*

Using the StatReporter Window

The StatReporter window appears when you click the StatReporter taskbar button, then click Restore on the pop-up menu.

Using this window is much like using a word processor: You can create customized reports by copying and pasting text and/or graphics panes, and adding your own text.

Using Dialog Boxes

This section covers topics such as setting system-wide edit preferences, accessing analyses and the different types of Options dialog boxes, and performing some basic tasks involving analyses.

Modifying Dialog Box Appearance

Most dialog boxes in STATGRAPHICS *Plus* have several kinds of controls, including textboxes, command buttons, checkboxes, and option (radio) buttons. Some contain labeled tabs you use by clicking the tab to turn to its page.

Microsoft Windows contains three settings that control the size of the type in the desktop: Windows Standard, Windows Standard (extra large), and Windows Standard (large). If the title on a dialog box truncates, you will need to reset that control. A truncated title will end abruptly followed by a series of dots; for example,

Box-and-Whisker Plot Op.....

To reset the control:

1. Click START... to display the Start menu.

2. Click SETTINGS... CONTROL PANEL... to display the Control Panel window.

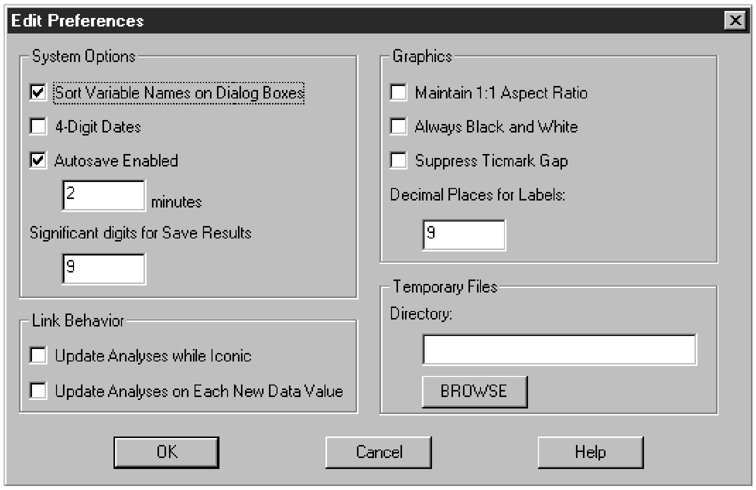
3. Double-click DISPLAY... to display the Display Properties dialog box.

4. Click the APPEARANCE... TAB to display the Appearance page. The top portion of the page shows the inactive window in the background, the active window in the foreground, and an OK message box in the front.

5. From the Scheme options, choose Windows Standard as the setting, then click OK.

Setting System-Wide Edit Preferences

Before you begin to use STATGRAPHICS *Plus* you might want to set system-wide edit preferences that will override the defaults for various aspects of the program. The Edit Preferences dialog box allows you to set behavioral preferences for some system, graphics, and link options. To access the dialog box, click Preferences on the Edit menu (see Figure 2-16).



Edit Preferences

System Options

☑ Sort Variable Names on Dialog Boxes

☐ 4-Digit Dates

☑ Autosave Enabled

2 minutes

Significant digits for Save Results

9

Link Behavior

☐ Update Analyses while Iconic

☐ Update Analyses on Each New Data Value

Graphics

☐ Maintain 1:1 Aspect Ratio

☐ Always Black and White

☐ Suppress Ticmark Gap

Decimal Places for Labels:

9

Temporary Files

Directory:

Figure 2-16. The Edit Preferences Dialog box

The System options portion of the dialog box lets you either retain or not retain settings for sorting variable names on dialog boxes, using four-digit dates, and enabling Autosave and setting specific times, in minutes. By default, variable names are sorted on dialog boxes. You can change that setting system-wide using this dialog box instead of changing it one by one later on, or changing it on the individual dialog boxes. Changing the setting for four-digit dates allows you to enter four-digit dates when you create or modify DataSheets instead of using the default of two-digit dates. The change takes place after you exit the current STATGRAPHICS *Plus* session, and re-enter the program.

The Graphics options portion of the dialog box lets you either retain or not retain a graphics aspect ratio of 1:1, which means that graphs will be created in a square shape; retain or not retain black and white graphs, which means you can override the default setting on the Profile page of the Graphics Options dialog box; and determine the number of decimal places that will appear in labels on charts.

The Link Behavior options portion of the dialog box lets you either update or not update analyses while they are in the nature of an icon, and update or not update analyses whenever a new data value is generated. Online Help contains detailed information about each of these options.

Using and Accessing Dialog Boxes and Options

This section explains the different types of dialog boxes and options you will use in STATGRAPHICS *Plus*, and how you access them.

Important Note: Beginning with Version 4, descriptions of the contents of the dialog boxes appear ***only*** in Online Help. To view the descriptions, click Help... on an Analysis dialog box, the Tabular and/or Graphical Options dialog boxes, or any other applicable dialog box.

To analyze data in STATGRAPHICS *Plus*, you will use a variety of dialog boxes and options. Some of the dialog boxes contain special list or text boxes that give you quick access to the variables in the open data file. You can type directly into an active text box or you can use the variables list to choose a variable name. A black arrow on a button indicates an active text box; a grey arrow indicates an inactive one.

In addition to dialog boxes, each analysis has a varying array of options from which you can choose to create reports and graphs, as well as pane and analysis options that are tied specifically to a report or graph. The dialog boxes and options are defined below.

- **Analysis Dialog Boxes**
  These dialog boxes are specific to an analysis. You use them to enter or choose the data that will be initially analyzed.

- **Analysis Options Dialog Boxes**
  These dialog boxes allow you to enter or choose options to vary the way the data are analyzed. These are named specifically for an analysis; for example, for the Weibull Analysis, this dialog box is called the Weibull Analysis Options dialog box. You access it by clicking the right mouse button on a text or graphics pane, then choosing Analysis Options... or Pane Options... from the pop-up menu. Analysis Options... are always specific to the analysis; the changes affect the analysis globally. Pane Options... are always specific to the pane that is currently open; the changes affect only that specific pane.

- **Tabular and Graphical Options Dialog Boxes**
  These dialog boxes contain a list of reports or graphs that are available for a specific analysis. You access them by clicking the Tabular or Graphical Options buttons on the Analysis toolbar.

- **Save Results Options Dialog Box**
  This dialog box allows you to save results calculated by an analysis to the datasheet. Check the items you want to save and select a name for each in the Target Variables column. Press OK to save the selected data to the datasheet. In Version 5, each analysis remembers the last selection of which results to save and under what names. An Autosave checkbox has been added to the dialog box. When checked, resuls will be resaved whenever the analysis recalculates its results, which can be important when running scripts. You access the dialog box by clicking the Save Results button on the Analysis toolbar.

Accessing dialog boxes is easy. First, choose the menu you want to use; then the submenu, if applicable; then the analysis. For example, to access the Weibull Analysis, which is part of the Life Data submenu, the access path is: DESCRIBE... LIFE DATA... WEIBULL ANALYSIS.... The steps for this path are shown in the instructions that follow.

To Access Analysis Dialog Boxes

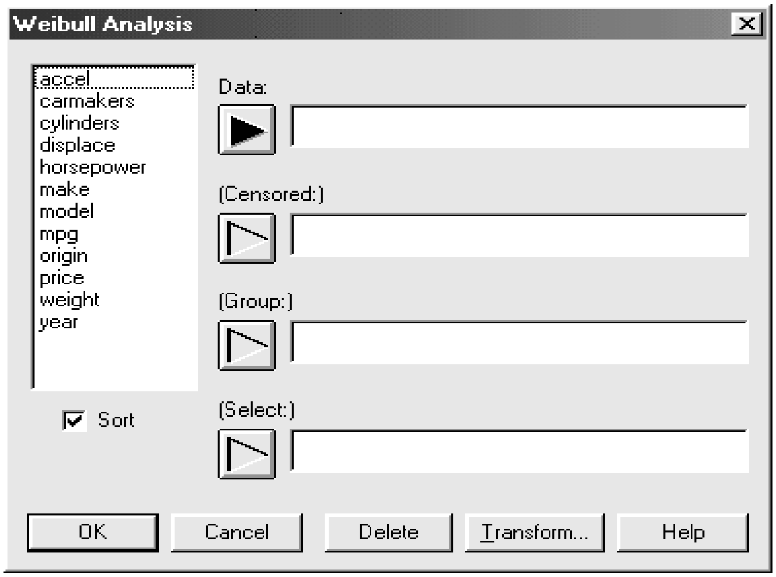
1. Choose DESCRIBE... LIFE DATA... WEIBULL ANALYSIS... to display the Weibull Analysis dialog box (see Figure 2-17).
2. Complete the dialog box, then click OK to display the first text and graphics panes of the Weibull Analysis.

To Access Tabular or Graphical Options Dialog Boxes

1. Click the Tabular or Graphical Options button on the Analysis toolbar to display the dialog box (see Figure 2-18).
2. Click the name of the Tabular or Graphical option you want to use, then click OK to display the report or graph.

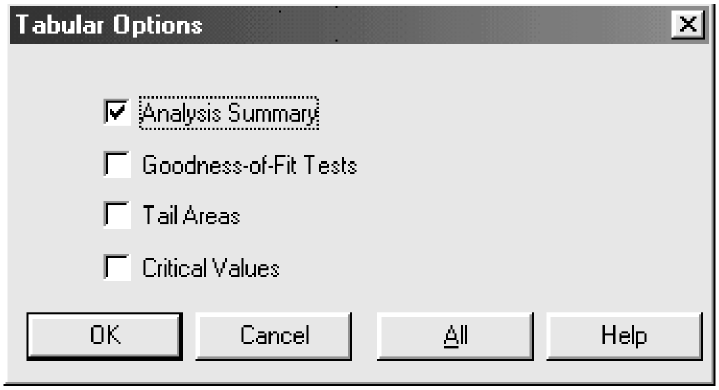
To Access Pane Options... or Analysis Options...

1. Click the right mouse button on a report or graph.
2. Choose either Pane Options... or Analysis Options... on the pop-up menu to display the appropriate Options dialog box (see Figure 2-19).



☑ Sort

Figure 2-17. The Weibull Analysis Dialog Box



☑ Analysis Summary

□ Goodness-of-Fit Tests

□ Tail Areas

□ Critical Values

Figure 2-18. The Tabular Options Dialog Box for the Weibull Analysis

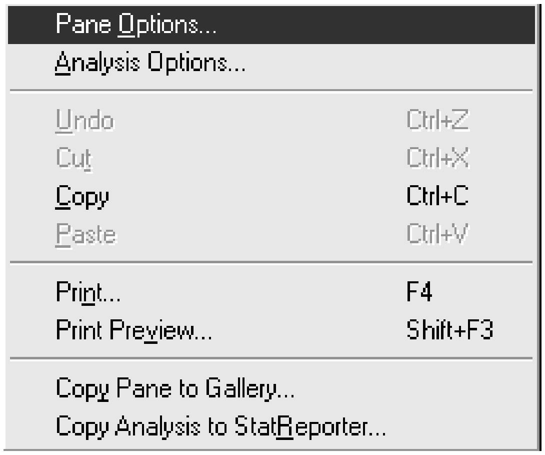


Figure 2-19. Pane... and Analysis...Options on a Pop-Up Menu

Working with Analyses

The Edit menu contains several commands you can use to modify an analysis at any time. These commands are: Change Text Font, Change Analysis Title, Copy Analysis.

To Change a Text Font

1. Access the analysis and maximize a text pane.

2. Choose EDIT... CHANGE TEXT FONT... to display the Font dialog box.

3. Make the changes. Choose a new font; change the font style from Regular, Italic, Bold, or Bold Italic; and view the changes.

The default text font in Version 5 is the same as the default Windows system font. In Version 4, the default text font was Courier New, size 9.

Note: Only select a non-proportionally spaced font, since output tables will not align properly using a proportionally spaced font.

4. Click OK.

To undo the changes, you must return to the Font dialog box and reverse your changes.

To Change Analysis Title

1. Access the analysis.

2. Choose EDIT... CHANGE ANALYSIS TITLE... to display the Change Analysis Title dialog box.

3. Type a new title into the Title text box, then click OK. The new title displays in the Analysis titlebar.

To Copy An Analysis

1. Choose EDIT... COPY ANALYSIS...; the program copies the analysis and displays the Analysis icon and name in a taskbar button.

3 Accessing and Using Files

STATGRAPHICS *Plus* is designed to handle a variety of file formats, including, STATGRAPHICS files created in other operating systems; spreadsheet files created in Lotus and Excel; database files created in SQL formats; and ASCII text files. In addition to data files, STATGRAPHICS *Plus* also handles files created using three special features in the product — StatFolios, the StatGallery, and the StatReporter — that are designed to speed up the way you work and cut down on time spent on repetitious data-entry tasks.

The first sections of this chapter explain the tasks of opening, saving, and closing files, as well as how you combine data files or StatFolios and sort a file. The remaining sections explain how you import files from other applications, how you use ODBC to query databases, and how to link and exchange files between other applications and STATGRAPHICS *Plus.*

Creating New Data Files

If the data you are analyzing are not already in files, use the DataEditor to enter the data and create a STATGRAPHICS *Plus* data file. *See the section, "Creating a DataSheet," in Chapter 4, Working with Data and DataSheets for information about how to create a new data file.*

Working with Existing Files

Note: In the sections that follow, there is discussion about the dialog boxes you use to open, close, and save a data file, StatFolio, StatGallery, or StatReporter. Because all these dialog boxes are similar except in specific name, the convention in the documentation is to refer to them as [*n*], where [*n*] equals the name of your selection; that is StatFolio, Data File, StatGallery, or StatReporter. The name *file* in the text is used interchangeably for all four types of files. The use of ellipses is to show the progression path.

Opening Existing Files

If you want to edit, add to, delete from, or perform an analysis on existing data, you must first retrieve the file that contains the data. An existing file is one you previously saved and includes data files, and files in StatFolios, the StatGallery, and the StatReporter. One way to open an existing file, is to use the Open command on the File menu.

After you access the Open [*n*]... dialog box, you need to provide information about the file. Online Help contains detailed descriptions of the options on all the dialog boxes, including the defaults and explanations of what you enter into the fields.

If you are using STATGRAPHICS *Plus* on a network, any number of users on the network can open the file and work with its contents. If multiple users have read/write privilege, processes should be established to ensure that concurrent use does not result in data loss.

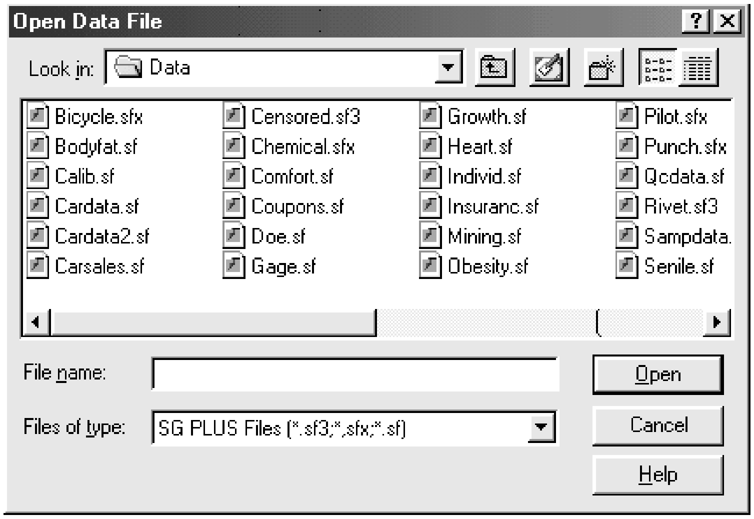
To Open an Existing File

1. Choose FILE... OPEN... OPEN [*n*]..., which displays the appropriate Open dialog box (see Figure 3-1).

2. Select or type the name of the file you want to open; then click Open. The program opens the file and places the appropriate icon in a Taskbar button.

Note: If you **make any recorded changes** to a file, then try to open another file or close the current file, a message displays asking if you want to save the current StatFolio, Data File, StatGallery, or StatReporter.

If you click Yes, and the file is yet untitled, another message box tells you to rename the file before you save it. Click OK and proceed as usual. When you click OK, the Save Data File As... dialog box displays, which you use to select or enter a name for the file.



Look in: Data

Files of type: SG PLUS Files (*.sf3;*.sfx;*.sf)

Figure 3-1. Open Data File Dialog Box

The Files of Type drop-down list in the dialog box shows the different types of files you can use with STATGRAPHICS *Plus;* they are listed in Table 3-1.

Table 3-1. Types of Files STATGRAPHICS *Plus* Uses

Name of File	Type of File	File Extension
StatFolio Files	StatFolios	*.sgp
Data Files	SG PLUS Files Created and/or Saved in STATGRAPHICS *Plus* for Windows, Versions 1, 2, and 3	*.sf , *.sfx, *.sf3
	SG PC Files Created and/or Saved in STATGRAPHICS for DOS	*.asf
	Executstat Files Created and/or Saved in Executstat	*.edf

Table 3-1. Continued

Name of File	Type of File	File Extension
	dBASE Database Files	*.dbf
	DIF Files	*.dif
	Lotus 1-2-3 Spreadsheet Files	*.wk*
	Excel Spreadsheet Files	*.xls
	ASCII Files	*.txt, *.* (ASCII Header)
StatGallery Files	StatGalleries	*.sgg
StatReporter Files	Rich Text Files	*.rtf

Saving Files

You can save a file anytime during a working session or when you finish a session of STATGRAPHICS *Plus.* You can save the file using the same name, or save it using a different name. Two commands on the File menu allow you to perform these tasks: Save... and Save As....

- **Save...**
  The Save... command saves changes to an existing file using the name with which it was last saved; that is, the file retains the name of the previously saved file. You can save a file as a StatFolio, a data file, a StatGallery, or a StatReporter. If you are saving a data file, you can also use the shortcut keys, **Shift+F11**.

- **Save As...**
  The Save As... command saves a file for the first time after you create it, saves it with a new name, saves it in a different format or a different directory or folder, or saves it on another disk. You can save a file as a StatFolio, a data file, a StatGallery, or a StatReporter. If you are using Save As... with a data file, you can also use the shortcut keys, **Shift+12**.

  Data files can be saved in one of three formats: standard SG Plus files (sf3), text files, and XML files. Text files are tab-delimited ASCII files that are easily readable by many programs including Excel. XML files easily transfer data over the Internet.

Note: Text files and XML files are intended only for exporting data for use by other programs. Continue to use SG Plus files for primary storage of data entered in STATGRAPHICS *Plus*.

To Save a File Using the Same Name

1. Choose FILE... SAVE... SAVE [*n*].... The program saves the file using the same name, and allows you to continue working.

The only exception to the above is for StatFolios. Using the path: FILE... SAVE... SAVE STATFOLIO..., a message displays asking you to rename the data file before saving the StatFolio. After you click OK, the Save Data File As... dialog box displays, which you use as usual.

To Save a File Using a Different Name

1. Select FILE... SAVE AS... SAVE [*n*] AS... to display the Save [*n*] As... dialog box.

2. Enter a new name for the file into the File Name text box; then click Save. The program saves the file and redisplays the data icon with the new file name in a Taskbar button.

Closing a File and Exiting

To Close a File

1. Select FILE... CLOSE... CLOSE [*n*]....

To Close a File and Exit STATGRAPHICS Plus

1. Select FILE... CLOSE... CLOSE [*n*]...; the program closes the file and removes the Taskbar buttons from the applicable window.

2. Select FILE... EXIT STATGRAPHICS.... The program displays a dialog box that asks if you want to save the current StatFolio.

3. Click Yes, No, or Cancel, depending on the action you want to take, and proceed accordingly.

Combining Existing Files

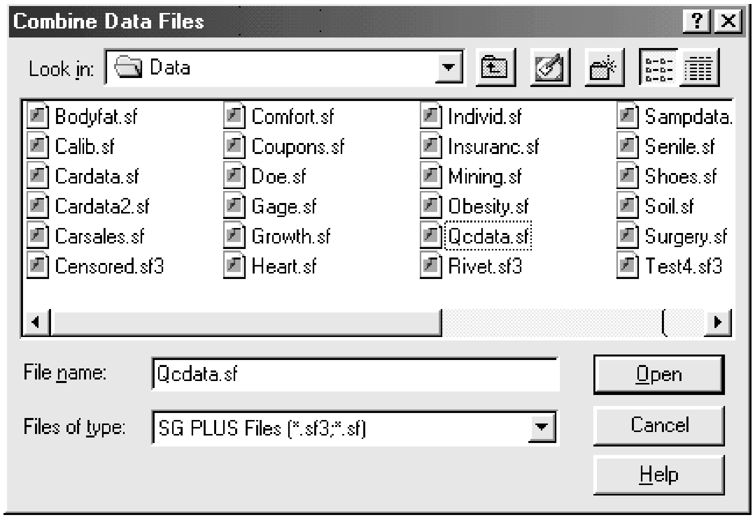
Sometimes you need to combine several files into a single file. You can combine data files or StatFolio files. You can also combine as many files as you like, but you must combine them two at a time.

Use the Combine... command on the File menu to combine the files. When you select the command and press OK, the Combine Data Files or Combine StatFolios dialog box appears.

After you access the Combine *[n]* dialog box you need to provide information about the files you want to combine. *See Online Help if you need information about this dialog box.*

To Combine Files or StatFolios

1. Open the first file you want to combine; for example, **cardata**.
2. Choose FILE... COMBINE... COMBINE [*n*]... to display the Combine [*n*]... dialog box (see Figure 3-2).



Combine Data Files

Look in: Data

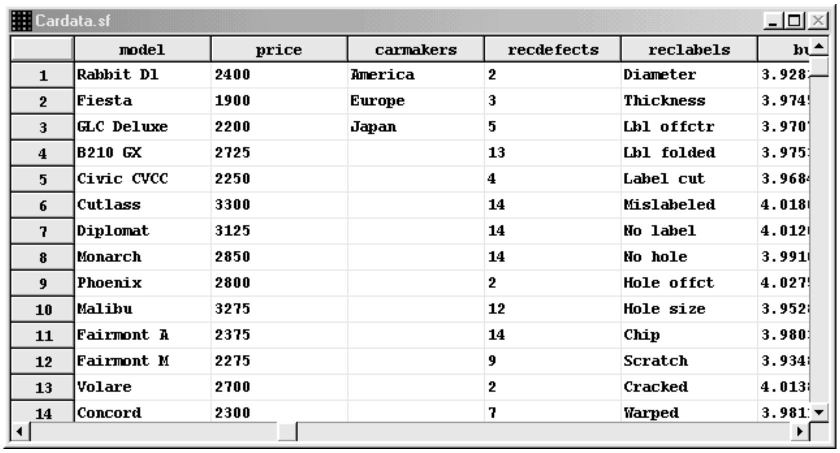
File name: Qcdata.sf

Files of type: SG PLUS Files (*.sf3;*.sf)

Figure 3-2. Combine Data Files Dialog Box

3. Open the second file you want to combine; for example, **QCDATA**; the program loads the data and combines the files.

4. Click the **cardata** taskbar button; then click Restore... to display the DataSheet that contains the files you combined (see Figure 3-3).



Cardata.sf

	model	price	carmakers	recdefects	reclabels	bu
1	Rabbit Dl	2400	America	2	Diameter	3.928
2	Fiesta	1900	Europe	3	Thickness	3.974
3	GLC Deluxe	2200	Japan	5	Lbl offctr	3.970
4	B210 GX	2725		13	Lbl folded	3.975
5	Civic CVCC	2250		4	Label cut	3.968
6	Cutlass	3300		14	Mislabeled	4.018
7	Diplomat	3125		14	No label	4.012
8	Monarch	2850		14	No hole	3.991
9	Phoenix	2800		2	Hole offct	4.027
10	Malibu	3275		12	Hole size	3.952
11	Fairmont A	2375		14	Chip	3.980
12	Fairmont M	2275		9	Scratch	3.934
13	Volare	2700		2	Cracked	4.013
14	Concord	2300		7	Warped	3.981

Figure 3-3. Untitled Datasheet Showing Combined Files

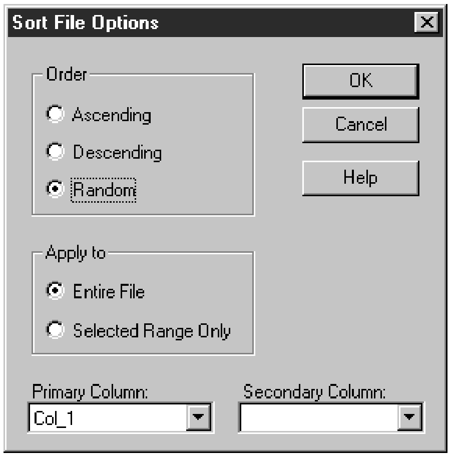
5. To retain the **cardata** name for the combined file, save the DataSheet as either a data file or as a StatFolio, using the directions above. Use FILE... SAVE AS... SAVE DATA FILE AS... to save the combined file with a new name.

   If you want to combine more files with the new file, repeat the process.

Sorting a File

Regardless of the order in which you enter values into the columns of a DataSheet, you can sort them into one column, two or more adjacent columns, or sort the entire DataSheet into ascending, descending or random order. You can sort the values on an entire DataSheet or by only the portion of it that you highlight.

To sort a file, use the Sort File Options dialog box, which you access by first accessing the DataSheet, then choosing EDIT... SORT FILE... from the Edit menu (see Figure 3-4). You can also access the dialog box by right clicking on a column of data, then choosing Sort File... from the pop-up menu.



Sort File Options

Order

○ Ascending

○ Descending

● Random

Apply to

● Entire File

○ Selected Range Only

Primary Column: Col_1

Secondary Column:

Figure 3-4.   Sort File Dialog Box

Importing Files

When you “import” data, you are actually opening data in a file created as another file type. Refer to Table 3-1 to review the file types that STATGRAPHICS supports.

If you are importing native files, the data are transferred directly into a DataSheet format. If you are importing files other than native files, you will have to specify parameters before you can import the data. To do this, use dialog boxes specific to the type of data you are importing. For example, if the file is in ASCII format, you will supply the parameters using the Read ASCII File dialog box; if the file is in Excel format, you will use the Read Excel File dialog box. STATGRAPHICS *Plus* now lets you read Excel spreadsheets through Excel `97. The Read Excel File dialog box also lets you define a character or string as the missing value parameter. During the import, that character or string is converted by STATGRAPHICS *Plus* into missing values (blanks).

To import spreadsheet data from noncompatible programs, first save the spreadsheet as an ASCII text file, then import the text file. *See the section, “To Import an ASCII File,” below.*

The steps below are for importing Excel files, then for importing an ASCII file. In each case, you will complete a dialog box: Read Excel File in the first case, and Read ASCII File in the last.

To Import Excel Files

1. Open STATGRAPHICS *Plus.*
2. Choose FILE... OPEN... OPEN DATA FILE... to display the Open Data File dialog box.
3. Using the Look In drop-down list, if necessary, choose the directory where the Excel file resides.
4. Choose EXCEL (*.XLS)... from the Files of Type... drop-down list.
5. Enter the name of the file into the File Name text box, then click OK to display the Read Excel File dialog box.
6. Complete the Read Excel File dialog box, then click OK to import the data. The dialog box lets you indicate if the variable names you want to import are stored in the first row of the Excel spreadsheet or if STATGRAPHICS should create default names. You can also enter the number of the worksheet you are importing, and/or provide a character or character string to be interpreted as a missing value.

To Import An ASCII File

1. Open STATGRAPHICS *Plus.*
2. Choose FILE... OPEN... OPEN DATA FILE... to display the Open Data File... dialog box.
3. Using the Look In drop-down list, if necessary, choose the directory where the ASCII file resides.
4. Choose ALL FILES (*.*)... from the Files of Type... drop-down list.
5. Enter the name of the file into the File Name text box, then click OK to display the Read ASCII File dialog box.
6. Complete the Read ASCII File dialog box, then click OK to import the data.

Using ODBC

Open Database Connectivity (ODBC) is a database-dependent interface from Microsoft Corporation. STATGRAPHICS *Plus* includes support for ODBC for 32-bit applications. The interface allows you to read data from a database file, such as one from Microsoft's Access or Excel, or from any other SQL-compliant database format for which you have an ODBC driver.

If you have never used ODBC, check the Windows Control Panel to see if you have the proper setup. If an ODBC icon appears, double-click it to see which data sources are set up. If the ODBC icon does not appear in the Windows Control Panel or if double-clicking the icon does not indicate you have the data sources you need, the ODBC command in STATGRAPHICS *Plus* will not work.

Version 5 ships with an ODBC driver. Thus, XML data files can be read using the Query Database procedure under the File menu. See the instructions below.

Important Note: The step-by-step instructions and explanations of the dialog boxes and the items on them in the remaining portions of this section, document the STATGRAPHICS *Plus* implementation of ODBC accessing Microsoft's Access database. If you previously used ODBC, the instructions and the names of the dialog boxes and the items on them may differ, depending on how ODBC was set up.

To Read XML Data Files

1. Using the ODBC Data Sources options under the Windows Control Panel, define an XML data source. Add a User DSN named "MML" using the "MERANT 3.60 32-bit XML" driver that STATGRAPHICS automatically installs in the Windows/System directory.

2. Choose FILE... OPEN... QUERY DATABASE (ODBC)... to display the Select Data Source dialog box.

3. Choose the XML file you want to open by clicking on the appropriate name, then clicking OK to display the Query Database dialog box.

Depending on the set up, a prompt may ask you to provide information such as a login ID and/or password, directory, and/or file information.

4. Enter the information, then click OK to display the Query Database dialog box.

Note: If you are working on a network and need help, see your system administrator.

5. Choose the name of the table (database) you want to open from the Table drop-down list. When you choose a table, the program displays the names of the fields in the Available Fields list box.
6. Choose the names of the fields you want to read into STATGRAPHICS *Plus* by double-clicking the field name. As you choose the names of the fields, they appear in the Selected Fields list box.
7. Click OK to display the Query Database (ODBC) - Use Rows dialog box
8. Complete the dialog box if you want to import a subset of the fields, then click OK. If you do not want to select a subset of the fields, click OK. The program imports the data you want to save and displays the name of the database in brackets in a data icon Taskbar button.

   As you complete the Query Database (ODBC) - Use Rows dialog box, keep these tips in mind.

   - If you do not make changes in the Use Rows text box, when you click OK, STATGRAPHICS *Plus* brings all the data into the DataEditor.
   - A semicolon is not necessary at the end of the statement you enter into the Use Rows text box.
   - STATGRAPHICS *Plus* uses implied **Select** and **Where** statements, so if you repeat them, a syntax error occurs.
   - If you use a text variable, use this syntax: “Description”=`software’, where Description is the name of the variable, which appears in double quotation marks. Use single quotation marks with the text string, software.
   - If you use a text variable, use this syntax: “temperature”100, where temperature is the name of the variable, which appears in double quotation marks; the numeric, 100, does not need single quotation marks.
   - If you use Boolean operators, you must use an ampersand (&) for the AND statement.

To Read SQL Databases

Note: The steps follow one set of dialog boxes. These will vary, depending on the ODBC setup on your machine.

1. Choose FILE... OPEN... QUERY DATABASE (ODBC)... to display the SQL Data Sources dialog box.

2. Choose the data source you want to open by clicking on the name, then clicking OK to display the Query Database dialog box.

   Depending on the set up, a prompt may ask you to provide information such as a login ID and/or password, directory, and/or file information.

3. Enter the information, then click OK to display the Query Database dialog box.

   Note: If you are working on a network and need help, see your system administrator.

4. Choose the name of the table (database) you want to open from the Table drop-down list. When you choose a table, the program displays the names of the fields in the Available Fields list box.

5. Choose the names of the fields you want to read into STATGRAPHICS *Plus* by double-clicking the field name. As you choose the names of the fields, they appear in the Selected Fields list box.

6. Click OK to display the Query Database (ODBC) - Use Rows dialog box

7. Complete the dialog box if you want to import a subset of the fields, then click OK. If you do not want to select a subset of the fields, click OK. The program imports the data you want to save and displays the name of the database in brackets in a data icon Taskbar button.

   As you complete the Query Database (ODBC) - Use Rows dialog box, keep these tips in mind.

   - If you do not make changes in the Use Rows text box, when you click OK, STATGRAPHICS *Plus* brings all the data into the DataEditor.

   - A semicolon is not necessary at the end of the statement you enter into the Use Rows text box.

   - STATGRAPHICS *Plus* uses implied **Select** and **Where** statements, so if you repeat them, a syntax error occurs.

   - If you use a text variable, use this syntax: “Description”=`software’, where Description is the name of the variable, which appears in double quotation marks. Use single quotation marks with the text string, software.

- If you use a text variable, use this syntax: “temperature”100, where temperature is the name of the variable, which appears in double quotation marks; the numeric, 100, does not need single quotation marks.
- If you use Boolean operators, you must use an ampersand (&) for the AND statement.

Note: To complete the SQL Data Sources, Query Databases Dialog Box, and Query Database (ODBC) - Use Rows dialog boxes, see their individual entries in Online Help where details for each of the items on all the dialog boxes are fully discussed.

Using OLE

OLE is a mechanism that allows you to create and edit documents that contain items or “objects” created by multiple applications. OLE was originally an acronym for Object Linking and Embedding. However, it is now referred to simply as OLE.

OLE documents, historically called *compound documents*, seamlessly integrate various types of data or *components*. Sound clips, spreadsheets, and bitmaps are typical examples of components found in OLE documents. In supporting OLE, STATGRAPHICS *Plus* allows you to use OLE documents without worrying about switching between different applications; OLE does the switching for you.

You use *container applications* to create compound documents and a *server application* or *component application* to create the items within the container documents. Any application you write can be a container, a server, or both.

OLE incorporates many different concepts that all work toward the goal of seamless interaction between applications. These areas include the following.

- ■ **Linking and Embedding**
  Linking and embedding are the two methods for storing items inside an OLE document that was created in another application.

- **In-Place Activation**
  Activating an embedded item in the context of the container document is called *in-place activation* or *visual editing*. The container application's interface changes to incorporate the features of the component application that created the embedded item. Linked items are never activated "in-place" because the actual data for the item is contained in a separate file, out of the context of the application containing the link.

Note: Linking and embedding and in-place activation provide the main features of OLE visual editing.

- **Automation OLE**
  OLE Automation allows one application to drive another application. The driving application is known as an *automation client* or *automation controller*, and the application being driven is known as an *automaton server* or *automation component.*

- **Compound Files**
  Compound files provide a standard file format that simplifies structured storing of compound documents for OLE applications. Within a compound file, "storages" have many features of directories and "streams" have many features of files. This technology is also called *structured storage*.

- **Uniform Data Transfer**
  Uniform Data Transfer (UDT) is a set of interfaces that allow data to be sent and received in a standard fashion, regardless of the actual method chosen to transfer the data. UDT forms the basis for data transfers by drag and drop. UDT now serves as the basis for existing Windows data transfer, such as the Clipboard and dynamic data exchange (DDE).

- **Drag and Drop**
  Drag and drop is an easy-to-use, direct-manipulation technique used to transfer data between applications, between windows within an application, or even within a single window in an application. You simply select the data to be transferred and drag it to the desired destination.

- **Component Object Model**
  The Component Object Model (COM) provides the infrastructure used when OLE objects communicate with each other. The MFC OLE classes simplify COM for the programmer.

Using DDE and OLE to Link and Exchange Files

STATGRAPHICS *Plus* supports two functions for linking and exchanging data with other applications:

- Dynamic Data Exchange (DDE)
- OLE (formerly Object Linking and Embedding).

DDE allows you to establish a link to data in another application using the Paste Link command. For example, you can copy a portion of a spreadsheet and paste-link it into a STATGRAPHICS *Plus* DataSheet. When the data in the source application is changed, the data will update in STATGRAPHICS *Plus*. The source application can remain open when STATGRAPHICS *Plus* updates, which differs from the StatLink... command, which updates from a closed file.

STATGRAPHICS *Plus* also supports OLE as a server application. In practice, this means that STATGRAPHICS *Plus* objects (text blocks or graphics) can be pasted or paste-linked into other applications. For example, you can copy a STATGRAPHICS *Plus* graphics pane and use the Paste Special... command to paste-link it into a word processing application. When changes are made to the graphic, they will also be made in the copy paste-linked into the word processor.

You can create a link by using one of the following commands.

- using the Paste Link... command
- using the Paste Special... command
- using the StatLink... command to tie StatFolios directly to data from spreadsheets, databases, and measuring devices such as digital micrometers via linking software.

Steps for creating a link using some of these methods follow.

To Create a Link Using the Paste Link... Command

1. Open a STATGRAPHICS *Plus* session.

2. Maximize a datasheet.

3. Open a spreadsheet application and load a spreadsheet file containing data. Highlight a section of data with the cursor.

4. **Choose** EDIT... COPY... **from the Edit menu in the spreadsheet application to copy the data.**

5. Return to STATGRAPHICS *Plus* and choose EDIT... PASTE LINK... from the Edit menu or datasheet pop-up menu. The data is displayed in the open datasheet.

   After an item has been linked, changes to the data made in the other application will also be made in the datasheet. Any analysis or StatFolio using that datasheet will also reflect the change.

6. When a DDE link is established, the FILE... LINKS... menu item is activated. You can use the dialog box associated with that menu item to update or change the links.

   Note: If the StatFolio you want to link was saved in versions of STATGRAPHICS *Plus* earlier than Version 4, you will have to resave it in Version 4 or 5 before you can use it in this type of link.

7. Choose SAVE AS... SAVE DATA FILE AS.... to save the datasheet and preserve the DDE link.

To Create a Link Using the Paste Special... Command

1. Save a STATGRAPHICS *Plus* analysis as a StatFolio.

2. Maximize the text or graphics pane in STATGRAPHICS *Plus*.

3. Choose EDIT... COPY... from the Edit menu.

4. Open the destination application and choose EDIT... PASTE SPECIAL... from the Edit menu. A dialog box asks you to select the type of object you want to paste.

5. Ensure that the PASTE LINK radio button is selected and SGWIN StatFolios Object is highlighted. When you click OK, the pane in STATGRAPHICS *Plus* is now displayed and linked with the destination application.

   When an item has been linked, it will reflect changes that are made in STATGRAPHICS *Plus*. For example, if a scatterplot is paste linked to a word processor page, change the number of points used to generate the plot in STATGRAPHICS *Plus*, then return to the word processor page, the plot is also updated there.

   Note: If the StatFolio you want to link was saved in versions of STATGRAPHICS *Plus* earlier than Version 4, you will need to resave it in Version 5 before you can use it in a link.

See Chapter 7, Using Special Features, for information about how you use the StatLink feature.

4 Working with Data and DataSheets

This chapter contains information about variables, how you assign formats to them and name them; how you create, modify, and edit a DataSheet; and how you recode data.

Working with Variables

Regardless of the type of statistical analysis you plan to perform, before you begin work you must organize the data. Organizing data includes collecting, classifying, and tabulating it. This section looks at the classification process; and how you create variables and DataSheets.

In STATGRAPHICS *Plus*, data are stored in variables, which are stored in DataSheets. A variable is defined as a characteristic or property of an individual population unit. A variable describes any finding (an attribute or a characteristic), that can change, can vary, or can be expressed as more than one value, or in various values or categories. A variable contains observations, or data values, that measure a certain characteristic of a population.

It is extremely important that you know the type of data you are working with because, although you can process various types of data in the analyses in STATGRAPHICS *Plus*, the results may be partially or totally meaningless if the analysis is not appropriate for that data. In some cases, the use of inappropriate data may cause the program to shut down.

Assigning Formats to Variables

The format of a variable affects the way the data appear in a DataSheet. When you are creating a new DataSheet, you can create variables in 10 different formats: Numeric, Character, Integer, Date, Month, Quarter, Time (in two formats), Fixed Decimal, and Formulas. The variables are shown on the Modify Column dialog box, which you use to assign a format to a variable.

The format includes the name, a comment that identifies the variable, the width required for the variable in the DataSheet, and a classification type, such as numeric or month.

Note: To change the width of a column, you simply enter another value for the width in the Width text box on the Modify Column dialog box.

- **Numeric**
  Numeric variables contain only numbers and have a floating decimal place. You can type as many numbers as you need after the decimal. You can also create numeric variables that contain numbers as codes, then store them in a single variable or use them to divide data into classes.

- **Character**
  Character variables contain characters and/or numbers that are treated as characters. You cannot run calculations on character variables. You can use character variables to define subgroups.

- **Integer**
  Integer variables contain numbers without decimal places.

- **Date**
  Date variables contain numbers in MM/DD/YY format. If you prefer to use a four-digit date, the format is MM/DD/YYYY. Use the Edit Preferences dialog box from the Edit menu to change the data format. You can use date variables as labels on a graph.

- **Month**
  Month variables contain numbers in MM/YY format. If you prefer to use a four-digit date, the format is MM/YYYY. Use the Edit Preferences dialog box from the Edit menu to change the data format.

- **Quarter**
  Quarter variables contain data in Q1/YY, Q2/YY, Q3/YY, and Q4/YY format. If you prefer to use a four-digit date, the format is Q1/YYYY, Q2/YYYY, Q3/YYYY, and Q4/YYYY. Use the Edit Preferences dialog box from the Edit menu to change the data format. You can use quarter variables as labels on a graph.

- **Time (HH:MM)**
  This type of Time variable contains numbers in a hour and minute format (HH:MM). You can use this Time variable as a label on a graph.

- **Time (HH:MM:SS)**
  This type of Time variable contains numbers in a hour, minute, and second format (HH:MM:SS). You can use this Time variable as a label on a graph.

- **Fixed Decimal**
  Fixed-decimal variables contain only numbers that have a fixed number of decimal places.

- **Formula**
  Formula variables contain data from existing columns you can use to perform calculations. The program greys out the resulting values and creates a link between the values used in the formula and those in the columns. You can also use the Formula column as a calculator and use numbers like 2145/3 to create a numeric answer for the first row of a column. *See "Updating a Formula," later in this chapter for instructions on how to do this.*

  You cannot edit the data in the Formula column; however, if you change the data in the columns used in the formula, the formula is updated. You can also update a formula using the Update Formulas... command on the Edit menu and the DataSheet pop-up menu.

Naming Variables

Follow these conventions when you name variables for STATGRAPHICS *Plus.*

- Each column (variable) in a DataSheet has a unique name that can contain from one to 32 characters.
- The name must begin with a letter (A-Z or a-z), an underscore (_), or a number sign (#). It cannot begin with a digit (0-9).
- The name can consist of letters (A-Z or a-z), digits (0-9), an underscore (_), a period (.), a blank ( ), as well as any of the following symbols: @, #, $, and %.
- The name **cannot** consist of any of the following characters: ‘ “ . < > ~ + - * / ^ = & | ( ) , ; !

- Embedded blanks in variable names are ignored; for example, Flow Rate and FlowRate are equivalent.

Creating a DataSheet

When you enter STATGRAPHICS *Plus* for the first time, you will begin work with a blank untitled DataSheet you use to set up and enter the data that will become the variables for an analysis.

Opening a Blank DataSheet

The first step in creating a DataSheet is to open a blank form.

To Open a Blank DataSheet

1. Open STATGRAPHICS *Plus*.
2. Click the data icon taskbar button.
3. Click Restore... on the pop-up menu to display the untitled DataSheet (see Figure 4-1).

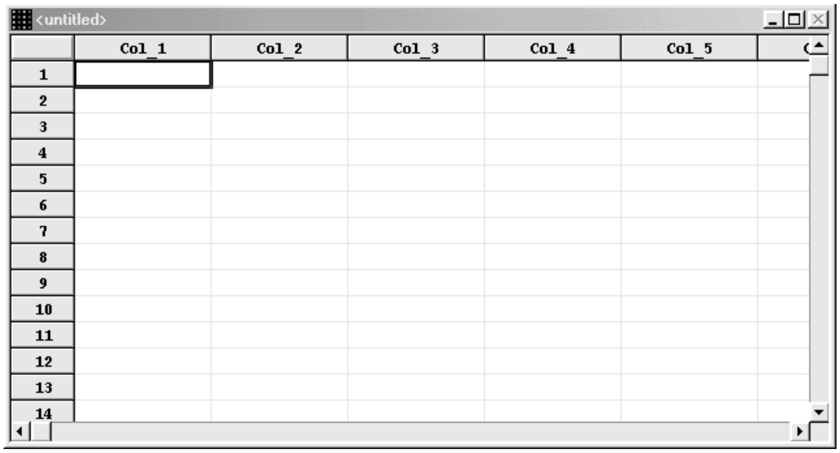


Figure 4-1. Untitled DataSheet

Setting Up a DataSheet

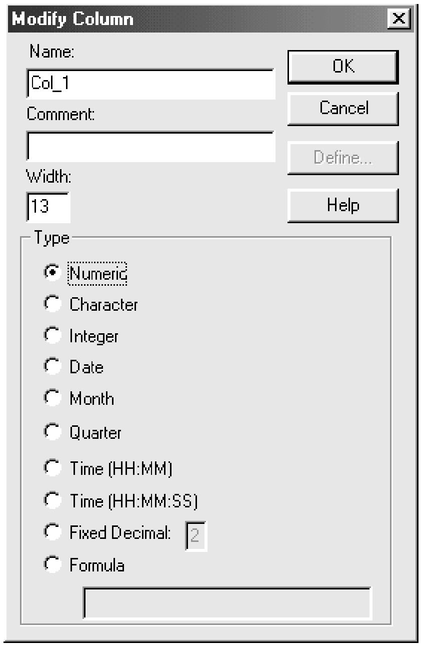
A DataSheet is a “spreadsheet-like” worksheet in the DataEditor that contains rows and columns. Cells in the rows and columns contain the values for the variables you will create. A DataSheet does not behave like a true spreadsheet (such as Microsoft’s Excel) because it does not have the functionality necessary to create results.

There are many ways you can set up a DataSheet. The steps below provide one method, which is to first create names for all the variables. You use the Modify Columns dialog box to do this.

To Create Names for Variables

1. Access a blank DataSheet.
2. Position and click the mouse pointer on Col_1 of the untitled DataSheet to highlight the column.
3. Access the Modify Column dialog box in one of three ways: double click on a column name, select Modify Column from the Edit menu, or use the DataSheet pop-up menu. (see Figure 4-2).
4. Type a name for the first variable (the title of the first column) in the Name text box.
5. Type a description about the variable in the Comment text box. The comment serves as a note to yourself (or to future users of this file), which explains something about the variable you are creating.
6. Enter an integer between 1 and 70 for the width of the first variable. See Online Help for the range of values you can use for each individual variable type.
7. Use Type options to choose the type of data you will be entering, then click OK to update and redisplay the DataSheet with the name of the variable appearing instead of Col_1 (see Figure 4-3).

   As this point you can either continue creating titles for the remaining variables or enter the values for the first variable, then create the next new variable name by following the steps above.



Name:
Col_1

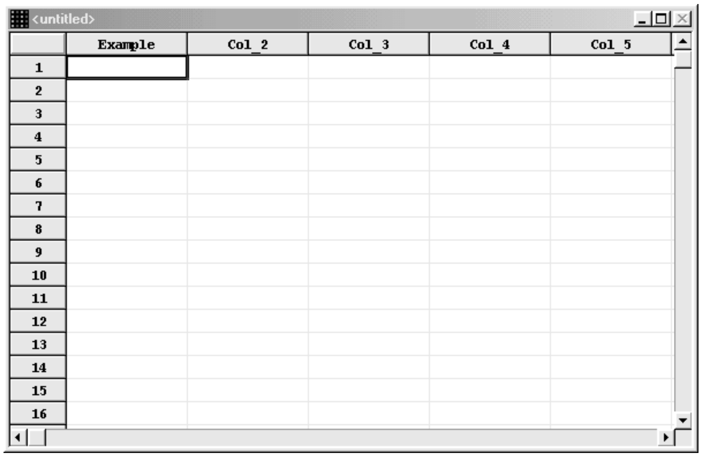
Comment:

Width:
13

Type

● Numeric
○ Character
○ Integer
○ Date
○ Month
○ Quarter
○ Time (HH:MM)
○ Time (HH:MM:SS)
○ Fixed Decimal: 2
○ Formula

Figure 4-2. Modify Column Dialog Box



	Example	Col_2	Col_3	Col_4	Col_5
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					

Figure 4-3. Redisplay of DataSheet with Name in Column 1

Entering Data

The steps below provide instructions for actually entering the data, including entering it by columns or rows, entering it by typing or using Copy and Paste, or entering it by creating values using the Generate Data dialog box. *For information about importing data from other software applications, see Chapter 3, Accessing and Using Files.*

When you enter the values, use the mouse or the arrow keys to move up and down through the cells. Use the mouse or the Tab and Shift-Tab keys to move right and left through the cells. If you enter more values than will display in a cell at one time, the DataSheet automatically scrolls the display across the cell.

The program organizes the values in the DataSheet (the representations of the variables) into a table of rows and columns. Each column represents a variable; each row represents one observation for each variable.

To Enter Data by Columns or Rows

1. Access the DataSheet.

2. Click the cell in which you want to begin entering data; then enter the first value.

3. Press the down arrow if you are entering the data by columns; then continue entering the values in a downward- or upward- column fashion. Press the right arrow if you are entering data by rows; then continue entering the values in a cross-wise right or left direction.

Entering Data by Typing the Values

When you enter data by typing it, you enter the values directly into the cells of the DataSheet. If you need to enter an observation that does not contain data (missing values), enter the missing value by leaving the appropriate cell empty (blank).

Entering Data Using Copy and Paste

Using this option, you can create a new DataSheet by copying all or portions of data from other DataSheets in STATGRAPHICS *Plus.*

To Enter Data from Other DataSheets

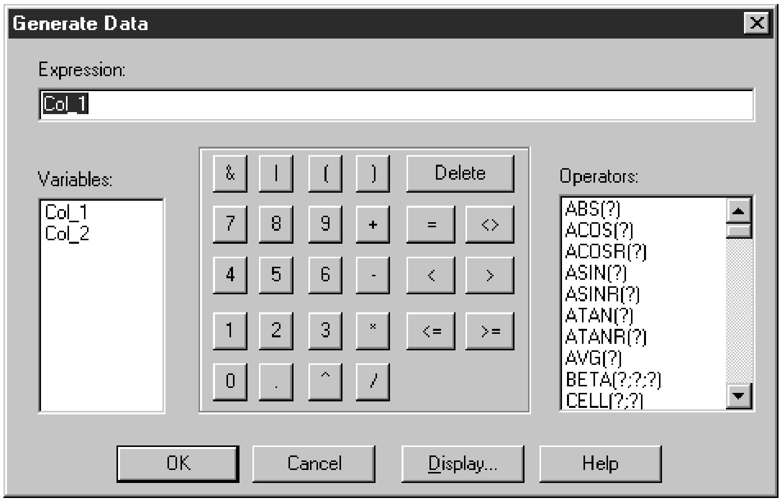
1. Open the DataSheet in STATGRAPHICS *Plus* that contains the data you want to use.
2. Highlight the data.
3. Choose EDIT... COPY... from the Menu bar.
4. Close the DataSheet; open the DataSheet that will contain the copied data.
5. Click the cell where you want to begin pasting the data.
6. Choose EDIT... PASTE... from the Menu bar. The program places the data in the cells of the new DataSheet.
7. Name and save the DataSheet.

To Enter Data from Other Applications

1. Access a new untitled DataSheet in STATGRAPHICS *Plus*.
2. Access and open the spreadsheet in another application that contains the data you want to use.
3. Highlight the data you want to copy.
4. Choose EDIT... COPY... from the Menu bar.
5. Click the cell of the new STATGRAPHICS *Plus* DataSheet where you want to begin pasting the data.
6. Choose EDIT... PASTE... from the Menu bar. The program places the new data in the cells of the DataSheet.
7. Name and save the DataSheet.

Generating Data for a DataSheet

You can also use operators to transform data to create values for a variable. The Generate Data dialog box lets you use variable names, operators, and/or a key pad to create a mathematical expression (see Figure 4-4).



Expression: Col_1

Figure 4-4. Generate Data Dialog Box

The mathematical expression can perform specific functions ranging from basic math to complex calculations. You create the expression by using, individually or in combination, the variable names and operators, as well as the key pad, which you can use to insert or delete numeric values.

In addition to the operator name, each item in the list of operators includes the correct syntax for the operator. For example, ABS(?) indicates that the ABS operator takes a single right argument, (?). Double-click the operator to display it in the Expression text box. *See Appendix B, Using Operators, and Online Help for descriptions and examples of all the operators.*

You can access the Generate Data dialog box by:

- using the DataSheet pop-up menu (the method outlined below)
- choosing EDIT... GENERATE DATA... from the Menu bar
- pressing the Transform... command button on Analysis dialog boxes.

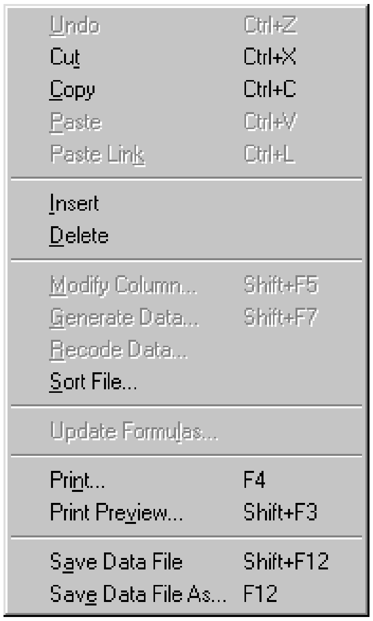


Figure 4-5. DataSheet Pop-Up Menu

To Access the Generate Data Dialog Box

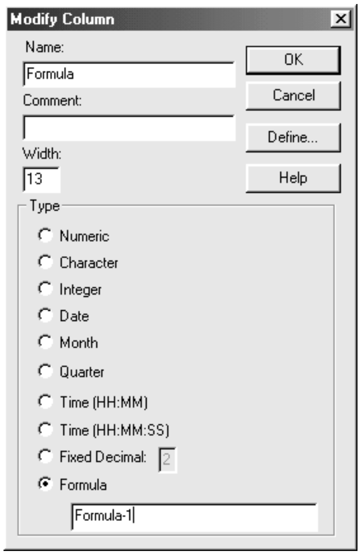
1. Place the mouse pointer on an open DataSheet and click the left button on the column heading to highlight a column.
2. Click the right button to display the DataSheet pop-up menu (see Figure 4-5).
3. Click Generate Data... on the pop-up menu to display the Generate Data dialog box.
4. Use the variable names, operators, and/or key pad, either individually or in combination, to create a function in the Expression text box.
5. Click OK to generate the data.

Updating a Formula

If a column in a DataSheet is made up entirely of numbers, you can use that data with the Update Formulas... command like a calculator to create numeric answers for that column.

To Update a Formula

1. Access an untitled DataSheet.
2. In Column 1, enter a column of numbers; then click on the Column 1 heading to select the column.
3. Access the Modify Column dialog box in one of three ways: double click on a column name, select Modify Column from the Edit menu, or use the DataSheet pop-up menu.
4. Enter a name for the column in the Name text box.
5. From the Type options, choose Formula.
6. In the Formula text box, type the name of the column, followed by a mathematical symbol indicating addition, subtraction, multiplication, or division (+,-,*,/) and a number that will be used with the mathematical symbol. For example, if you want to subtract 1 from each value in the column, you would enter the name of the column, followed by -1 (see Figure 4-6). Each time you access the Modify Column dialog box and click OK, the numbers are recalculated using the formula. This is repeated until you change the formula.



Name: Formula

Comment:

Width: 13

Type

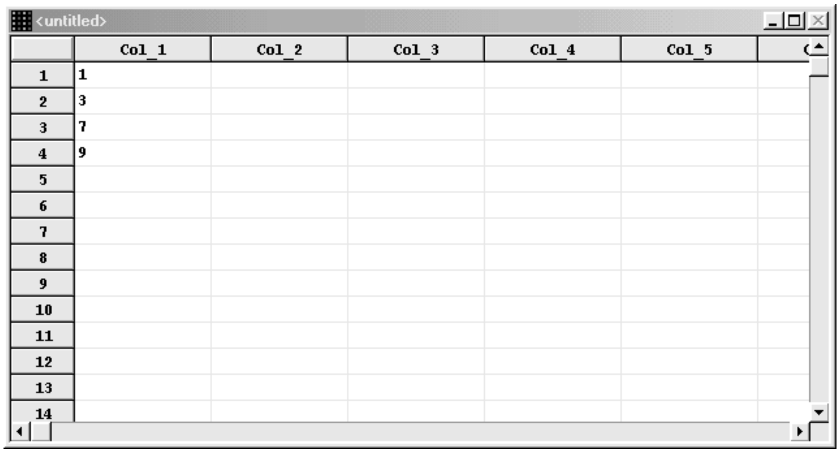
○ Numeric
○ Character
○ Integer
○ Date
○ Month
○ Quarter
○ Time (HH:MM)
○ Time (HH:MM:SS)
○ Fixed Decimal: 2
● Formula

Formula-1

Figure 4-6. Modify Column Dialog Box with a Formula for Subtracting 1

Try this example.

1. Access an untitled DataSheet.
2. In Column 1, enter ***1***, ***3***, ***7***, and ***9*** (see Figure 4-7); then click on the Column 1 heading to select the column.



	Col_1	Col_2	Col_3	Col_4	Col_5
1	1				
2	3				
3	7				
4	9				
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					

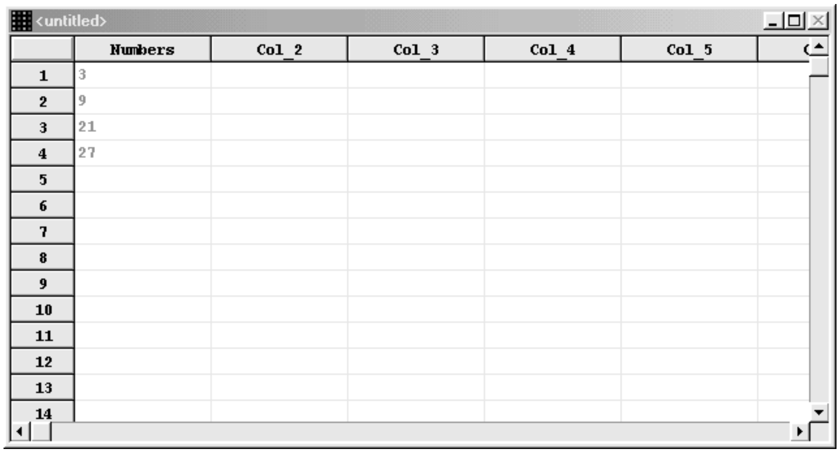
Figure 4-7. Untitled DataSheet with Values Entered

3. Access the Modify Column dialog box in one of three ways: double click on a column name, select Modify Column from the Edit menu, or use the DataSheet pop-up menu.
4. Enter ***Numbers*** for the name of the column in the Name text box.
5. From the Type options, choose Formula.
6. In the Formula text box, type ***Numbers*3***.
7. Click OK to update the values in the column (see Figure 4-8).

   Note: Formulas you create using the Update Formulas... command are not dynamically linked.

Modifying and Editing a DataSheet

After you create, name, and save a DataSheet, you can modify or edit it in several ways.



	Numbers	Col_2	Col_3	Col_4	Col_5
1	3				
2	9				
3	21				
4	27				
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					

Figure 4-8. DataSheet with Updated Values

Modifying a DataSheet

If you are modifying a variable using the Modify Column dialog box, you can rename it, change the comments about it, change the column width, change the classification type, or change the formula. These tasks were discussed in previous sections of this chapter.

Editing a DataSheet

If you are using commands on the Edit menu or the DataSheet pop-up menu, there are basic editing tasks you can perform using either of them. This includes undoing the last editing action or modification; cutting, copying, or pasting a highlighted entry; inserting or deleting a cell, row, or column; and sorting a file.

To Undo an Entry

Whenever possible, you can undo (reverse) the last editing action or modification.

1. Click the right button to display the DataSheet pop-up menu.
2. Click Undo... to reverse the last change and redisplay the DataSheet with the original data entered into the cell.

To Cut an Entry

1. Select the entry you want to cut by clicking on the cell or group of cells.
2. Click Cut... on the DataSheet pop-up menu to cut the entry or entries you selected and redisplay the DataSheet with the selected entry(ies) removed.

To Copy and Paste An Entry

1. Select the entry you want to copy by clicking on the cell or group of cells.
2. Click Copy... on the DataSheet pop-up menu to copy the entry and place it on the clipboard but leave it unchanged in the DataSheet.
3. Place the mouse pointer in the first position of the location into which the program will copy the entries.
4. Click Paste... on the DataSheet Pop-up menu to copy the entry into its new location.

To Insert a Column

1. Place the mouse pointer on the column that appears after the location in which you want to insert the new column.
2. Click the left mouse button to highlight the column; then click the right mouse button to display the DataSheet pop-up menu.
3. Click Insert... on the DataSheet pop-up menu to insert a new column before the selected column.

To Delete a Column

1. Place the mouse pointer on the column that you want to delete.
2. Click the left mouse button to highlight the column; then click the right mouse button to display the DataSheet pop-up menu.
3. Click Delete... on the DataSheet Pop-up menu to delete the column.

Recoding Data

The Recode data feature allows you to change the upper and lower limits of the data to create a new value. For example, suppose you are using the Tabulation Analysis to create a Piechart. Because your dataset contains 50 values and STATGRAPHICS *Plus* allows you to use only 20, you must change the limits to create fewer values, which would allow you to create the Piechart. You will use the Recode Data feature to create four categories of the following:

0-20
20-30
30-40
40-50.

1. Copy the **mpg** variable into Column 1 of a new DataSheet.

2. Type ***Groups*** into the Name text box on the Modify Column dialog box to name the new variable.

3. Click the left mouse button on the column of data; then click the right button on Recode Data... on the DataSheet pop-up menu to display the Recode Data dialog box.

4. Enter the following data into the Lower Limit, Upper Limit, and New Value text boxes. Figure 4-9 shows the completed dialog box.

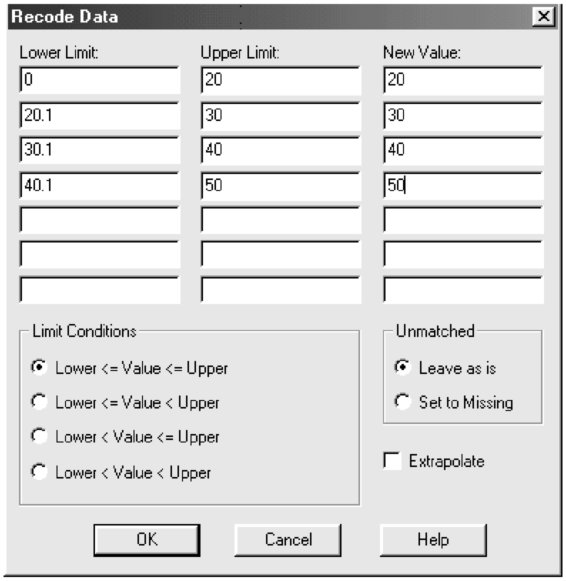
0	***20***	***20***
20.1	***30***	***30***
30.1	***40***	***40***
40.1	***50***	***50***

5. Click OK to recode the data.

6. Access the Tabulation Analysis. From the menus, choose: DESCRIBE... CATEGORICAL... TABULATION... to display the Tabulation Analysis dialog box.

7. Enter the **Groups** variable into the Data text box; then click OK to display the Analysis Summary and a Barchart in the Analysis window.

8. Click the Piechart option on the Graphical Options dialog box; then click OK to display the Barchart (see Figure 4-10). Notice that the data have been plotted into four groups within the new ranges.



Recode Data

Lower Limit:	Upper Limit:	New Value:
0	20	20
20.1	30	30
30.1	40	40
40.1	50	50

Limit Conditions

- ● Lower <= Value <= Upper
- ○ Lower <= Value < Upper
- ○ Lower < Value <= Upper
- ○ Lower < Value < Upper

Unmatched

- ● Leave as is
- ○ Set to Missing

□ Extrapolate

Figure 4-9. Completed Recode Data Dialog Box

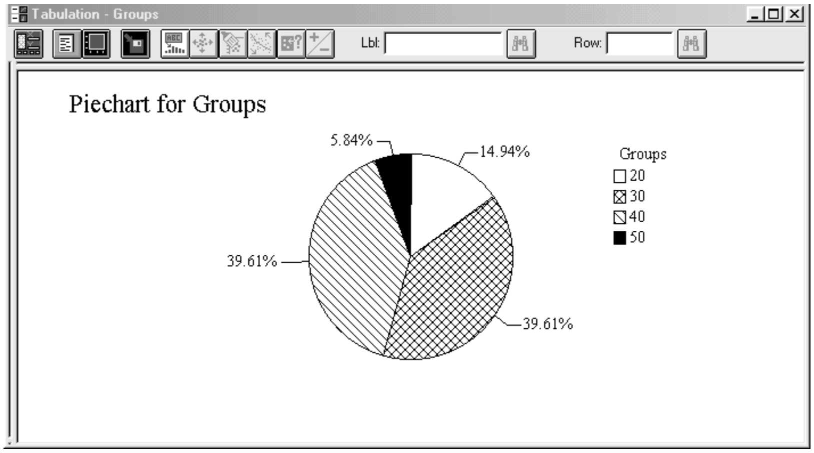


Figure 4-10. Piechart with Four Groups within New Ranges

5 Working with Graphs and Graphics Options

Statistical graphics, like statistical calculations, are only as good as what goes into them. Excellent statistical graphics should not only communicate complex ideas clearly, precisely, and efficiently, they should also provide powerful analytical tools for exploring the data. Tufte (1983) says that graphical displays should:

- show the data
- cause [you] to think about substance rather than methodology
- present many numbers in a small space
- make large datasets coherent
- encourage the comparison of different pieces of data
- reveal data at several levels of detail.

The first section of this chapter explains how you access, open, and save graphs. The second discusses the Graphics Options tabbed dialog box. Other portions of the chapter explain how you use buttons to add text; jitter, brush, smooth, identify, display, include, and exclude points; and how you perform basic graphics tasks, such as resize a graph and use zoom features. The final section explains how you set system-wide graphics preferences and how you set and save user preference profiles.

Accessing, Opening, and Saving Graphs

After starting STATGRAPHICS *Plus*, open the data file you want to use; then select an analysis from the menus on the Application toolbar.

Accessing and Opening a Graph

Although you can access and open graphs using several methods, the steps below outline one easy method.

To Access and Open a Graph

1. Choose the analysis you want to use from the Plot, Describe, Compare, Relate, Special, or SnapStats menus to display the appropriate Analysis dialog box.
2. Complete the Analysis dialog box, then click OK to display a text and a graphics pane in an Application window.

Saving a Graph

There may be times when you want to save a graph in a standard format so you can re-create it in other software programs. You can save graphis in several formats: metafile (*.wmf), JPEG 24-bit color (.jpg), TIF Color (.tif), PNG High Color (.png), Windows BMP High Color (.bmp), and Encapsulated PostScript (.eps). The metafile format saves the image as well as all the information you used to create the graph. Use the Save Graph dialog box to save the graph.

To Save a Graph

1. Double-click the graph to maximize it.
2. Choose FILE... SAVE GRAPH... from the Menu bar (or press F3) to display the Save Graph dialog box (see Figure 5-1) or right click on the graph to display the pop-up menu and click Save Graph.
3. Complete the dialog box, then click Save... to save the graph.

Changing and Enhancing Graphs

The most effective way to describe, explore, and/or summarize a set of numbers is to look at pictures of the numbers. “A picture is worth a thousand

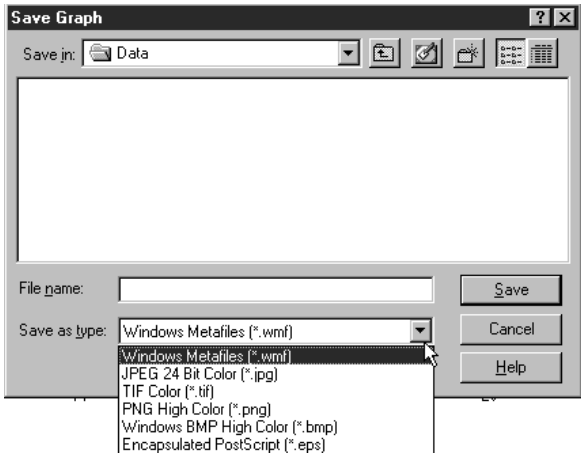

Save Graph
Save in:
Data
File name:
Save as type:
Windows Metafiles (*.wmf)
Windows Metafiles (*.wmf)
JPEG 24 Bit Color (*.jpg)
TIF Color (*.tif)
PNG High Color (*.png)
Windows BMP High Color (*.bmp)
Encapsulated PostScript (*.eps)
Save
Cancel
Help

Figure 5-1. Save Graph Dialog Box

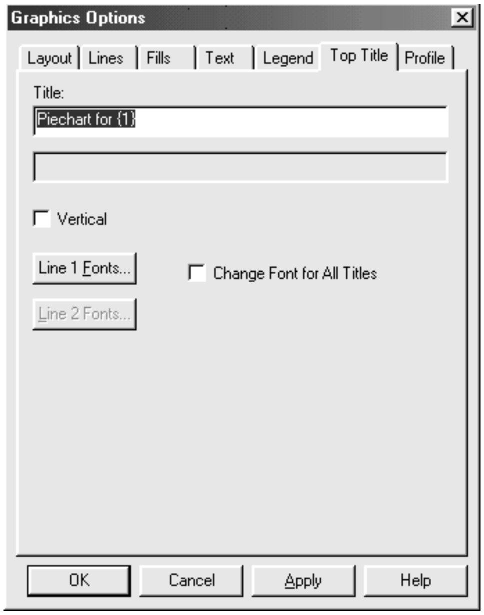
words" has particular significance in statistics because a graph or chart provides the simplest and most efficient display of data. Being able to customize graphs and charts through the use of points, lines, numbers, symbols, words, and color make them the simplest and most powerful way to visualize statistical information.

After you create and view a graph, you may want to change it, add information to it, or enhance it in other ways. You can change or enhance all the attributes that make up a graph such as its title, legends, axes, grid, points, or lines.

Using the Graphics Options Tab Dialog Box

STATGRAPHICS *Plus* contains a Graphics Options dialog box you use to access different pages through the use of tabs. The tabs are graph-specific; for example, if you are creating a Piechart and want to make changes to it, you first create the Piechart, then access the Graphics Options dialog box, which displays with all the Tab Pages applicable for modifying a Piechart (see Figure 5-2).

You can now add 3D effects to certain aspects of many graphs. These include the appearance of the axes, lines, bars, and pie slices. The 3D features are saved in the graphics profile and controlled by a new "3D effects” checkbox in the Layout, Lines and Fills pages of the tabbed graphics options dialog box.



Title: Piechart for {1}

□ Vertical

□ Change Font for All Titles

Figure 5-2. Graphics Options Dialog Box with Tab Pages Appropriate for a Piechart

If you make changes to a graph, leave any tab page, then decide you do not want to make the changes, clicking the right mouse button, then Undo on the pop-up menu, reverses all the changes.

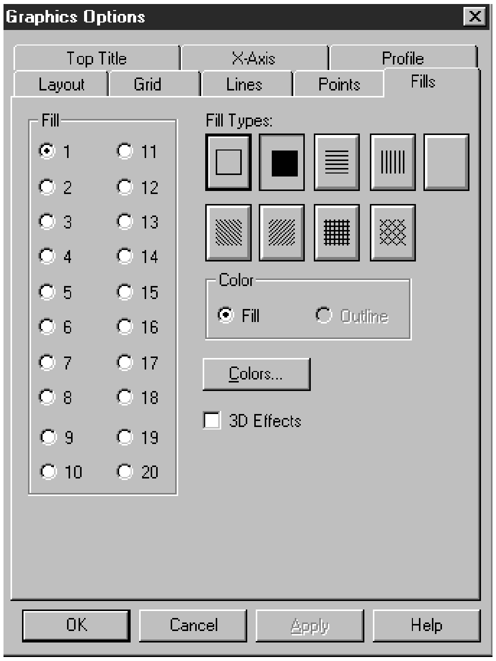
The tab pages are discussed in alphabetical order.

Using the Fill Tab Page

If you view and print a graph in color, you can coordinate the color of the outline and fill area of a graph. If you view and print the graph in black and white, you can use different fill patterns to achieve the same effect. You use the Fill tab page on the Graphics Options dialog box to change the style and color of fill patterns to distinguish the bars, boxes, wedges, and other fill areas on a graph. *See Online Help for descriptions of all the text boxes and options for the tab pages.*

To Change the Style and Color of Fill Patterns

1. Double-click the graph to maximize it.
2. Place the mouse pointer on a fill area on the graph.
3. Click the left button on the graph, the right to display the pop-up menu, then the left on Graphics Options to display the Graphics Options dialog box, which should open to the tab page that corresponds to the area of the graph you selected in Step 2 (see Figure 5-3).



Fill

● 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10
○ 11 ○ 12 ○ 13 ○ 14 ○ 15 ○ 16 ○ 17 ○ 18 ○ 19 ○ 20

Color

● Fill ○ Outline

□ 3D Effects

Figure 5-3. Fill Tab Page

4. In the Fill options portion of the dialog box, click the number of the area you want to change.

5. Choose one of the nine Fill types: Empty (no fill pattern), Solid, Horizontal Lines, Vertical Lines, Blank (to turn off the corresponding item), Left- and Right-Facing Diagonal Lines, Cross-Hatch Lines, or Diagonal Cross-Hatch Lines.
6. For the Color option, click either Fill or Outline to indicate whether you want color applied to the fill pattern or to the outline of the filled area.
7. Click the Colors... button to display the Color dialog box.
8. Choose the color you want to use for the fill or outline, then click OK.
9. Click Apply, then OK on the Fills tab page to process the changes.

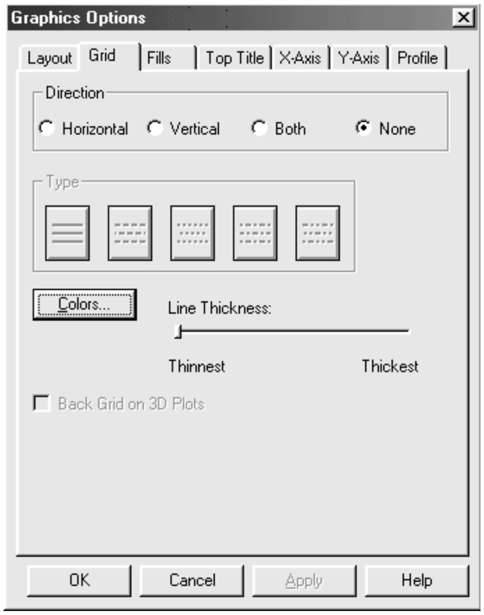
Using the Grid Tab Page

A grid, which is a series of lines that extend from an axis to an area on a plot, helps guide your eye from a point on a graph to its corresponding value or category on an axis. Using the proper grid, or using no grid, can make points easier to see, which visually enhances a graph.

The Grid tab page allows you to change the appearance of a grid by changing its direction, changing the type of the lines on it, changing its color and/or line thickness, and indicating if a back grid should display on three-dimensional plots.

To Change the Direction, Style, and Color of Grid Lines

1. Double-click the graph to maximize it.
2. Place the mouse pointer on an empty portion inside the graph frame.
3. Click the left mouse button on the graph, right click to display the pop-up menu, then left click on Graphics Options to display the Graphics Options dialog box. Click the Grid tab page (see Figure 5-4).
4. Choose one of the four Direction styles: Horizontal, Vertical, Both, or None.
5. Click the Colors... button to display the Color dialog box.
6. Choose the color you want to use.
7. Place the mouse pointer on the slider control for the line thickness and move the control from Thinnest to Thickest to change the width of the line. Release the control. **Note:** The line thickness feature is available only for solid lines.



Direction

○ Horizontal ○ Vertical ○ Both ● None

Type

□ Back Grid on 3D Plots

Figure 5-4. Grid Tab Page

8. If the Back Grid on 3D Plots option is active and you are using a 3D plot that would look better with a back grid, click the box.

9. Click Apply, then OK on the Grid tab page to process these changes.

Using the Label Tab Page

The Label tab page is available for plots, such as the Pareto Chart in the Quality and Design product, for which you want to create labels. Labels refer to text that appears on a graph as an identifier. A series of text boxes appear on the page that allow you to enter names for the labels. The letter between the braces, for example, {A}, is an internal system reference to variable names. The Fonts... command lets you change the font.

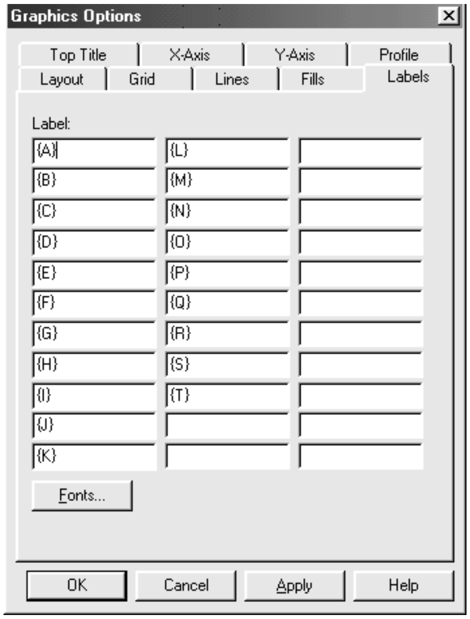
To Change the Name of a Label

1. Double-click the graph to maximize it.

2. Move the mouse pointer to the label you want to change.

3. Click the left mouse button to select the name of the label; markers appear at the corners of the selected label.

4. Click the left mouse button on the graph, right click it to display the pop-up menu, then left click on Graphics Options to display the Graphics Options dialog box opened to the Label tab page (see Figure 5-5).



Label:

{A}	{L}	
{B}	{M}	
{C}	{N}	
{D}	{O}	
{E}	{P}	
{F}	{Q}	
{G}	{R}	
{H}	{S}	
{I}	{T}	
{J}		
{K}		

Figure 5-5. Label Tab Page

5. In the Label text box, enter the new name for the label, then click OK to process the change.

6. Continue entering names until you finish.

7. Click the Fonts... button to display the Windows Font dialog box.

8. Choose the font, font style, size, color, and other attributes, then click OK.

9. Click Apply and OK on the Label tab page to process the changes.

Using the Layout Tab Page

The Layout tab page allows you to customize a graph by choosing a style and color for the tickmarks, the type of frame that will surround the graph, the color of the background and/or border of the graph, and the thickness of the axis line.

To Change Tickmarks

1. Double-click the graph to maximize it.
2. Place the mouse pointer on an empty portion inside the graph frame.
3. Click the left mouse button, then the right on Graphics Options on the pop-up menu to display the Graphics Options dialog box opened to the Layout page (see Figure 5-6).

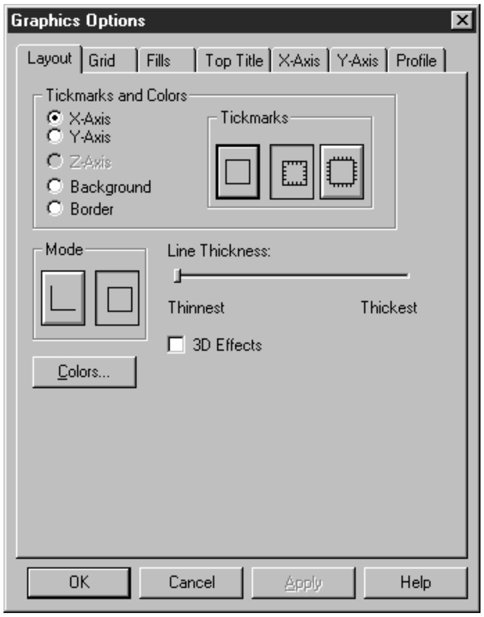


Figure 5-6. Layout Tab Page

4. Click the option that corresponds to the axis whose tickmarks you want to change (X-, Y-, or Z-Axis).

5. Choose one of the three Tickmark styles: No Tickmarks, Tickmarks Inside, or Tickmarks Outside.
6. Select the 3D Effects box, if desired.
7. Click the Colors... button to display the Color dialog box.
8. Choose the color you want to use, then click OK on the Color dialog box.
9. Click Apply, then OK on the Layout tab page to process all the changes.

To Change the Background and/or Border of a Graph

1. Double-click the graph to maximize it.
2. Place the mouse pointer on an empty portion inside the graph frame.
3. Click the left mouse button on the graph, the right to display the pop-up menu, then the left on Graphics Options to display the Graphics Options dialog box, opened to the Layout tab page.
4. Click either the background or the border option.
5. Select the 3D Effects box, if desired.
6. Click the Colors... button to display the Color dialog box.
7. Choose the color you want to use, then click OK on the Color dialog box.
8. Click Apply, then OK on the Layout tab page to process all the changes.

To Change the Shape of a Graph

1. Double-click the graph to maximize it.
2. Place the mouse pointer on an empty portion inside the graph frame.
3. Click the left mouse button, then the right on Graphics Options on the pop-up menu to display the Graphics Options dialog box opened to the Layout tab page.
4. Choose one of the two shape styles: L-Shaped or Square.
5. Click Apply, then OK on the Layout tab page to process the change.

Using the Legend Tab Page

Many of the analyses you use to plot data for two or more variables contain a legend; for example, a Multiple X-Y-Z Plot or a Barchart. A legend lists and explains the symbols on a graph. You use the Legend tab page on the Graphics Options dialog box to change the title of a legend, change the names of the individual variables that apply to a legend, and to change the font attributes such as the style, point size, and color. A series of text boxes appear on the page that allow you to enter names for the legends. The letter between the braces, for example, {A}, is an internal system reference to variable names. The Fonts... command lets you change the font.

To Change a Legend Title and Variable Names

1. Double-click the graph to maximize it.

2. Move the mouse pointer to the legend title you want to change.

3. Click the left button, then the right on Graphics Options on the pop-up menu to display the Graphics Options dialog box opened to the Legend tab page (see Figure 5-7).

4. Enter a title into the Title text box.

5. Click in the {A} Legend text box, then enter a name for the first variable.

6. Continue entering names until you finish.

7. Click the Fonts... button to display the Windows Font dialog box.

8. Choose the color you want to use, then click OK.

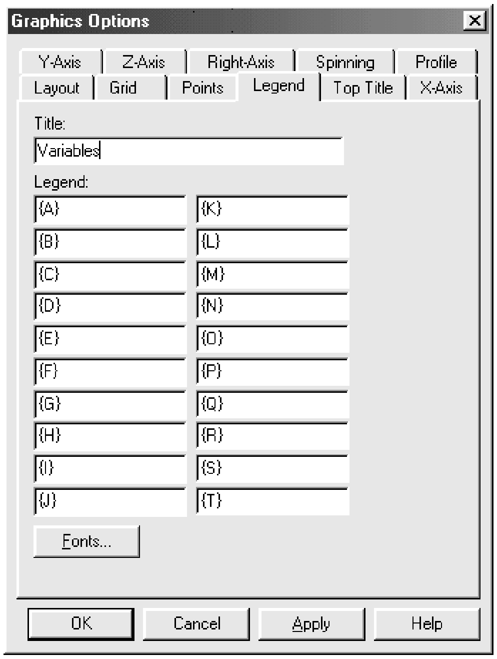
9. Click Apply and OK on the Legend tab page to process the changes.

Using the Line Tab Page

Changing the look of the lines that appear on a graph is useful when you need to enhance and clarify the graphical results. You use the Line tab page on the Graphics Options dialog box to set the style and color of the lines on a graph.

To Change the Type, Thickness, and Color of Lines

1. Double-click the graph to maximize it.



Title: Variables

Legend:

{A}	{K}
{B}	{L}
{C}	{M}
{D}	{N}
{E}	{O}
{F}	{P}
{G}	{Q}
{H}	{R}
{I}	{S}
{J}	{T}

Figure 5-7.   Legend Tab Page

2. Place the mouse pointer on the line you want to change.

3. Click the left button on the graph, the right to display the pop-up menu, then the left on Graphics Options to display the Graphics Options dialog box opened to the Line tab page (see Figure 5-8).

4. In the Line options portion of the dialog box, click the number of the line you want to change.

5. Choose one of the five Line styles: Solid, Dashed, Dotted, Dash-Dot, or Dash-Dot-Dot.

6. If you chose a Solid line, place the mouse pointer on the slider control for the line thickness and move the control from Thinnest to Thickest to change the width of the line. Release the control.

7. Click the Colors... button to display the Color dialog box.

8. Choose the color you want to use, then click OK.

9. Click Apply, then OK on the Lines tab page to process the changes.

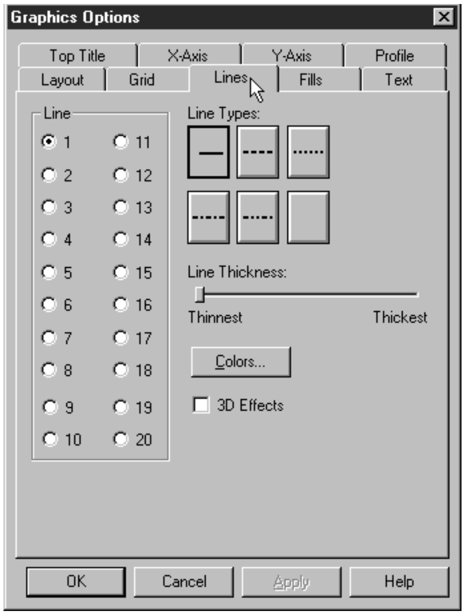


Figure 5-8. Line Tab Page

Using the Point Tab Page

When you create two- and three-dimensional scatterplots or similar types of plots, it is helpful to be able to change the shape, size, and color of the individual points. You use the Point tab page on the Graphics Options dialog box to set the style and color of points.

To Change the Shape, Size, and Color of Points

1. Double-click the graph to maximize it.

2. Place the mouse pointer on the point you want to change.

3. Click the left button on the graph, the right to display the pop-up menu, then the left on Graphics Options to display the Graphics Options dialog box opened to the Points tab page (see Figure 5-9).

4. In the Point options portion of the dialog box, click the number of the point you want to change.

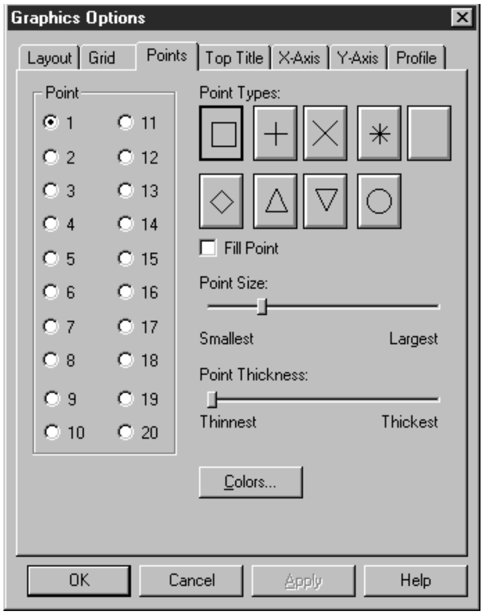


Figure 5-9. Points Tab Page

5. Choose one of the nine Point types: Square, Plus, X, Star, Blank, Diamond, Up Arrowhead, Down Arrowhead, or Circle.

6. Click the Fill Point check box or leave it blank to indicate if you want filled points. An unchecked box (the default) means the points will not be filled.

7. Place the mouse pointer on the slider control for the Point Size and move the control from Smallest to Largest to change the size of the point, then release the control.

8 Place the mouse pointer on the slider control for the Point Thickness and move the control from Thinnest to Thickest, then release the control.

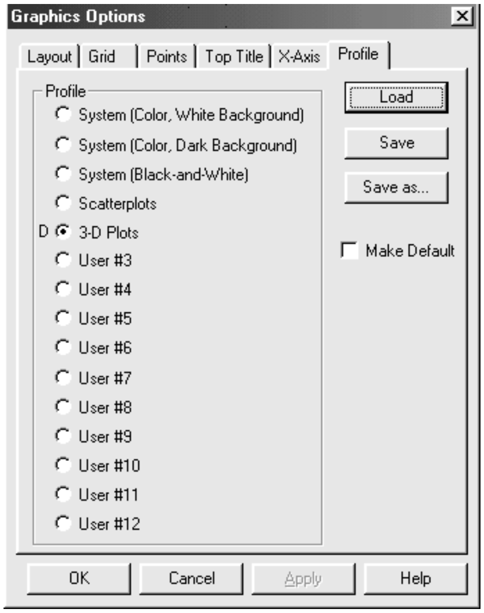
9. Click the Colors... button to display the Color dialog box.

10. Choose the color you want to use, then click OK.

11. Click Apply, then OK on the Points tab page to process the changes.

Using the Profile Tab Page

The Profile tab page on the Graphics Options dialog box lets you set up and save graphics profiles (see Figure 5-10). The first two options, System (Color, White Background) and System (Color, Dark Background) describe the two predefined choices for choosing the color and background of the graphs you produce; the choices are self-explanatory. System (Black and White) means that the graphs will always display in black and white format.



Profile

○ System (Color, White Background)
○ System (Color, Dark Background)
○ System (Black-and-White)
○ Scatterplots
D ● 3-D Plots
○ User #3
○ User #4
○ User #5
○ User #6
○ User #7
○ User #8
○ User #9
○ User #10
○ User #11
○ User #12

□ Make Default

Figure 5-10. Profile Tab Page

You can set up or edit up to 12 user profiles. When you set them up or edit them, you can use either numbers, letters, names, or a combination of them to create the name of the profile.

To Set Up a User Profile

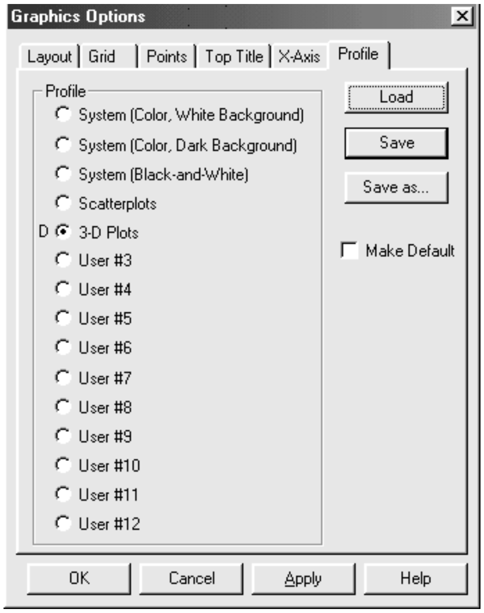
1. Access the Graphics Options dialog box.

2. Select your graphics preferences using the tabs on the Graphics Options dialog box.

3. Return to the Profile tab page; use Load..., Save..., or Save As.... Load loads the profile; Save, saves it; and Save As saves the profile as a specific name.

   For example, you might set up one profile for scatterplots, and another one for 3-dimensional plots. After choosing your preferences for each type of plot:

4. Click the first User # option.
5. Click Save As... to display the Save Profile dialog box.
6. In the text box, type ***Scatterplot*** to save the profile for scatterplots; click OK; the program displays the current graphics settings saved in a saved profile message. Click OK.
7. Repeat Steps 4, 5, and 6, using ***3-D Plots*** as the name for the User #2 option. When you click OK, names are displayed on the Profile tab page as the names for the two user options (see Figure 5-11).



Profile

○ System (Color, White Background)
○ System (Color, Dark Background)
○ System (Black-and-White)
○ Scatterplots
D ● 3-D Plots
○ User #3
○ User #4
○ User #5
○ User #6
○ User #7
○ User #8
○ User #9
○ User #10
○ User #11
○ User #12

□ Make Default

Figure 5-11. Profile Dialog Box with Two User Profiles

If you want to make one of the profiles the default:

8. Click the option button next to the setting you want to make the default.

9. Click Make Default..., then Load.... The D default indicator to the left of the option buttons, moves to the left of the new profile, indicating the default.

Using the Right Axis Tab Page

Some plots, such as the Multiple X-Y-Z Plot, contain a right axis. The Right Axis tab page lets you change the scale for this axis.

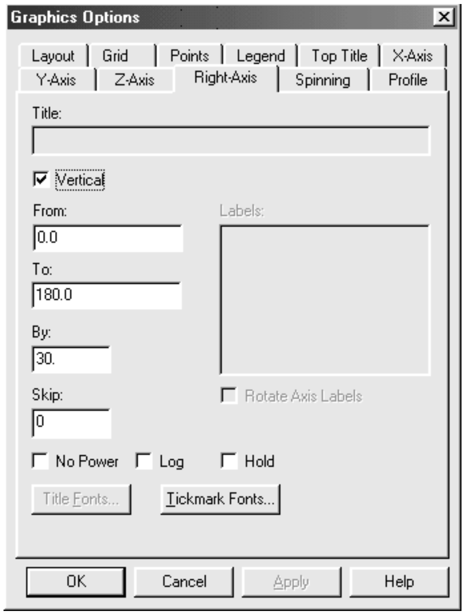
To Change the Scaling for the Right Axis

1. Double-click the graph to maximize it.

2. Position the mouse pointer on one of the values on the right axis.

3. Click the left button to select the axis; markers appear at the corners of the axis.

4. Click the left button on the graph, the right to display the pop-up menu, then the left on Graphics Options to display the Graphics Options dialog box opened to the Right Axis tab page (see Figure 5-12).



Graphics Options

Title:

☑ Vertical

From: 0.0

To: 180.0

By: 30.

Skip: 0

Labels:

☐ Rotate Axis Labels

☐ No Power ☐ Log ☐ Hold

Figure 5-12. Right Axis Tab Page

5. In the From text box, enter a number that will determine the minimum value for the axis.
6. In the To text box, enter a number that will determine the maximum value for the axis.
7. In the By text box, enter a number that will determine the distance between each tickmark on the axis scale.
8. In the Skip text box, enter the number of tickmarks that should be skipped when the axis is created.
9. Use the No Power check box to indicate if the axis scale for the tickmarks should be calculated using a power of 10.
10. Use the Log check box to indicate if the axis scale for the tickmarks should be calculated using decimal log scaling.
11. Use the Hold check box to indicate if the current scaling should be retained, even if you change the data in the Analysis dialog box.
12. Click the Tickmark Fonts... button to display the Font dialog box, then choose the font you want to use for the tickmarks, and click OK.
13. Click Apply and OK on the Right Axis tab page to process the changes.

Using the Spinning Tab Page

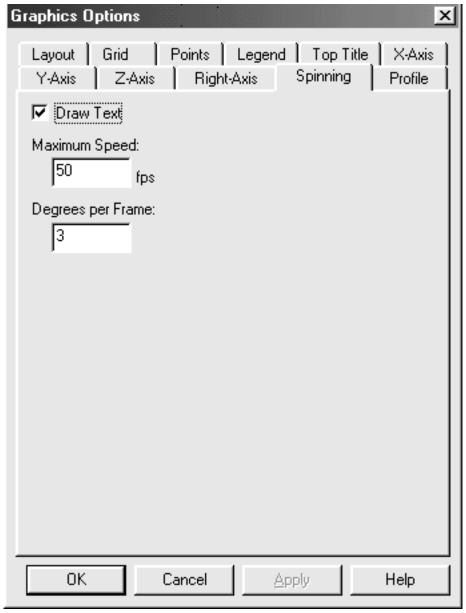
You can view three-dimensional graphs from different horizontal and vertical angles by spinning a graph to change the location of the viewpoint. To set preferences for spinning, use the Spinning tab page of the Graphics Options dialog box (see Figure 5-13).

You can set your preferences to indicate if text should be displayed on the graph while it is spinning; and to enter values that control the number of frames per second (fps) and the degrees per frame the graph will display while it is spinning.

To actually spin the graph, click the Smooth/Rotate button on the Analysis toolbar. *See the section "Using the Smooth/Rotate Button" later in this chapter for information about smoothing.*

To Spin Graphs

1. Double-click a three-dimensional graph to maximize it.



☑ Draw Text

Maximum Speed: 50 fps

Degrees per Frame: 3

Figure 5-13. Spinning Tab Page

2. Click the Smooth/Rotate button on the Analysis toolbar (the third button from the right), to display a slidebar/button combination on the toolbar that allows you to spin the graph (see Figure 5-14).

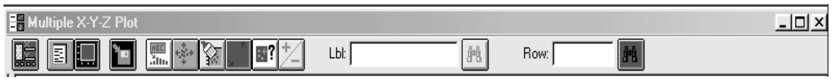


Figure 5-14. Smooth/Rotate Button

3. *Spin the Graph Horizontally:* Click the Horizontal button (the horizontal, double-headed arrow) to begin spinning the graph horizontally.

Spin the Graph Vertically: Click the Vertical button (the vertical double-headed arrow) to begin spinning the graph vertically.

4. Click the button a second time or click on the graph to stop the spinning. Use the slidebars beside the two buttons to control the horizontal and vertical angles of the graph. The range of the horizontal angle is from -90 to +90 degrees. The range of the vertical angle is from -180 to +180 degrees.

5. Click the right button, then click the left on the Reset Scaling/ Viewpoint... command on the pop-up menu to return the graph to its original viewpoint.

Using the Text Tab Page

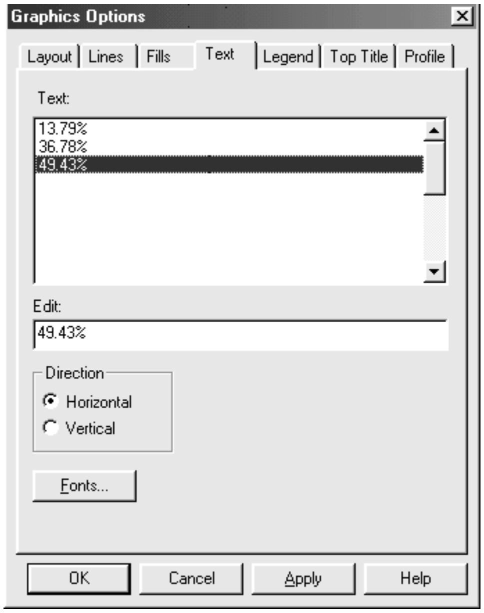
On certain graphs or if you add text to a graph, the Text tab page allows you to edit the text.

To Add Miscellaneous Text to a Graph

1. Double-click the graph to maximize it.
2. Place the mouse pointer on an empty portion inside the graph frame.
3. Click the left button, then the right on Graphics Options on the pop-up menu to display the Graphics Options dialog box opened to the Text tab page (see Figure 5-15).
4. In the Text text box, enter the text you want to add to the graph.
5. Click one of the Direction options to choose the way the text will display, either horizontally or vertically.
6. Click the Fonts... button if you want to make changes to the font or its attributes.
7. Click OK to process the changes.

Using the Top Title Tab Page

Top title refers to the title that displays at the top of a graph. You can change the entire name of the title or add or delete words from it. You use the Top Title tab page in the Graphics Options dialog box to make the text changes; to indicate the direction in which the title will appear; to change fonts for the one- or two-line title; and to indicate if the font should be changed for both titles if you are using a two-line title.



Text:
13.79%
36.78%
49.43%

Edit: 49.43%

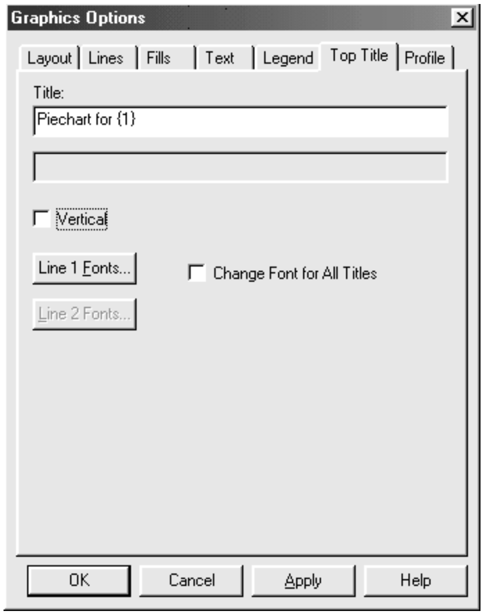
Direction
● Horizontal
○ Vertical

Figure 5-15. Text Tab Page

To Change a Top Title

1. Double-click the graph to maximize it.
2. Place the mouse pointer on the title you want to change.
3. Left click on the graph, right click to display the pop-up menu, then the left on Graphics Options to display the Graphics Options dialog box opened to the Top Title tab page (see Figure 5-16).
4. In the first Title text box, enter either a new title or make changes to the current title.

   Note: If the plot you are working with contains a title on a second line, the second text box will be activated and you can make changes there also.
5. Click the Vertical check box if you want the title to display vertically on the graph. If the box is checked, the title will display in that direction.
6. Click the Line 1 Fonts... button to display the Font dialog box you use to change the font for the text in line 1.



Title: Piechart for {1}

☐ Vertical

☐ Change Font for All Titles

Figure 5-16.   Top Title Tab Page

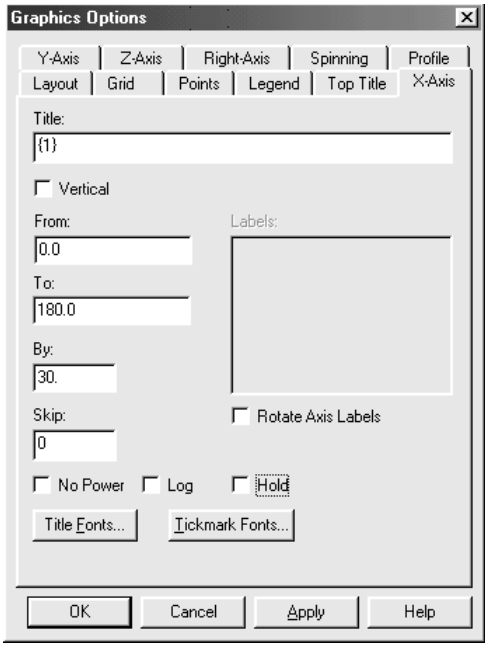
7. Click the Line 2 Fonts... button to display the Font dialog box you use to change the font for the text in line 2.
8. Click the Change Font for All Titles check box if you want to change to a font that will be used for all the titles on the graph.
9. Click Apply and OK to process the changes.

Using the X-Axis, Y-Axis, and Z-Axis Tab Pages

The scales against which a graph is plotted are known as axes. The X- Y-, and Z-Axis tab pages are identical except for defaults. You must change each axis independently; for example, if you want to make changes to the X-axis and the Y-axis, you must first access the X-Axis tab page, make the necessary changes, then access the Y-Axis tab page and make the necessary changes on that page.

To Change the Axis Scaling

1. Double-click the graph to maximize it.
2. Position the mouse pointer on one of the values on the axis whose scale you want to change (the X-, Y-, or Z-axis).
3. Click the left button to select the axis; markers appear at the corners of the selected axis.
4. Left click on the graph, right click to display the pop-up menu, then left click on Graphics Options to display the Graphics Options dialog box opened to the *n* Axis tab page (*n* indicates the selected axis). Figure 5-17 shows the X-Axis tab page; however, all three pages are identical except for their title.



Title: {1}

□ Vertical

From: 0.0

To: 180.0

By: 30.

Skip: 0

Labels:

□ Rotate Axis Labels

□ No Power □ Log □ Hold

Figure 5-17. X-Axis Tab Page

5. Enter a title for the axis in the Title text box.
6. Indicate if the title should display vertically or horizontally.
7. In the From text box, enter a number that will determine the minimum value for the axis.

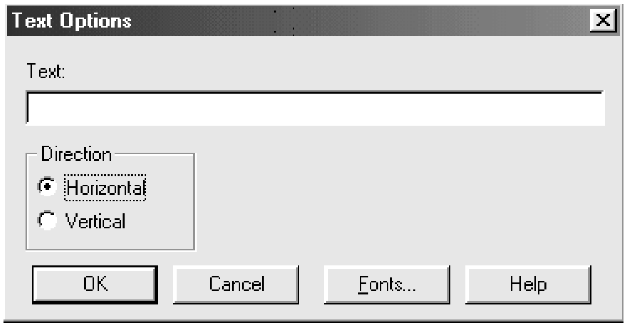
8. In the To text box, enter a number that will determine the maximum value for the axis.
9. In the By text box, enter a number that will determine the distance between each tickmark on the axis scale.
10. In the Skip text box, enter the number of tickmarks that should be skipped when the axis is created.
11. In the No Power check box, indicate if the axis scale for the tickmarks should be calculated using a power of 10.
12. In the Log check box, indicate if the axis scale for the tickmarks should be calculated using decimal log scaling.

13. In the Hold check box, indicate if the current scaling should be retained, even if you change the data in the Analysis dialog box.
14. Click the Title Fonts... button to display the Font dialog box you use to change the font for the title.
15. Click the Tickmark Fonts... button to display the Font dialog box you use to change the font for the tickmarks.
16. Click Apply and OK on the *n* Axis tab page to process the changes.

Using Graphics Task Buttons

In addition to using the Graphics Options tab pages to change and enhance graphics, you can use the special graphics task buttons located on the Analysis toolbar: Add Text, Jittering, Brushing, Smooth/Rotate, Set Points Labels, Locate Labels, Locate Row, and Include/Exclude.Using the Add Text Button

The Add Text button becomes available when you maximize a graph in an Analysis window. When you click the button, the Text Options dialog box displays. Enter the miscellaneous text you want to add and indicate the direction in which the text will display (see Figure 5-18). Use the Text tab page on the Graphics Options dialog box to edit the text.



Text Options

Text:

Direction

● Horizontal

○ Vertical

Figure 5-18. Text Options Dialog Box

Using the Jittering Button

The Jittering button is available when you maximize a Scatterplot in an Analysis window. When you plot data that contain many repeated values, it is likely that some points will occupy the same position. This causes two or more points to be plotted one on top of the other, which makes the scatterplot difficult to interpret.

Jittering prevents this and also allows you to see the density of the points in various locations on the plot. Jittering adds a small, random offset to each point. You can control the amount of horizontal and/or vertical offset that will affect each point. Use the Jittering dialog box to set the offsets.

To Jitter Points

1. Double-click the plot to maximize it.

2. Click the Jittering button on the Analysis toolbar (the button with four outward-pointing arrows) to display the Jittering dialog box (see 5-19).

3. Click and move the slider control horizontally and/or vertically until you reach the point on the bar that represents the offset you want to use for the points. Release the slider control.

4. Click OK to redisplay the plot with the points jittered.

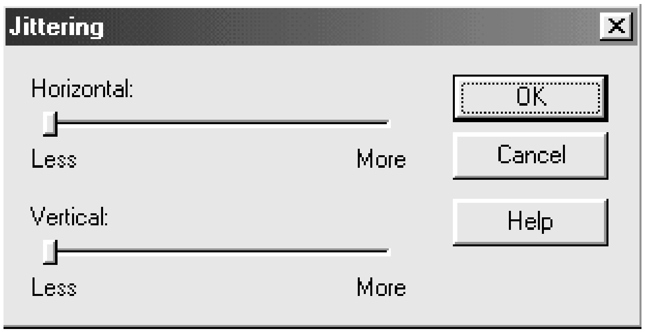


Figure 5-19. Jittering Dialog Box

Using the Brushing Button

The Brushing button becomes available when you maximize a Scatterplot in an Analysis window. Clicking the button displays the Brushing dialog box, which lets you select the variable that will be used to brush the points. Brushing shows the influence an added variable has on the data or a range of data. The added variable is shown in red. The red points should fall between the intervals if the left value is less than the right. If the right value is less than the left, the red points should fall outside the intervals.

When you click the Brushing button, the Brushing dialog box displays, which lets you choose or enter the name of a variable that will be used to brush the data. Label and Row text boxes on the Analysis toolbar are also replaced with horizontal and vertical slider controls and text boxes that display the minimum and maximum values.

To Brush Points

1. Double-click the plot to maximize it.
2. Click the Brushing button on the Analysis toolbar (a paint brush with scattered points) to display the Brushing dialog box (see Figure 5-20).
3. Choose or enter the name of the variable that will be used for brushing.
4. Click OK to redisplay the plot with the values for the added variable shown on the plot in red. Notice that the Label and Row text boxes on the Analysis toolbar have been replaced with two slider controls and text boxes (see Figure 5-21).

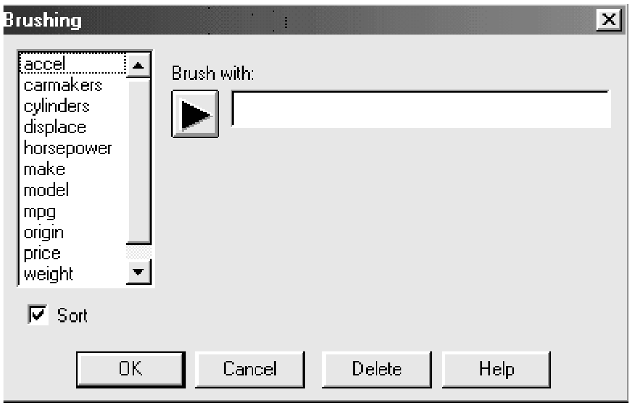


Figure 5-20. Brushing Dialog Box

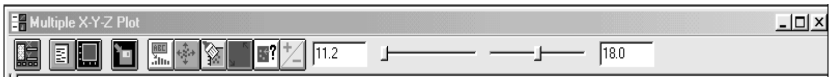


Figure 5-21. Label and Row Text Boxes on Analysis Toolbar

5. Move the slider controls or enter values into the text boxes to minimize or maximize the amount of brushing. The minimum and maximum values appear in the text boxes next to the slider controls. As you move the slider controls, the values for the brushing variable increase or decrease.

Note: This feature is not available if you choose the black and white option on the Edit Preferences dialog box.

Using the Smooth/Rotate Button

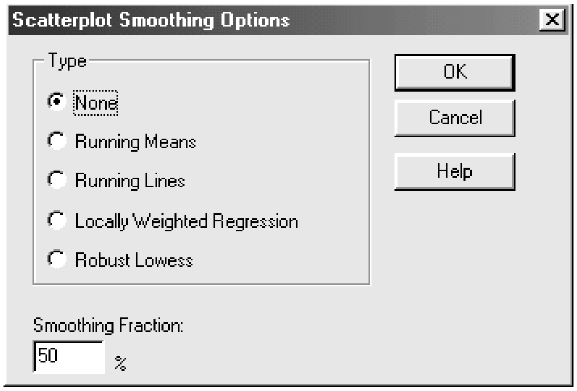
This button lets you smooth points on a plot or rotate (spin) a three-dimensional graph; it is available when you maximize a two- or three-dimensional plot in an Analysis window. The Spinning tab page on the Graphics Options dialog box explains how you use the spinning option and lets you set preferences for it. This section explains how you use the button to smooth points.

The pattern of points on two-dimensional plots, such as an X-Y Scatterplot, are often easier to see if you use a smoothing technique. The goal of smoothing is to reduce the fluctuations in the data to make long-term trends more apparent. You use the Smooth/Rotate button and the Scatterplot Smoothing Options dialog box to select smoothing options for these types of scatterplots.

Each of the smoothing options uses a slightly different method to estimate the relationship between Y and X. For each value of X, the program selects *k* points in the neighborhood of X, where *k* is determined from the smoothing fraction you enter into the Scatterplot Smoothing Options dialog box. For example, if the smoothing fraction is 30 percent, and there are 100 observations, 30 points in the neighborhood of X are selected. The Y values at the selected points are used to estimate the smoothed line for that value of X. The first two steps are repeated for each value of X. The smoothing options differ in how they define the "neighborhood" of X and also in the way they estimate the smoothed line.

To Smooth Points

1. Double-click the plot to maximize it.
2. Click the Smooth/Rotate button on the Analysis toolbar to display the Scatterplot Smoothing Options dialog box (see Figure 5-22).



Scatterplot Smoothing Options

Type

● None
○ Running Means
○ Running Lines
○ Locally Weighted Regression
○ Robust Lowess

Smoothing Fraction: 50 %

Figure 5-22. Scatterplot Smoothing Options Dialog Box

3. Choose one of the smoothing methods: None, Running Means, Running Lines, Locally Weighted Regression, or Robust Lowess.

4. Enter a number into the Smoothing Fraction text box that will be used to determine the size of the smoothing window.
5. Click OK to apply the smoothing method and redisplay the plot.

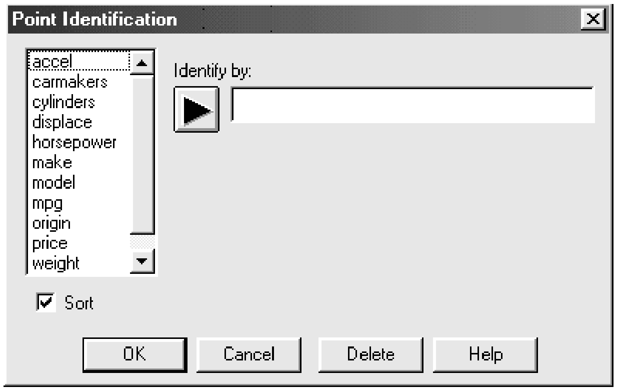
Using the Set Point Labels Button

This button becomes available when you maximize a Scatterplot in an Analysis window. Clicking the button displays the Point Identification dialog box that you use to enter or choose the name of a variable that will be used to identify the point.

When you create two- and three-dimensional scatterplots or similar types of plots, it is helpful to be able to identify and display individual points. For example, you may want to identify a point by its row number or by its label. In addition to displaying the values you use to create the plot (for example, the X-, Y-, and Z-axes), you can also select another variable to use as a label for each of the points. You can also highlight a row in the DataSheet that corresponds to a point on a plot to see other related values in the dataset.

To Identify a Point by Its Label

1. Double-click the plot to maximize it.
2. Click the Set Point Labels button on the Analysis toolbar (the scatterplot with a question mark) to display the Point Identification dialog box (see Figure 5-23).



☑ Sort

Figure 5-23. Point Identification Dialog Box

3. Enter or choose the name of the variable that contains the labels for the points.

4. Click OK to redisplay the plot.

5. Enter the name of the label in the Label (Lbl:) text box on the Analysis toolbar. You must enter the name exactly as it appears in the variable. For example, if you want to see the points that represent names of automobiles, select the **make** variable, then enter the name of an automobile (Ford, for instance) in the Label text box.

6. Click the Locate Label button on the Analysis toolbar (the binoculars to the right of the Label (Lbl:) text box. The color or shape of the points that represent the label whose name you entered changes, and the row number of the DataSheet displays in the Rows text box. If more than one point has the same label, more than one point will change; however, the row number will not display in the Row text box.

   If you want to identify points for additional labels, repeat Steps 3, 4, and 5, except enter the name of a different label in the Label text box.

To Display a Label by Its Point

1. Double-click the plot to maximize it.

2. Click the Set Point Labels button on the Analysis toolbar to display the Point Identification dialog box.

3. Enter or choose the name of the variable that contains the labels for the points.

4. Click OK to redisplay the plot.

5. Click the point whose label you want to display; the label for the point appears in the Label (Lbl:) text box.

   If you want to display additional labels for the same point, repeat Steps 2 and 3, except enter or choose a different variable name on the Point Identification dialog box.

   To display several labels at the same time, you can use operators in the dialog box. For example, for a file that contains the makes and models of cars, you could enter:

   Juxtapose (make,model)

to display both the make and model of the selected point. *For more information on operators, see Appendix B, Using Operators, or Online Help.*

To Locate a Point by Its Row and Axis

1. Double-click the plot to maximize it.

2. Click the left mouse button on a point to select it. As you do this, a pop-up box displays that contains the Row number then the point's location on the axis or axes.

This behavior is similar to that of the Locate... command on the Graphics pop-up menu. In the latter case, when you click Locate..., a light blue line appears on the plot. You can move the line in either a left or right direction to see the location of the points in the pop-up box.

Using the Exclude/Include Button

This button becomes available when you maximize a scatterplot in an Analysis window. Clicking the button excludes from the analysis, the point you select. When you select a point, then click the button, the program marks the point with a colored X. You can exclude as many points as you like. If you select a point that you previously excluded and click the button, that point is included in the analysis.

Note: This button is available for scatterplots where the point values are used in a calculation.

Performing Miscellaneous Graphics Tasks

In addition to using the tab pages on the Graphics Options dialog box to modify graphs, you can also perform other miscellaneous graphics tasks, such as repositioning text, resizing graphs, and using the Zoom and Locate features.

Repositioning Text

You can reposition text you added to graphs or existing text.

To Reposition Text on Graphs

1. Double-click the graph to maximize it.
2. Move the mouse pointer to the text you want to reposition.
3. Click the left button to select the text; markers appear at the corners of the selected text.
4. Click and drag the text to its new position.
5. Release the mouse button to release the markers, then click the button a second time outside the text area to remove the markers.

Resizing Graphs

You can change the vertical, horizontal, and/or diagonal dimensions of a graph to make it larger or smaller.

To Resize a Graph

1. Double-click the graph to maximize it, then click the graph to select it; markers appear at the corners of the graph.
2. *For Vertical Changes:* Place the mouse pointer on the top or bottom of the graph, then click.

   For Horizontal Changes: Place the mouse pointer on either side of the graph, then click.

   For Diagonal Changes: Place the mouse pointer on any corner of the graph, then click.
3. When the shape of the pointer changes to a double-headed arrow, drag the mouse until the graph is the size and/or shape you want, then release the button.

   Repeat this process as many times as necessary to achieve the size you want.

Using Zoom Features

The Graphics pop-up menu contains two commands that make it easier to view graphs: Zoom In... and Undo Zoom....

To Use Zoom In...

1. Access a plot that contains points such as a Scatterplot.
2. Click the right button on the portion of the graph that contains the points you want view at a closer range, then click Zoom In... from the pop-up menu. When you make the selection; a check mark appears in front of the command name and enables a zoom rectangle or rubber-band box.
3. Draw a rectangle around the points you want to view. You can use Zoom In... in a sequence of 10 times.

To Use Undo Zoom...

1. Access the Graphics pop-up menu, then click Undo Zoom... to cancel the action.

   You can use Undo Zoom... in a sequence of 10 times. When you reach the maximum number, the Undo Zoom... feature is greyed out on the pop-up menu.

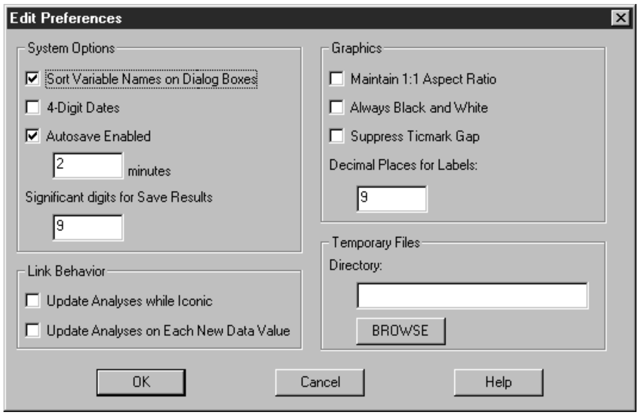
Using the Locate Feature

The locate feature is accessed from the right mouse button popup menu. Selecting the locate menu item causes a single locate line to appear on graphs with a scalable x-axis or two lines forming a cross-hair display on graphs with both a scalable x-axis and a scalable y-axis.

Setting System-Wide Graphics Preferences

The Edit Preferences dialog box contains options you use to select specific system-wide preferences about the graphics you create. For example, you can choose to always maintain a 1:1 aspect ratio, always display graphs in black and white, and indicate the number of decimal places that will appear in numeric legends.

These options appear in the Graphics portion of the Edit Preferences dialog box, which you access from the Edit menu (see Figure 5-24). The settings you choose affect all the analyses and remain in effect system-wide until you change them from this dialog box or override them. *See Online Help for details about each of these options.*



Edit Preferences

System Options

☑ Sort Variable Names on Dialog Boxes

☐ 4-Digit Dates

☑ Autosave Enabled

2 minutes

Significant digits for Save Results

9

Link Behavior

☐ Update Analyses while Iconic

☐ Update Analyses on Each New Data Value

Graphics

☐ Maintain 1:1 Aspect Ratio

☐ Always Black and White

☐ Suppress Ticmark Gap

Decimal Places for Labels:

9

Temporary Files

Directory:

Figure 5-24. Edit Preferences Dialog Box

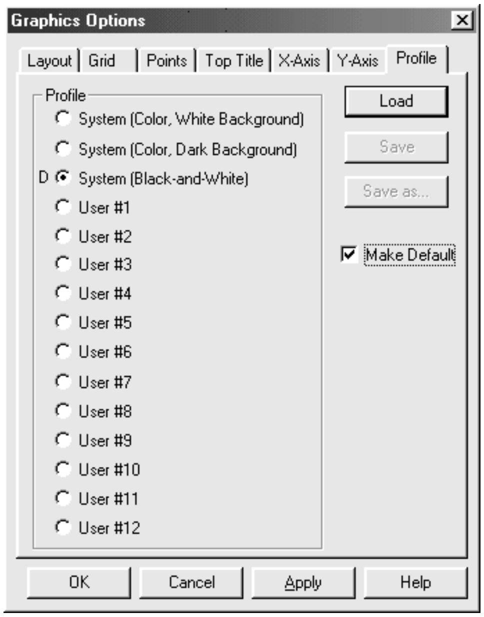
Setting and Saving User Preference Profiles

Often the type of task you perform on a regular basis requires that you set specific preferences for the graphics you produce. The Profile tab page on the Graphics Options dialog box allows you to select and save your preferences (see Figure 5-25).

The Profile page contains options you use to select preferences about the graphics you create. You can indicate if the profile will display graphs, system-wide, in color with a white background, in color with a dark background, or black and white. You cannot override the black and white setting indicated on the Edit Preferences dialog box. *See the "Using the Profile Tab Page," earlier in this chapter for more information.*

References

Tufte, E. R. 1990. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.



Profile

○ System (Color, White Background)

○ System (Color, Dark Background)

D ● System (Black-and-White)

○ User #1

○ User #2

○ User #3

○ User #4

○ User #5

○ User #6

○ User #7

○ User #8

○ User #9

○ User #10

○ User #11

○ User #12

☑ Make Default

Figure 5-25. Profile Tab Page with User Preferences

6 Printing, Publishing and The StatFolio Start-Up Script

This chapter describes how you print the various types of windows in STATGRAPHICS *Plus*, and how you use the various printing commands: Print Preview..., [Print] Setup..., and Page Setup....

The chapter also describes how to publish a StatFolio for viewing with a browser as well as view the published files using the user's default browser. Additionally a new option has been added under the edit menu titled "StatFolio Start-Up Script." This allows the user to define a script which will be run whenever the current StatFolio is loaded, either manually through the File menu or on program startup via the command line.

Printing Various Types of Windows

STATGRAPHICS *Plus* contains several types of windows you can print: Comments, DataSheet, Analysis, StatAdvisor, StatGallery, and StatReporter.

To print any of these windows, you first complete a dialog box that is specific to the type of window you want to print. The dialog boxes are Print Comments, Print DataSheet, Print Analysis, Print StatAdvisor, Print StatGallery, and Print StatReporter.

You can also use the Print Preview... command to preview the windows before you print them. *See "Using the Print Preview... Command," later in this chapter for complete information.*

There are several methods you can use to access these dialog boxes:

- Choose FILE... PRINT... from the Menu bar.
- Click the Print button on the Application toolbar.
- Click the Print... command on the pop-up menu that appears whenever you click the right mouse button on a pane.

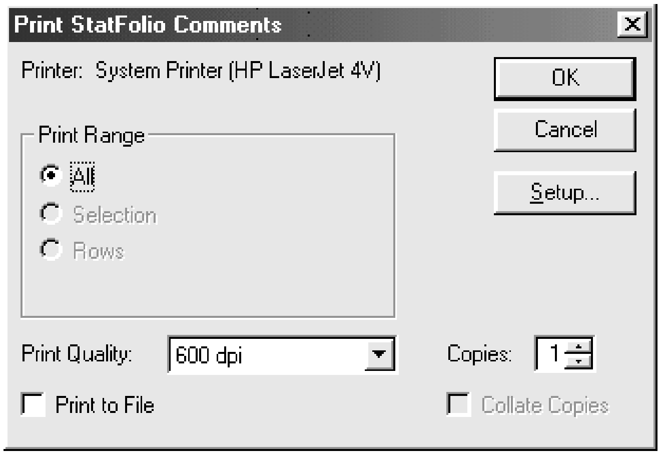
In this chapter, the step-by-step instructions use the first method.

Printing a [StatFolio] Comments Window

The [StatFolio] Comments window contains a record you can create for a StatFolio. You use it much like you would use a notebook. Printing the window provides a hard-copy record of the information.

To Print a [StatFolio] Comments Window

1. Click the Comments icon on the Taskbar button, then click Restore on the pop-up menu to display the Comments window.
2. Choose FILE... PRINT... from the Menu bar to display the Print StatFolio Comments dialog box (see Figure 6-1).



Printer: System Printer (HP LaserJet 4V)

Print Range

● All
○ Selection
○ Rows

Print Quality: 600 dpi

Copies: 1

□ Print to File

□ Collate Copies

Figure 6-1. Print [StatFolio] Comments Dialog Box

3. Choose the print range, select the print quality, indicate the number of copies, and if you want to print to a file. If your printer allows you to collate, indicate if you want the copies collated.
4. Click OK to print the window.

   The Setup... command on the dialog box displays the Print Setup dialog box, which allows you to select a different printer, and choose a paper size and source. *See "Using the [Print] Setup... Command," later in this chapter for complete information.*

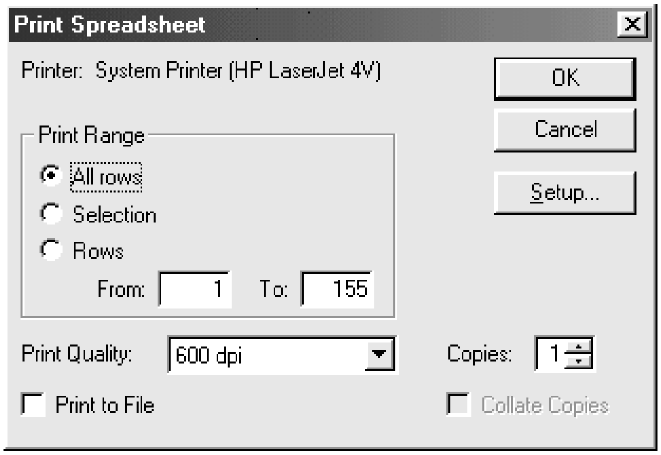
If you need help completing the dialog boxes, Online Help contains descriptions for all the options and text boxes.

Printing a DataSheet Window

The DataSheet window contains the DataSheet data for the file you select. You can print the entire DataSheet or only the portions you highlight.

To Print a DataSheet Window

1. Choose FILE... OPEN... OPEN DATA FILE... from the Menu bar to display the Open Data File dialog box.
2. Select the file you want to open by clicking on its name, then click the Open button. The program opens the file and displays the Data icon with the file name in a Taskbar button.
3. Click the Taskbar button that contains the name of the file you just opened.
4. Click Restore on the pop-up menu to load the data and display the DataSheet.
5. Choose FILE... PRINT... from the Menu bar to display the Print DataSheet dialog box (see Figure 6-2).



Print Spreadsheet

Printer: System Printer (HP LaserJet 4V)

Print Range

● All rows
○ Selection
○ Rows
From: 1 To: 155

Print Quality: 600 dpi

Copies: 1

□ Print to File

□ Collate Copies

Figure 6-2. Print DataSheet Dialog Box

6. Choose the print range, select the print quality, indicate the number of copies, and if you want to print to a file. If your printer allows you to collate, indicate if you want the copies collated.

7. Click OK to print the window or the portion of it you selected.

Printing an Analysis Window

An Analysis window is the window that appears after you complete an Analysis dialog box for an analysis you select from one of the five STATGRAPHICS *Plus* menus.

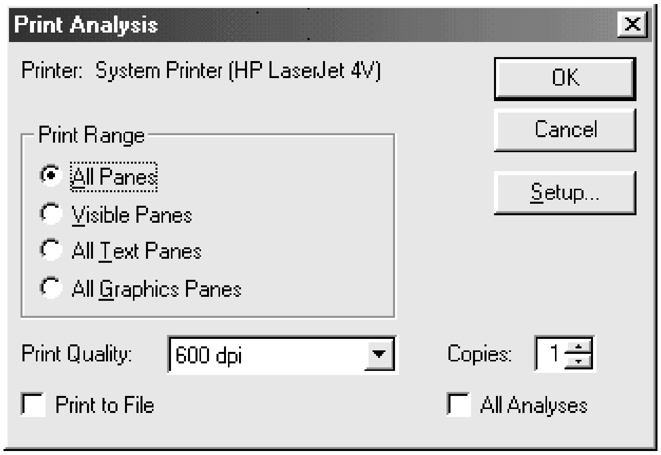
You can print a single maximized Analysis pane, all panes, visible panes, all text panes, or all the graphics panes.

To Print an Analysis Window

1. Choose the analysis you want to use from one of the five menus.

2. Complete the Analysis dialog box and click OK to display the Analysis Summary and the first graph in the text and graphics panes of the Analysis window.

3. Choose FILE... PRINT... from the Menu bar to display the Print Analysis dialog box (see Figure 6-3).



Printer: System Printer (HP LaserJet 4V)

Print Range
● All Panes
○ Visible Panes
○ All Text Panes
○ All Graphics Panes

Print Quality: 600 dpi

Copies: 1

□ Print to File

□ All Analyses

Figure 6-3. Print Analysis Dialog Box

4. Choose the print range, select the print quality, indicate the number of copies and if you want to print to a file. If you want to print all the analyses in the StatFolio, click that check box.

5. Click OK to print the window.

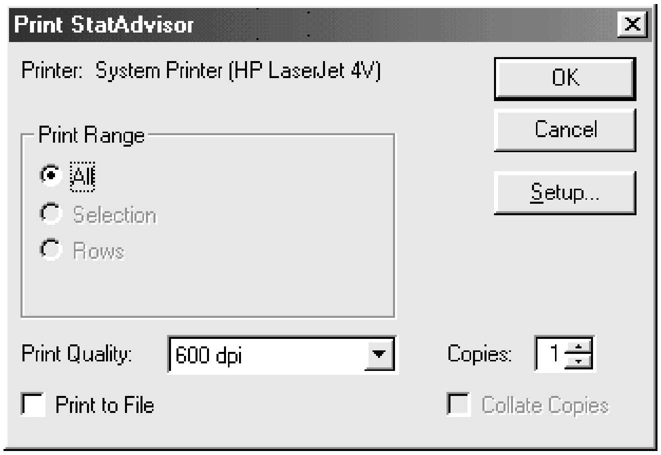
Printing a StatAdvisor Window

The StatAdvisor window contains information that helps you understand and interpret the results of a statistical analysis. You can print this information.

To Print a StatAdvisor Window

1. Open and click on the StatAdvisor window you want to print.

2. Choose FILE... PRINT... from the Menu bar to display the Print StatAdvisor dialog box (see Figure 6-4).



Print StatAdvisor

Printer: System Printer (HP LaserJet 4V)

Print Range

● All

○ Selection

○ Rows

Print Quality: 600 dpi

Copies: 1

□ Print to File

□ Collate Copies

Figure 6-4. Print StatAdvisor Dialog Box

3. Choose the print range, select the print quality, indicate the number of copies and if you want to print to a file. If your printer allows you to collate, indicate if you want the copies collated.

4. Click OK to print the StatAdvisor window.

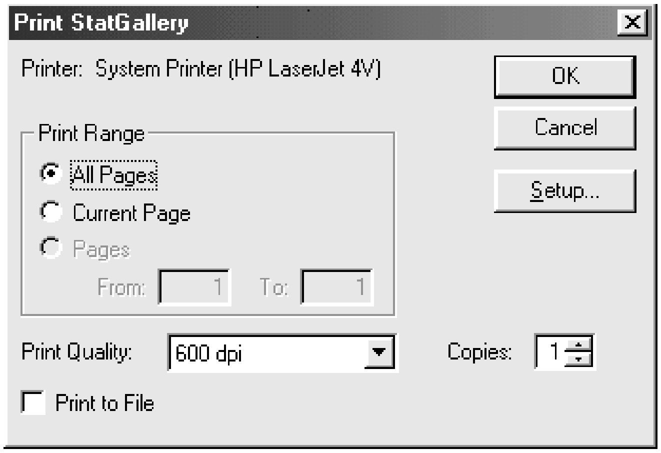
Printing a StatGallery Window

STATGRAPHICS *Plus* allows you to copy text and graphics to a StatGallery window so you can either view or print them. You can print the current page, all the pages, or a range of pages. *For information about using the StatGallery, see Chapter 7 in this manual.*

To Print a StatGallery Window

1. Complete, then click the StatGallery window you want to print.

2. Choose FILE... PRINT... from the Menu bar to display the Print StatGallery dialog box (see Figure 6-5).



Print StatGallery

Printer: System Printer (HP LaserJet 4V)

Print Range

● All Pages

○ Current Page

○ Pages

From: 1 To: 1

Print Quality: 600 dpi

Copies: 1

□ Print to File

Figure 6-5. Print StatGallery Dialog Box

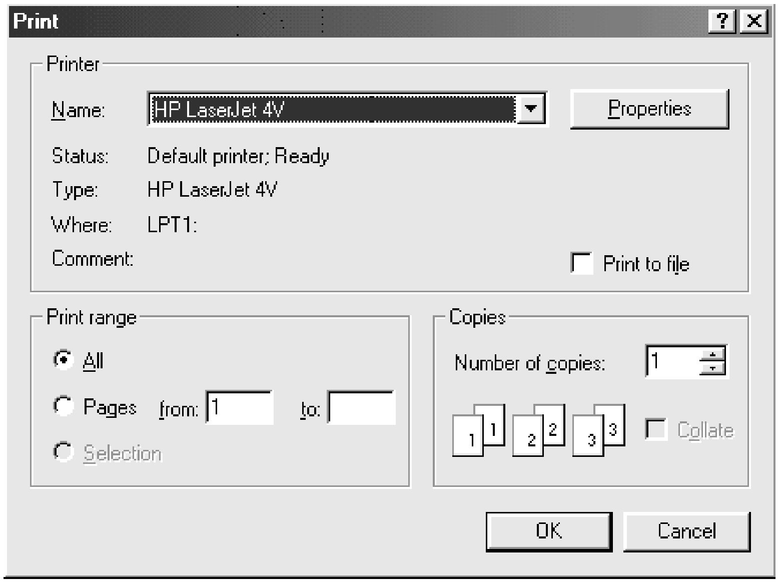
3. Choose the print range, select the print quality, indicate the number of copies, and if you want to print to a file.

4. Click OK to print the StatGallery.

Printing a StatReporter

1. Complete the StatReporter, then click the window you want to print.

2. Choose FILE... PRINT... from the Menu bar to display the Print dialog box (see Figure 6-6).



Print

Printer

Name: HP LaserJet 4V

Status: Default printer; Ready

Type: HP LaserJet 4V

Where: LPT1:

Comment:

□ Print to file

Print range

● All

○ Pages from: 1 to:

○ Selection

Copies

Number of copies: 1

□ Collate

Figure 6-6. Print StatReporter Dialog Box

3. Choose the print range, indicate the number of copies and if you want to print to a file. If your printer allows you to collate, indicate if you want the copies collated. This is a standard Windows Print dialog box.

4. Click OK to print the report.

Using the Print Preview... Command

The Print Preview... command on the File menu, allows you to view pages to see their format before you actually print them. Among the command's features are the capabilities to display multiple pages, to zoom in and out of the pages, and to print from the command. You can preview one pane at a time; for example, if the text and graphics options for an analysis are an Analysis Summary and a Scatterplot, you can preview only one of them at a time (see the steps below for an example of how you do this).

There are two methods you can use to access this command:

- Choose FILE... PRINT PREVIEW... from the Menu bar.
- Click the Print Preview... command on the pop-up menu that appears whenever you click the right mouse button on a pane.

To Use the Print Preview... Command

1. Access the analysis you want to use from one of the five menus.
2. Complete the Analysis dialog box, then click OK to display the Analysis Summary and the first graphical option.
3. Click the Tabular and/or Graphical Options buttons on the Analysis toolbar to select the other options you want to create.
4. Click OK to redisplay the Analysis window with the new options.
5. Choose FILE... PRINT PREVIEW... from the File menu on the Application toolbar to display the first page in the Print Preview window (see Figure 6-7).

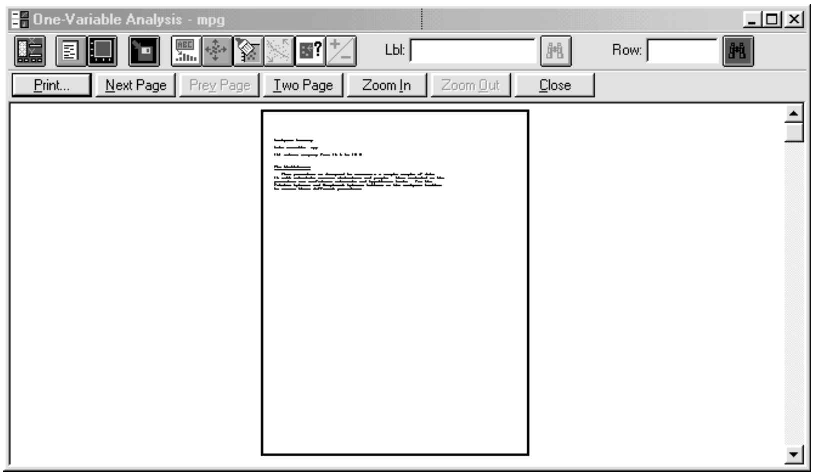


Figure 6-7. Print Preview Window

6. Click the Next Page button on the Preview window toolbar to preview the remaining pages, if there are any.

7. Click one of the other buttons on the Preview window toolbar to perform other actions.

8. Click either Close or Print on the Preview window toolbar when you have finished previewing the pages; the program will either close the Preview window and redisplay the Analysis window, or display the Print Analysis dialog box.

   The buttons on the Preview window help you navigate Print Preview.

   - **Print**
     This button closes the Preview window and displays a Print dialog box. The name of the dialog box will vary depending on the type of material you are previewing.

   - **Next Page**
     This button displays the next page of the material you are previewing if there is more than one page.

   - **Prev Page**
     This button displays the previous page of the material you are previewing if there is more than one page.

   - **Two Page**
     This button displays two side-by-side pages of preview material.

   - **Zoom In**
     This button moves the focus closer to the material on the page. You can click the Zoom button twice in one sequence.

   - **Zoom Out**
     This button returns the preview material closer to its original size.

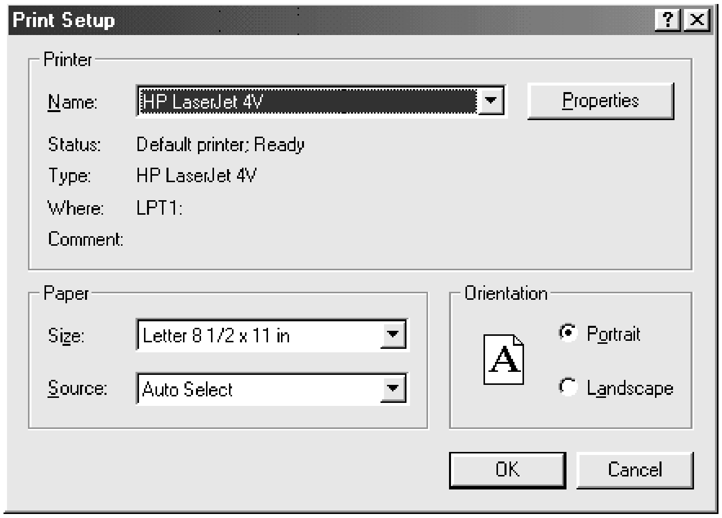
   - **Close**
     This button closes the preview material without printing it.

Using the [Print] Setup... Command

If you want to print to a printer other than to the default printer (the printer you designated as the default when you installed your Windows program), use the Print Setup... command from the File menu **or** the Setup... command on the Print [*n*] dialog box(es).

To Use the [Print] Setup... Command

1. Choose FILE... PRINT SETUP... from the Menu bar to display the Print Setup dialog box (see Figure 6-8).



Name: HP LaserJet 4V

Status: Default printer; Ready

Type: HP LaserJet 4V

Where: LPT1:

Comment:

Size: Letter 8 1/2 x 11 in

Source: Auto Select

● Portrait

○ Landscape

Figure 6-8. Print Dialog Box

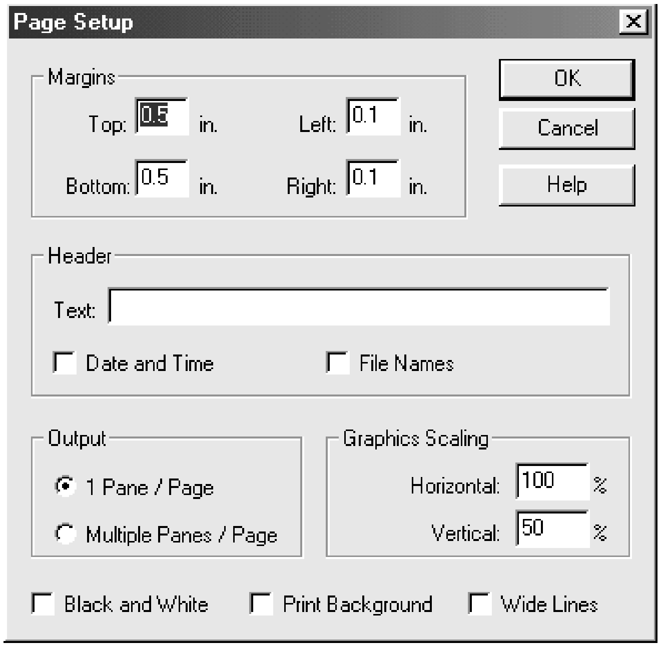
2. Choose the name of the printer you want to switch to, the paper size and source, and the print orientation, then click OK to save the settings.

Using the Page Setup... Command

STATGRAPHICS *Plus* uses default settings to define the print attributes, however, you can choose different values for the attributes. The changes you make remain in effect until you change them again. The settings are retained even when you start and end other sessions.

To Use the Page Setup... Command

1. Choose FILE... PAGE SETUP...from the Menu bar to display the Page Setup dialog box (see Figure 6-9).



Margins

Top: 0.5 in. Left: 0.1 in.

Bottom: 0.5 in. Right: 0.1 in.

Header

Text:

□ Date and Time □ File Names

Output

● 1 Pane / Page

○ Multiple Panes / Page

Graphics Scaling

Horizontal: 100 %

Vertical: 50 %

□ Black and White □ Print Background □ Wide Lines

Figure 6-9. Page Setup Dialog Box

2. Change the settings for the top, bottom, left, and right margins; enter the text that will be used for the header; indicate if the date, time, and file name should be part of the heading; choose the number of panes that will be printed per page; enter values for the horizontal and vertical graphics scaling; and indicate if the graphics should be converted to a gray scale, if the background should be printed, and if the line width should be doubled.
3. Click OK to save the settings.

Publishing

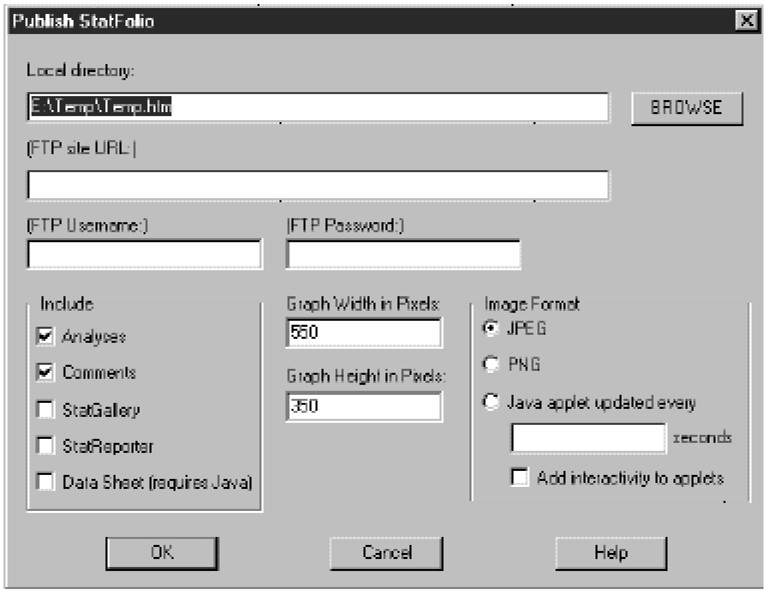
You can now publish StatFolios and view the published results. Access these two menu items using the File Menu. *StatPublish* lets you publish a StatFolio using a browser. *View Published Results* lets you view the published files using the default browser.

Publishing a StatFolio

To publish a StatFolio, complete the Publish StatFolio dialog box.

To Publish a StatFolio

1. Choose FILE...STATPUBLISH...from the Menu bar to display the Publish StatFolio dialog box (see Figure 6-10).


Publish StatFolio

Local directory: E:\Temp\Temp.htm

(FTP site URL:)

(FTP Username:)

(FTP Password:)

Include

☑ Analyses

☑ Comments

□ StatGallery

□ StatReporter

□ Data Sheet (requires Java)

Graph Width in Pixels: 550

Graph Height in Pixels: 350

Image Format

● JPEG

○ PNG

○ Java applet updated every ____ seconds

□ Add interactivity to applets

Figure 6-10. Publish StatFolio Dialog Box

2. Specify the Local directory on your hard disk where you want to save all html and image files.
3. Specify an optional site on a web server where the files will be moved after they are saved. An FTP server must be available on this site.
4. Specify an FTP Username and Password, if needed.
5. Select the items to be published. Choices include Analyses, Comments, StatGallery, StatReporter, and Data Sheet. Note that the Data Sheet is

published as a scrollable spreadsheet that requires the browser to have Java enabled in order to view it.

6. Specify the Width and Height of graphs (in pixels) when displayed on the web pages.

7. Specify the format for image files. The Image format may be either:

   - 24-bits per pixel JPEG files. This is a static image that does not change.

   - PNG files. This is a static image that does not change.

   - Java applets optionally updated periodically. If a duration k in seconds is specified, the image files are reloaded every k seconds. In addition, if you check the "Add interactivity to applets" box, users will be able to identify points on the published graphs by clicking on them using the mouse.

   Note that any use of Java is browser-dependent, particularly if web pages are printed. Currently, only Internet Explorer 5 does a reasonable job of printing applets.

8. Click OK.

   A dialog box appears notifying you that the StatFolio was pubished successfully.

9. Click OK.

   When a StatFolio is published, several files are created:

   - Table of Contents - an html file listing each item that has been published, with links.

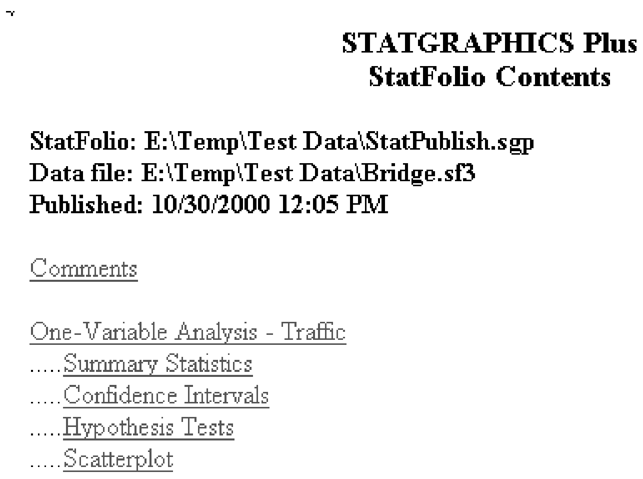
   - An html file for the Comments window, for each page of the StatGallery, for the StatReporter, and for each analysis window.

   - Image files for the Comments window, for each page of the StatGallery, for images in the StatReporter, for each pane in the analysis windows, and for the datasheet. File names are generated automatically and replace any existing files.

   The html files are meant to provide a quick way to view the contents of a StatFolio. You may imbed links to the image files into their own html documents. The file names will not change unless the StatFolio is modified by adding new analyses or panes within an analysis.

   A typical table of contents page appears below (see Figure 6-11):

View Published Results



STATGRAPHICS Plus
StatFolio Contents

StatFolio: E:\Temp\Test Data\StatPublish.sgp
Data file: E:\Temp\Test Data\Bridge.sf3
Published: 10/30/2000 12:05 PM

Comments

One-Variable Analysis - Traffic
.....Summary Statistics
.....Confidence Intervals
.....Hypothesis Tests
.....Scatterplot

Figure 6-11. Sample Table of Contents File

This menu option starts up the default browser and loads the table of contents for the published StatFolio.

To View Published Results

1. Choose FILE...VIEW PUBLISHED RESULTS...from the Menu bar to display the published results.

The results appear.

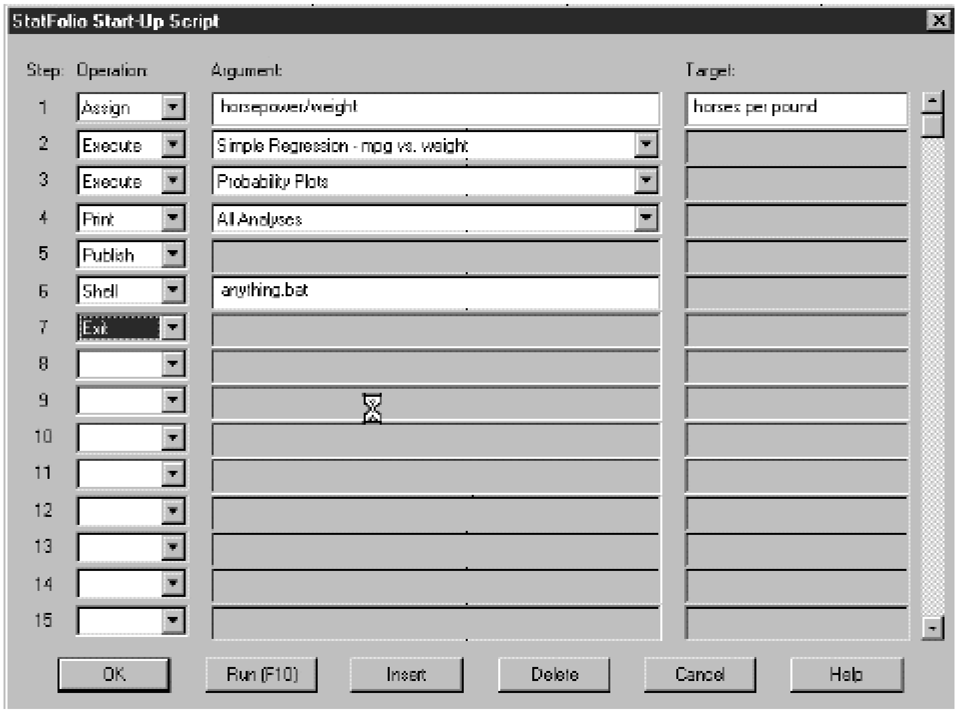
StatFolio Start-Up Scripts

The StatFolio Start-Up Script lets you define a script that will be run whenever the current StatFolio is loaded, either manually through the File menu or on program startup via the command line.

Using StatFolio Start-Up Scripts

To Use StatFolio Start-Up Scripts

1. Choose EDIT...STATFOLIO START-UP SCRIPT...from the Menu bar to display the StatFolio Start-Up Script dialog box (see Figure 6-12).



Step:	Operation:	Argument:	Target:
1	Assign	horsepower/weight	horses per pound
2	Execute	Simple Regression - mpg vs. weight	
3	Execute	Probability Plots	
4	Print	All Analyses	
5	Publish		
6	Shell	anything.bat	
7	Exit		
8			
9			
10			
11			
12			
13			
14			
15			

Figure 6-12. StatFolio Start-Up Script Dialog Box

2. Specify any of 6 Operations by clicking on the down arrow in the Operation box:

   Execute - runs the selected analysis.

   Assign - assigns the specified STATGRAPHICS expression to a target column of the data file, which may or may not exist. If it does, the contents of the column are replaced.

Print - prints the specified item, which may be “All Analyses”, “StatReporter”, “StatGallery”, “Data File”, “StatFolio Comments”, or a specified analysis. The current default print settings are used (no dialog box is displayed).

Publish - publishes the StatFolio to the last location used. An error message appears if the StatFolio has never been published.

Shell - instructs Windows to execute the specified file name.

Exit - causes STATGRAPHICS to terminate. This command is ignored except on initial loading of the program.

By properly configuring the script, you can make the program load a StatFolio, perform analyses, print or publish the results, and then exit.

For example, the script displayed in Figure 6-12 above could be executed by entering

```
sgwin.exe statfolio.sgp
```

in the Windows Run dialog box or by clicking on statfolio.sgp on the Windows Documents menu.

Note: A special switch allows any Exit operation in the script to be bypassed by typing

```
sgwin.exe /b statfolio.sgp
```

when loading STATGRAPHICS.

3. Click OK.

7 Using Special Features

This chapter explains six features that are unique to STATGRAPHICS *Plus*: the StatAdvisor, StatFolios, StatGallery, StatLink, StatReporter, and StatWizard. No other statistical software package offers work-saving tools such as these. They are all easy to use, yet powerful in the results they offer.

Special Features

- **StatAdvisor**
  This tool lets you display and print a short and easy-to-understand interpretation of the reports and graphics in a statistical analysis.

- **StatFolios**
  StatFolios are designed to eliminate the need for re-entering data and information you use on a regular basis.

- **StatGallery**
  This tool lets you copy an unlimited number of text panes and up to 100 graphics panes on multiple pages so you can view or print them.

- **StatLink**
  This tool provides the ability to tie a StatFolio directly to a data source so when you open the StatFolio, the data source is automatically queried and the analyses in the StatFolio are automatically updated.

- **StatReporter**
  An intermediate tool between a notepad and a “complete” word processor, this feature allows you to combine reports, created using tabular options, with graphics and your own notes in a report format that is available from within STATGRAPHICS *Plus.*

- **StatWizard**
  This new tool aids new or casual users by helping them select the correct analysis for their data using one of two options: selecting analyses from a standard menu or choosing an option from a Quick Pic list.

This chapter provides details about each of these features, and provides step-by-step instructions for using them.

Using the StatAdvisor

The StatAdvisor provides statistical advice that

- helps you interpret the results of an analysis
- highlights possible flaws and problem areas
- determines if the results are statistically significant
- adds credibility to reports.

The interpretation is a simple explanation of the data-sensitive results; that is, the interpretation varies according to the variables you use in an analysis. For example, when you use the StatAdvisor after you perform a Multiple Regression analysis, the interpretation indicates which variables appear to add significantly to the fitted model (see Figure 7-1).

You can access the StatAdvisor by either clicking the

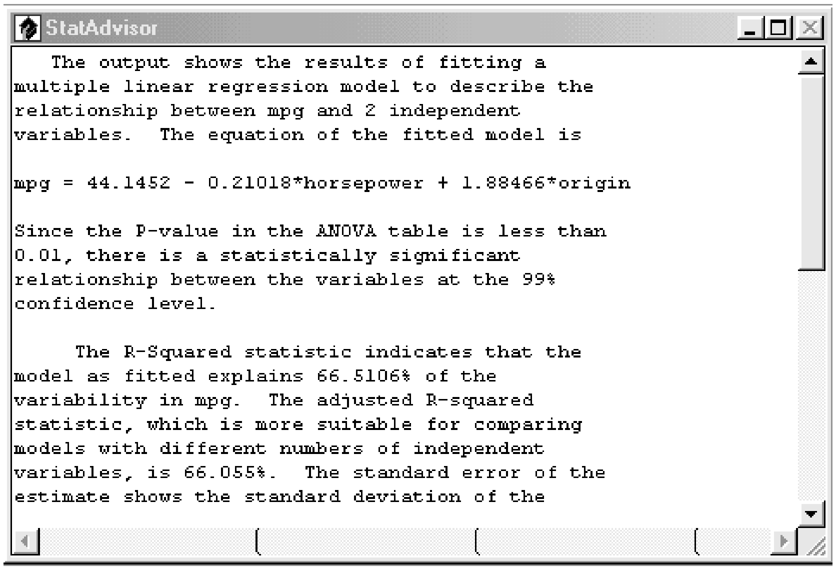
- StatAdvisor icon button on the Application toolbar, or
- StatAdvisor taskbar on the Application window.

Displaying the StatAdvisor in Text Panes

The View menu on the Application toolbar determines whether or not the StatAdvisor appears in the text panes.

To Display the StatAdvisor in Text Panes

1. Choose VIEW... from the Menu bar to display the drop-down menu, which has three choices: Toolbar, Status Bar, and StatAdvisor.



```
   The output shows the results of fitting a
multiple linear regression model to describe the
relationship between mpg and 2 independent
variables.  The equation of the fitted model is

mpg = 44.1452 - 0.21018*horsepower + 1.88466*origin

Since the P-value in the ANOVA table is less than
0.01, there is a statistically significant
relationship between the variables at the 99%
confidence level.

     The R-Squared statistic indicates that the
model as fitted explains 66.5106% of the
variability in mpg.  The adjusted R-squared
statistic, which is more suitable for comparing
models with different numbers of independent
variables, is 66.055%.  The standard error of the
estimate shows the standard deviation of the
```

Figure 7-1. StatAdvisor Interpretation for a Multiple Regression Analysis

2. Click the option you want to use; a check mark appears in front of your selection. In this case, click the StatAdvisor option.

The next analysis you run will display the StatAdvisor in the text pane.

Using the StatAdvisor Pop-Up Menu

The pop-up menu that displays when you click the right mouse button on a StatAdvisor pane contains three commands: Print..., Print Preview..., and Copy to Gallery.... The first two are self-explanatory; however, the latter is important because it lets you copy this interpretation into a StatGallery where it can add credibility to the results.

Using StatFolios

StatFolios eliminate the need for creating macros for repetitive tasks, for re-entering data, or for re-creating an analysis you will use another time. StatFolios let you perform the same analyses on numerous sets of data, or rerun analyses using new data or a new variable. You can save an entire set of analyses, which also saves the variables, graphics settings, and results from the various options. When you print the StatFolio, you can print one analysis at a time, or all of the analyses at once.

Using StatFolios greatly improves productivity because there is no need to repeat the same steps year after year. Instead, you can use them to automatically update analyses to reflect new data. When you save individual StatFolios with unique names, you can compare them yearly. And copying StatFolios into a StatGallery makes it easy to compare year-to-year data and to view graphics side-by-side.

To Create a StatFolio

1. Open and perform an analysis using the tabular and graphical options.
2. Save the analysis using FILE... SAVE AS... SAVE STATFOLIO AS..., which displays the Save StatFolio As... dialog box.
3. Type a name for the StatFolio in the File Name text box, and make any necessary changes, such as the directory and drive in which the file will reside.
4. Click Save to save the StatFolio.

Using the StatGallery

The StatGallery is a flexible tool you use primarily for archival purposes, while the StatReporter lets you “publish” reports that can include a StatGallery plus text you customize for the report. To access the StatGallery, click the StatGallery icon taskbar button in the Application window.

The StatGallery feature provides many benefits, including the ability to:

- Archive graphics for any given dataset; for example, you can make a text change in the Analysis Summary and copy the summary into a StatGallery; return to the DataEditor, change the data, refresh the Analysis Summary by minimizing it to a taskbar button, then restore it. The text pane will reflect the new data and the "custom" text will be gone.
- Copy analyses to the StatGallery and use the Save Graph... command from the File menu to save the StatGallery as a metafile.
- Perform the same analysis on the same data for a consecutive interval of time, save the analyses with unique names, then copy the same graph from the time intervals into the StatGallery and compare them side by side, perhaps on a monthly, quarterly, or yearly basis.

The StatGallery also lets you copy up to 100 graphics panes and an unlimited number of text panes (on multiple pages) to panes in the StatGallery window so you can view or print them.

A StatGallery window consists of nine panes, known as *splitter* windows, which are arranged in three columns. The panes are numbered in this order:

1	2	3
4	5	6
7	8	9

and grouped into columns in this order:

Column 1 = Panes 1, 4, 7

Column 2 = Panes 2, 5, 8

Column 3 = Panes 3, 6, 9

Four panes are usually visible, but using the Arrange Panes... command on the StatGallery pop-up menu allows you to use the StatGallery Options dialog box to change the arrangement of the panes and the columns. The bars separating the nine splitter windows allow you to change the pane height and width (see Figure 7-2).

The area under the StatGallery title bar displays the current page number and four page buttons: Next Page, Prev Page, First Page, and Last Page.

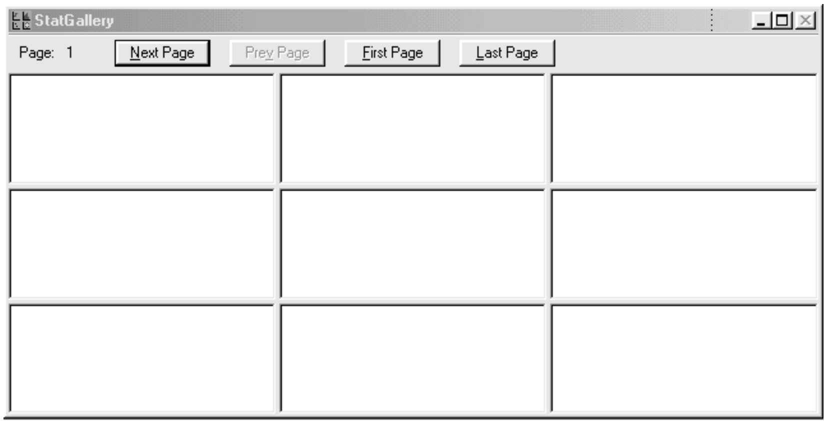


Figure 7-2. StatGallery Window with Nine Splitter Windows

- **Next Page**
  Adds another page to the StatGallery window.

- **Prev Page**
  Returns to the previous page of the StatGallery.

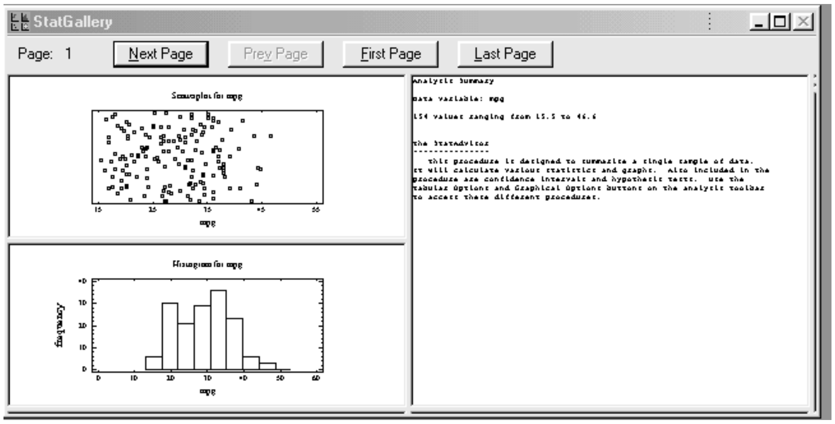
- **First Page**
  Returns to the first page of the StatGallery.

- **Last Page**
  Moves to the last page of the StatGallery.

The program copies the text and graphics panes you select to a secondary clipboard known as the *Gallery* clipboard, then allows you to place the panes in a variety of positions in the splitter windows within the StatGallery window. You can place items on the Gallery clipboard from within an analysis, within the StatAdvisor, or within a Comments window.

Each splitter window can contain:

- a single page of text from an analysis, the StatAdvisor, or the Comments window (see Figure 7-3).
- a single graph or multiple overlaid graphs from an analysis.



```
Analysis Summary

Data variable: mpg

154 values ranging from 15.5 to 46.6


The StatAdvisor
---------------
   This procedure is designed to summarize a single sample of data.
It will calculate various statistics and graphs.  Also included in the
procedure are confidence intervals and hypothesis tests.  Use the
Tabular Options and Graphical Options buttons on the analysis toolbar
to access these different procedures.
```

Figure 7-3. StatGallery Window with Different Items in Splitter Panes

Using the StatGallery Pop-Up Menu

This pop-up menu displays when you click the right mouse button on a pane in the StatGallery window.

- **Undo**
  Cancels the last action.

- **Cut**
  Removes the current contents of the pane and places it on the Gallery clipboard.

- **Copy**
  Copies individual StatGallery panes and places them on the Gallery clipboard or lets you paste them into the StatReporter or into other software applications.

- **Paste**
  Places the current contents of the Gallery clipboard into the selected pane. If you placed the item on the clipboard using the Cut... or Copy... commands, the item retains the properties it had before you placed it on the clipboard.

- **Paste Link**
  Maintains a link between the text or graph you place into the StatGallery and the analysis. If you make changes to the analysis, the linked text and the graphics pane are automatically updated.

- **Add Item**
  Opens a drawing toolbar that lets you to place lines, unfilled or filled rectangles and ellipses, or text on a graph (see Figure 7-4). Click one of the drawing buttons then click on the graph to use all but the last button. The Text Options dialog box displays when you click the last button (it looks somewhat like the letter A); it lets you enter the text you want to add. You must maximize the graphics pane before you can use the Add Item... command. **Note:** This option is available only for unlinked panes.

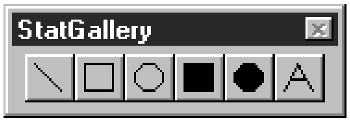


Figure 7-4. StatGallery Toolbar

- **Modify Item**
  Allows you to select an option that you want to change on an unlinked graph, then access a graphics Options dialog box; for example, if you want to change the layout, points, or point fills, the Layout Options, Point Options, or Fills Options dialog box would display. These dialog boxes are described in Online Help.

- **Delete Item**
  Deletes items you select in the StatGallery window; for example, lines, titles, legends, and added items.

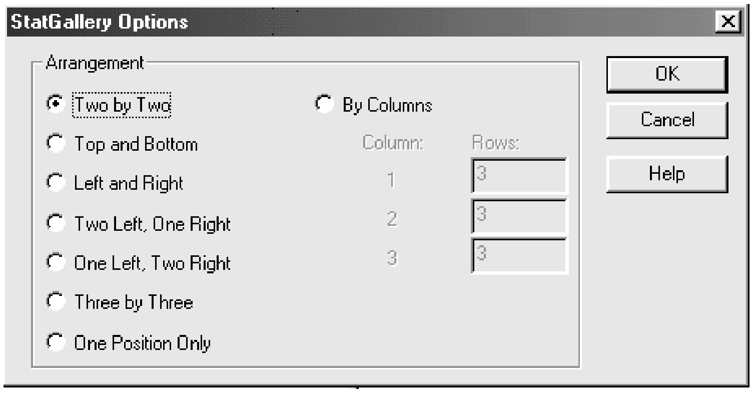
- **Unlink**
  Unlinks text or graphs you linked to an analysis using the Paste Link... command.

- **Clear Page**
  Clears the contents of the current page of the StatGallery window.

- **Delete Page**
  Deletes the current page from the StatGallery window.

- **Insert Page**
  Inserts a 2x2 page before the current page of the StatGallery window.

- **Arrange Panes**
  Displays the StatGallery Options dialog box, which allows you to select different arrangements for the panes on the current page of the StatGallery (see Figure 7-5).



StatGallery Options

Arrangement

● Two by Two
○ Top and Bottom
○ Left and Right
○ Two Left, One Right
○ One Left, Two Right
○ Three by Three
○ One Position Only

○ By Columns

Column:	Rows:
1	3
2	3
3	3

Figure 7-5. StatGallery Options Dialog Box

- **Clear Gallery**
  Erases the entire contents of the StatGallery.

- **Print**
  Displays the Print StatGallery dialog box, which allows you to select printing options. *See Chapter 6, Printing, Publishing and the StatFolio Start-Up Script, for information about this dialog box and printing in general.*

- **Print Preview**
  Displays the Print Preview window, which allows you to view pages to see their format before you actually print them. *See Chapter 6, Printing, Publishing and the StatFolio Start-Up Script, for information about Print Preview.*

- **Save StatGallery**
  Saves the current StatGallery.

- **Save StatGallery As**
  Saves the current StatGallery with a different name.

You can use the Copy to Gallery... command to place the text and graphs onto the Gallery clipboard, then use the Paste... or Paste Link... commands to copy the items into a splitter window.

Using the Paste Link... command automatically updates the StatGallery when you make changes within an analysis. The program checks the analysis to see if changes were made to the data, then automatically updates the data already there. For example, if you create a StatFolio one week and copy it into a StatGallery, then the next week make changes to that data, the program recalculates the analysis within the StatFolio, then automatically updates the StatGallery.

To Use the StatGallery

1. Access the text or graphics pane you want to move into a StatGallery window.

2. Click the left mouse button on the item you want to copy to the window, click the right to display the pop-up menu, then click the left on Copy Pane to Gallery... on the pop-up menu.

3. Click the left button on the StatGallery icon taskbar button, then click Restore... on the pop-up menu to display the StatGallery window.

4. Click the right button in one of the splitter windows to display the pop-up menu, then choose the task you want to perform.

To Modify the Arrangement of Panes in a StatGallery Window

1. Click the right mouse button on a StatGallery window.

2. Click the left on Arrange Panes... on the pop-up menu to display the StatGallery Options dialog box.

The dialog box lets you choose from seven pane positions, or from column and row positions. If you choose the By Column option, in the Row text boxes, enter the number of panes you want to appear in each of the three columns. Figure 7-6 shows a StatGallery window arranged into three columns with two panes in Column 1, one pane in Column 2, and three panes in Column 3.

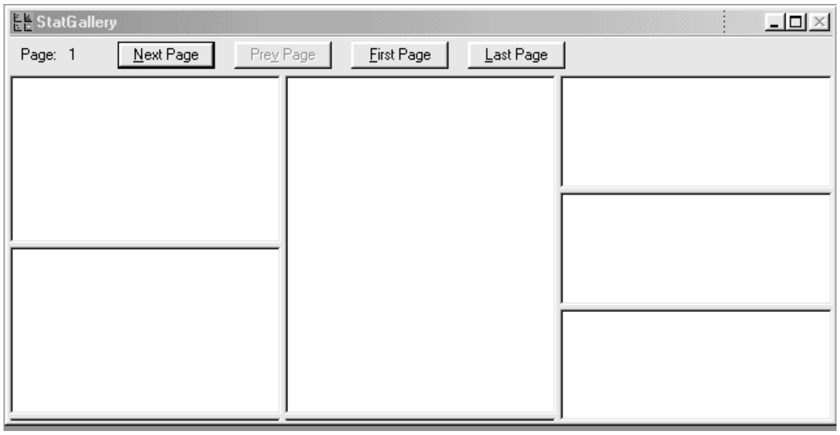


Figure 7-6. StatGallery Window Arranged into Three Columns

3. Choose one of the arrangement options, then click OK to rearrange the panes.

Overlaying Graphs

Overlaying one graph on top of another can be beneficial, especially in evaluating the results of an analysis. For the overlay to be helpful, use graphs that have approximately the same scaling.

To Overlay Graphs

1. Open and perform an analysis, choosing plots that will have approximately the same scaling.

2. Click the Copy Pane to Gallery... command from the pop-up menu.

3. Click the StatGallery taskbar button, then Restore... on the pop-up menu.

4. Click either the Paste or Paste Link command to copy the first graph into the StatGallery.

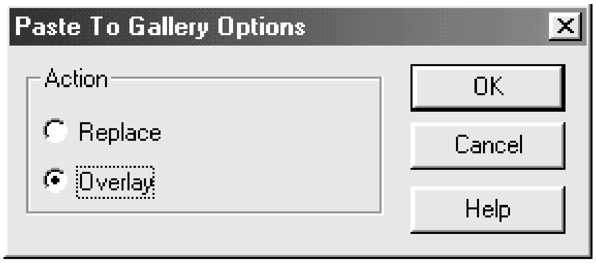
5. Minimize the StatGallery window.

6. Create the second plot.

7. Click the Copy Pane to Gallery... command on the pop-up menu.

8. Click the StatGallery taskbar button, then Restore... on the pop-up menu.

9. On the same pane that contains the first graph, click either the Paste or Paste Link command to copy the second graph into the StatGallery; the Paste to Gallery Options dialog box displays (see Figure 7-7).



Paste To Gallery Options

Action

○ Replace

● Overlay

Figure 7-7. Paste to Gallery Options Dialog Box

10. Click the Overlay option, then OK to overlay the second plot on top of the first.

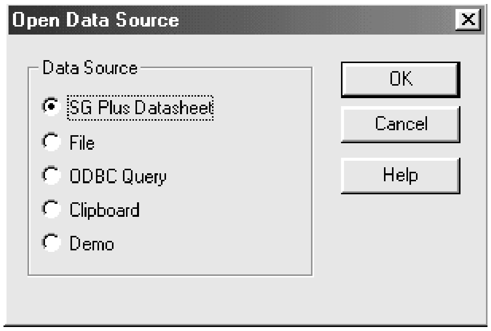
Using StatLink

StatLink is a key enhancement to STATGRAPHICS *Plus*. It provides the ability to tie a StatFolio directly to a data source so when a StatFolio is opened, an automatic query begins and an automatic update is made to the analyses in the StatFolio. In addition, the feature provides the capability to poll the data source automatically at regular intervals.

To use StatLink, from the menus, choose: FILE... STATLINK..., which displays a submenu of items: Change Data Source, Display Status, Start Polling, Stop Polling, and Update Now.

Using the Change Data Source Option

This option displays the Open Data Source dialog box (see Figure 7-8) that lets you choose the data source you want to use. You have five options.



Open Data Source

Data Source

● SG Plus Datasheet

○ File

○ ODBC Query

○ Clipboard

○ Demo

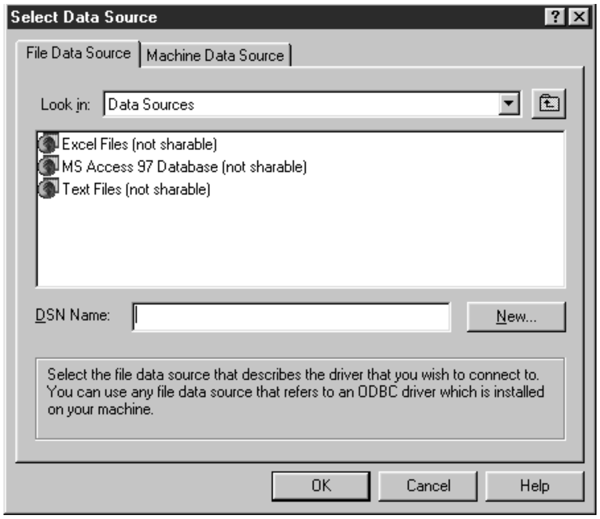
Figure 7-8.   Open Data Source Dialog Box

- **SG Plus DataSheet**
  Links data to a DataSheet. To check to see if the link was made, use StatLinks to display the Status message box.

- **File**
  Displays the Open Data File dialog box so you can select the file you want to use.

- **ODBC Query**
  Displays the Select Data Source dialog box (see Figure 7-9). Use the dialog box to select the data source that describes the driver you want to connect to. You can use any data source that refers to an ODBC driver that is installed on your computer.

- **Clipboard**
  Displays the Read Clipboard dialog box that lets you indicate if the variable names will be from the first row of data or if defaults should be created by the program (see Figure 7-10).



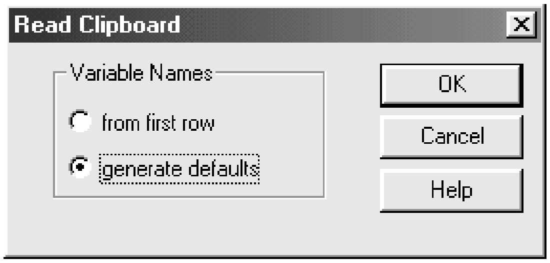
Look in: Data Sources

Excel Files (not sharable)
MS Access 97 Database (not sharable)
Text Files (not sharable)

DSN Name:

Select the file data source that describes the driver that you wish to connect to. You can use any file data source that refers to an ODBC driver which is installed on your machine.

Figure 7-9. Select Data Source Dialog Box



Variable Names

○ from first row

● generate defaults

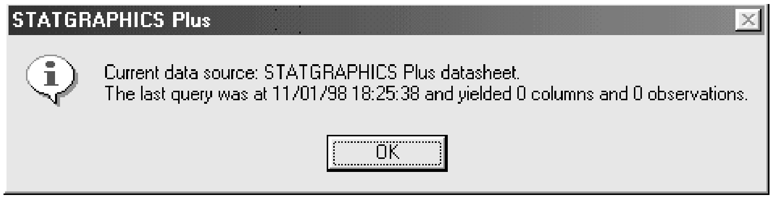
Figure 7-10. Read Clipboard Dialog Box

- **Demo**
  Generates a single column of random numbers.

After you make a selection, the program reads the data and locks the DataSheet to prevent other users from making changes.

Using the Display Status Menu Item

This item displays a message box that shows the last time the data was read and the number of rows and columns obtained from the data source (see Figure 7-11).



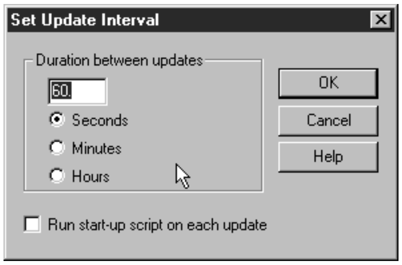
STATGRAPHICS Plus

Current data source: STATGRAPHICS Plus datasheet.
The last query was at 11/01/98 18:25:38 and yielded 0 columns and 0 observations.

Figure 7-11. Display Status Message Box

Using the Start and Stop Polling Menu Items

The first item displays the Set Update Interval dialog box, which lets you choose the duration of time between updates (see Figure 7-12). After the data are read, the program automatically updates the data in all the Analysis windows and StatGalleries to update their views.



Set Update Interval

Duration between updates

60

● Seconds
○ Minutes
○ Hours

□ Run start-up script on each update

Figure 7-12. Set Update Interval Dialog Box

The Stop Polling menu item immediately and automatically halts the polling.

Using the Update Now Menu Item

This item immediately rereads the data and updates all the analyses.

Using the StatReporter

The StatReporter feature lets you “publish” a customized report. Described as an “intermediate” tool between a notepad and a “complete” word processor, it lets you combine reports generated from tabular options, graphics, your own notes, and even interpretations from the StatAdvisor, into a report format. Your report can be further customized by using Cut... and Paste... to rearrange the text and graphs, changing the style of the text font, as well as its point size and color.

When you make changes to graphs in an analysis, the changes are automatically made in the StatReporter. You can also use the StatReporter to copy captions, annotate, copy items, and work back and forth between Word and STATGRAPHICS *Plus*.

To access the StatReporter, click the StatReporter taskbar on the Application window.

The StatReporter window has its own special toolbar (see Figure 7-13). The items on the toolbar let you select a font and change its point size; use bold, italic, underlining, or color; indicate if the text will be aligned left, right, or centered; use bullets; add the data and time; and find words or phrases in the document you are creating. The buttons follow standard Windows usage.

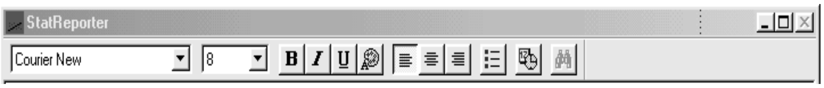


Figure 7-13. StatReporter Toolbar

When you click the right mouse button on the StatReporter window, the StatReporter pop-up menu displays (see Figure 7-14).

- **Undo**
  Cancels the last action.

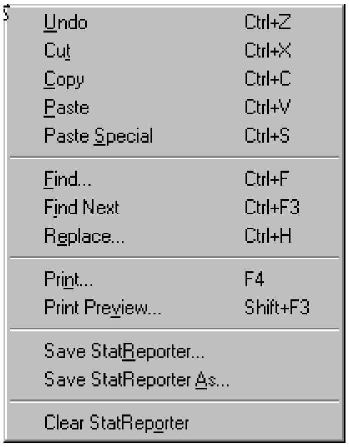


Figure 7-14. StatReporter Pop-Up Menu

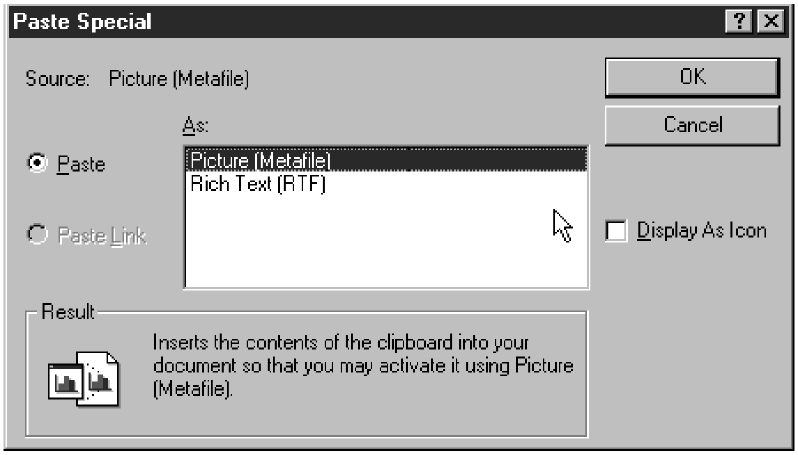
- **Cut**
  Removes the current contents of the pane and places it on the Gallery clipboard.

- **Copy**
  Copies an element in the StatReporter and pastes it into other applications.

- **Paste**
  Places the current contents of the Gallery clipboard into the StatReporter. If you placed the item on the clipboard using the Copy... command, the item retains the properties it had before you placed it on the clipboard.

- **Paste Special**
  Available only after you save a StatFolio. Lets you insert StatFolios, pictures (metafile format), or unformatted text using the Paste Special dialog box (see Figure 7-15).

- **Find**
  Displays the Find dialog box that lets you enter the word or phrase that you are trying to locate (see Figure 7-16). You can indicate if you want to match the whole word or if you want to match the case. Find is case-sensitive when you select Match Case; for example, to find StatAdvisor, be sure to specify Match Case and StatAdvisor.



Source: Picture (Metafile)

As:

● Paste

○ Paste Link

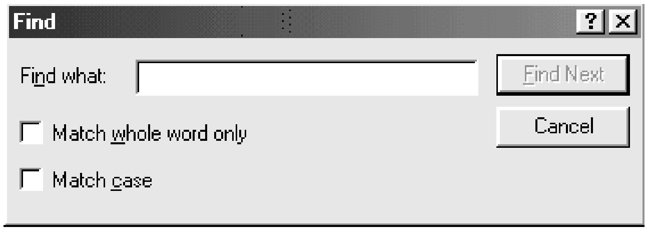
Picture (Metafile)
Rich Text (RTF)

□ Display As Icon

Result

Inserts the contents of the clipboard into your document so that you may activate it using Picture (Metafile).

Figure 7-15. Paste Special Dialog Box



Find what:

□ Match whole word only

□ Match case

Figure 7-16. Find Dialog Box

- **Find Next**
  Finds the next occurrence of the last search. The behavior is the same as that for the Find dialog box.

- **Replace**
  Displays the Replace dialog box that lets you enter the word or phrase you want to find and the word or phrase that it will be replaced with (see Figure 7-17). You can indicate if you want to match the whole word or if you want to match the case.

- **Print**
  Displays the Print dialog box that lets you choose the print range, indicate the number of copies and, if you want to, print to a file. If your printer

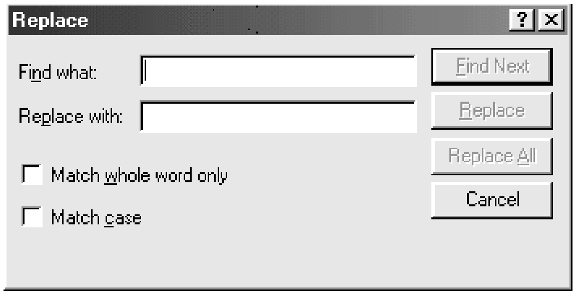


Figure 7-17. Replace Dialog Box

allows you to collate, indicate if you want the copies collated. This is a standard Windows Print dialog box. *For more information, see Chapter 6, Printing, Publishing and the StatFolio Start-Up Script.*

- **Print Preview**
  Displays the Print Preview window that shows the first page in the Print Preview window. *For more information, see Chapter 6, Printing, Publishing and the StatFolio Start-Up Script.*

- **Save StatReporter**
  Saves the current document.

- **Save StatReporter As**
  Saves the current document using a different name.

- **Clear StatReporter**
  Clears the StatReporter window.

To Use the StatReporter

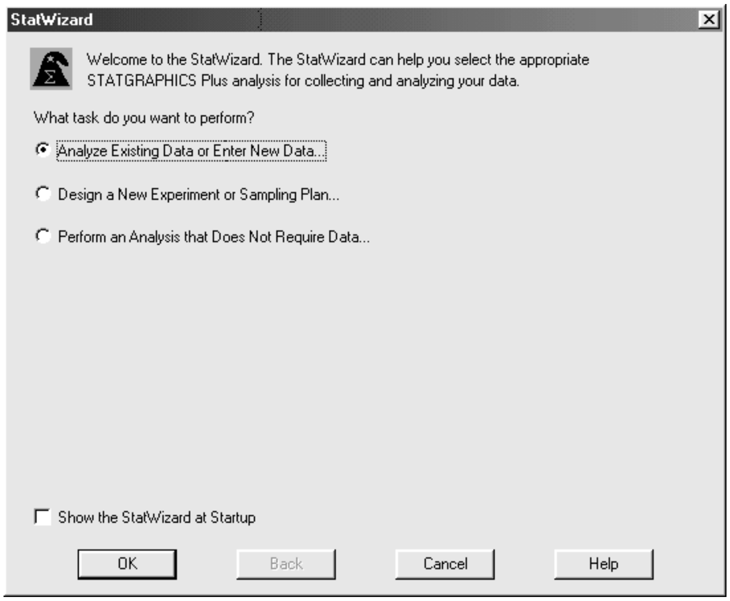
1. Access the text or graphics pane you want to move into a StatReporter window.
2. Click the right mouse button, then choose Copy Analysis to StatReporter... from the StatReporter pop-up menu.
3. Click the left button on the StatReporter icon taskbar button, then click Restore... on the pop-up menu to display the StatReporter with the text and graphics panes from the analysis copied into the window.

To Copy One Pane to the StatReporter

1. Access the text or graphics pane you want to move into a StatReporter window.
2. Click the right mouse button, then choose Copy... from the pop-up menu.
3. Click the left button on the StatReporter icon taskbar button, then click Restore... on the pop-up menu to display the StatReporter window.
4. Click the right button, choose Paste... or Paste Special... from the pop-up menu. The analysis pane is copied into the StatReporter.

Using the StatWizard

The StatWizard feature is especially useful to novice or new users who need help matching data with analyses. Figure 7-18 shows the StatWizard at Startup dialog box. The default behavior is that the StatWizard opens every time you open STATGRAPHICS *Plus*. To turn it off, click the Show the StatWizard at Startup checkbox on the initial dialog box. If you turn the StatWizard off, the dialog box shown in Figure 7-18 displays when you click the StatWizard icon on the Application toolbar when there is no data in the datasheet. If you have entered or loaded data into the datasheet, the dialog box shown in Figure 7-19 displays. This dialog box lets you choose analyses from a standard menu, a Quick Pic list, or a SnapStats list. Depending on the task you choose, a different series of dialog boxes will display.



StatWizard

Welcome to the StatWizard. The StatWizard can help you select the appropriate STATGRAPHICS Plus analysis for collecting and analyzing your data.

What task do you want to perform?

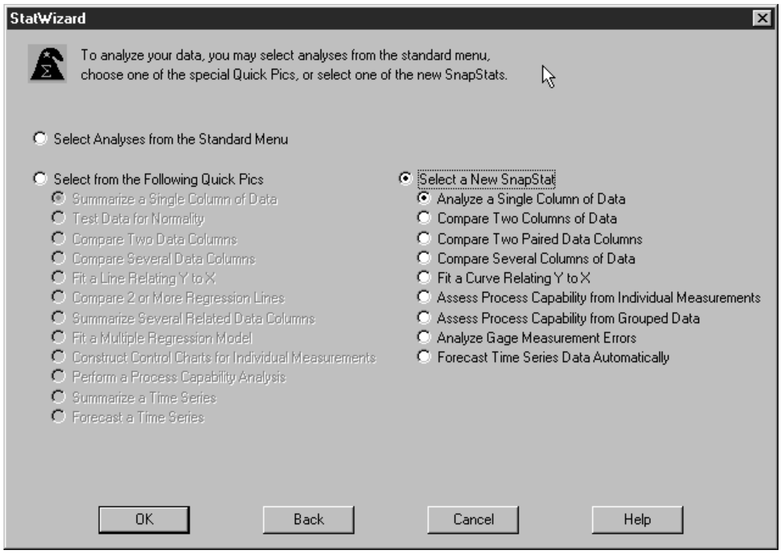
● Analyze Existing Data or Enter New Data...

○ Design a New Experiment or Sampling Plan...

○ Perform an Analysis that Does Not Require Data...

□ Show the StatWizard at Startup

Figure 7-18. StatWizard at Startup Dialog Box



StatWizard

To analyze your data, you may select analyses from the standard menu, choose one of the special Quick Pics, or select one of the new SnapStats.

○ Select Analyses from the Standard Menu

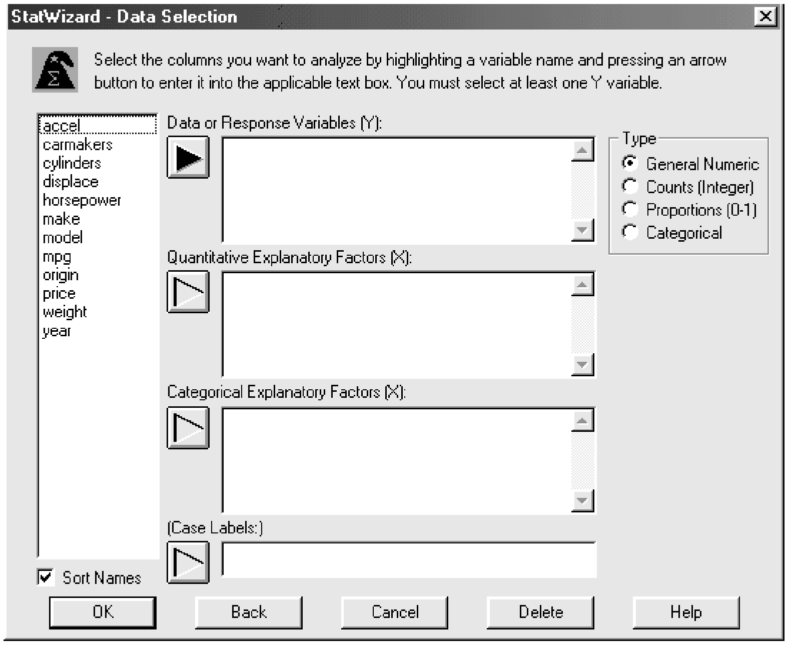
○ Select from the Following Quick Pics
- ● Summarize a Single Column of Data
- ○ Test Data for Normality
- ○ Compare Two Data Columns
- ○ Compare Several Data Columns
- ○ Fit a Line Relating Y to X
- ○ Compare 2 or More Regression Lines
- ○ Summarize Several Related Data Columns
- ○ Fit a Multiple Regression Model
- ○ Construct Control Charts for Individual Measurements
- ○ Perform a Process Capability Analysis
- ○ Summarize a Time Series
- ○ Forecast a Time Series

● Select a New SnapStat
- ● Analyze a Single Column of Data
- ○ Compare Two Columns of Data
- ○ Compare Two Paired Data Columns
- ○ Compare Several Columns of Data
- ○ Fit a Curve Relating Y to X
- ○ Assess Process Capability from Individual Measurements
- ○ Assess Process Capability from Grouped Data
- ○ Analyze Gage Measurement Errors
- ○ Forecast Time Series Data Automatically

Figure 7-19. StatWizard that Displays after Clicking the StatWizard Button while There are Data in a DataSheet

Selecting Analyses from the Standard Menu

When you use this option, you complete a series of dialog boxes that describe the data you are using. Because the dialog boxes are data-driven, they are too numerous to document separately, but four that are typical are shown here (see Figures 7-20, 7-21, 7-22, and 7-23).


StatWizard - Data Selection

Select the columns you want to analyze by highlighting a variable name and pressing an arrow button to enter it into the applicable text box. You must select at least one Y variable.

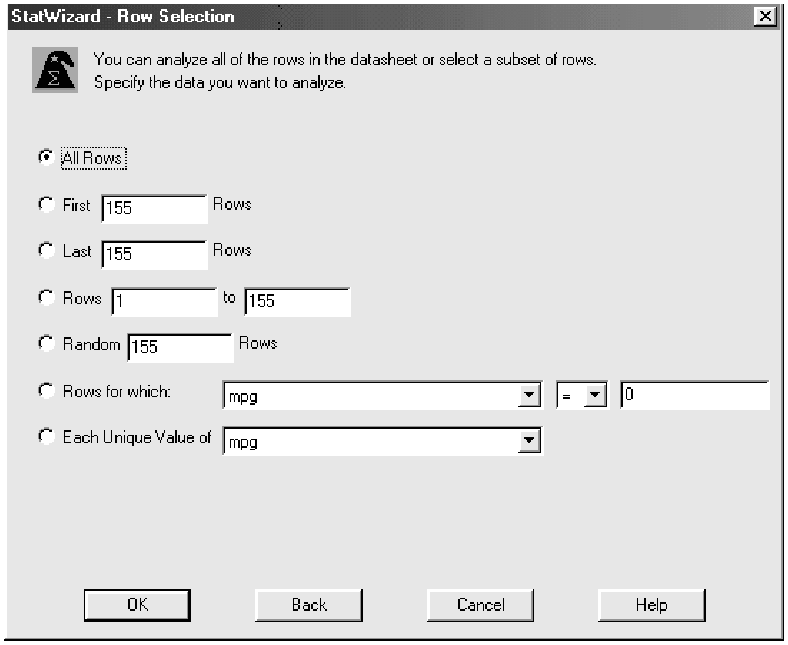
Type
● General Numeric
○ Counts (Integer)
○ Proportions (0-1)
○ Categorical

☑ Sort Names

Figure 7-20.   StatWizard - Data Selection Dialog Box

The StatWizard - Data Selection dialog box in Figure 7-20 lists the names of the variables in the file you have chosen. You are asked to enter a variable name for the Response Variable, Quantitative Explanatory Factor, the Categorical Explanatory Factor, and the Labels. You also choose the type of variable you are using — general numeric, counts (integers), proportions (0-1), or categorical.

You use the StatWizard - Row Selection dialog box when you want to analyze all the rows in the DataSheet or when you want to select a subset of the rows (see Figure 7-21).



StatWizard - Row Selection

You can analyze all of the rows in the datasheet or select a subset of rows.
Specify the data you want to analyze.

● All Rows

○ First 155 Rows

○ Last 155 Rows

○ Rows 1 to 155

○ Random 155 Rows

○ Rows for which: mpg = 0

○ Each Unique Value of mpg

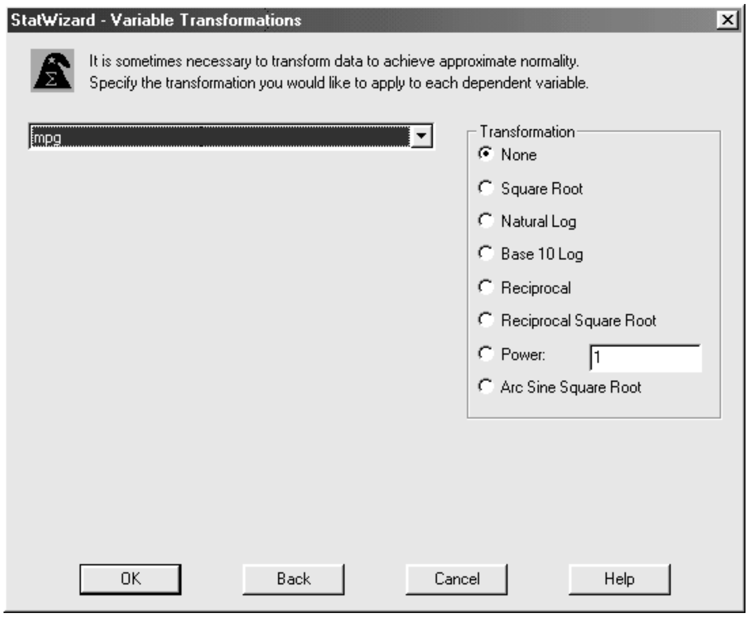
Figure 7-21. StatWizard - Row-Selection Dialog Box

The StatWizard - Variable Transformations dialog box is used when it is necessary to transform data to achieve an approximate normality (see Figure 7-22). You first select the name of the variable you want to transform, then choose the method that will be used to perform the transformation.

Figure 7-23 shows the StatWizard - Analysis Selection dialog box, which states the type of variable you have selected and asks you to select the analyses you want to perform.

Selecting Analyses from the Quick Pic List

The Quick Pic list contains 12 tasks, such as “Summarize a Single Column of Data,” “Compare Two Data Columns,” and “Fit a Multiple Regression Model.” When you choose a task, then click OK, the appropriate Analysis dialog box displays. For example, when you choose, “Summarize a Single Column of Data,” the One-Variable Analysis dialog box displays.



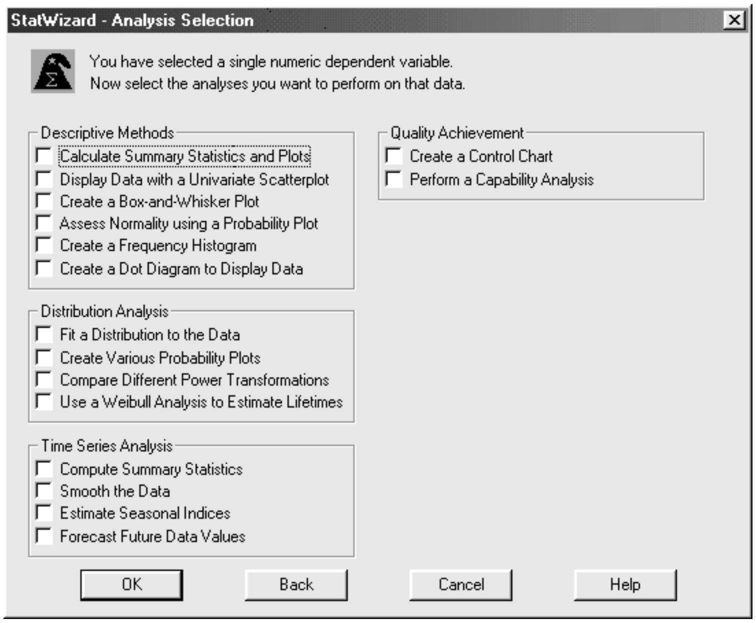
StatWizard - Variable Transformations

It is sometimes necessary to transform data to achieve approximate normality.
Specify the transformation you would like to apply to each dependent variable.

mpg

Transformation
- ● None
- ○ Square Root
- ○ Natural Log
- ○ Base 10 Log
- ○ Reciprocal
- ○ Reciprocal Square Root
- ○ Power: 1
- ○ Arc Sine Square Root

Figure 7-22. StatWizard - Variable Transformations Dialog Box



StatWizard - Analysis Selection

You have selected a single numeric dependent variable.
Now select the analyses you want to perform on that data.

Descriptive Methods
- □ Calculate Summary Statistics and Plots
- □ Display Data with a Univariate Scatterplot
- □ Create a Box-and-Whisker Plot
- □ Assess Normality using a Probability Plot
- □ Create a Frequency Histogram
- □ Create a Dot Diagram to Display Data

Distribution Analysis
- □ Fit a Distribution to the Data
- □ Create Various Probability Plots
- □ Compare Different Power Transformations
- □ Use a Weibull Analysis to Estimate Lifetimes

Time Series Analysis
- □ Compute Summary Statistics
- □ Smooth the Data
- □ Estimate Seasonal Indices
- □ Forecast Future Data Values

Quality Achievement
- □ Create a Control Chart
- □ Perform a Capability Analysis

Figure 7-23. StatWizard - Analysis Selection Dialog Box

Selecting Analyses from the SnapStats List

The SnapStats list contains 9 tasks:

- Analyze a Single Column of Data
- Compare Two Columns of Data
- Compare Two Paired Data Columns
- Compare Several Columns of Data
- Fit a Curve Relating Y to X
- Assess Process Capability from Individual Measurements
- Assess Process Capability from Grouped Data
- Analyze Gage Measurement Errors
- Forecast Time Series Data Automatically

When you choose a task, then click OK, the appropriate Analysis dialog box displays. For example, when you choose, “Analyze a Single Column of Data,” the One Sample Analysis dialog box displays.

8 Using Basic Plots

Introduction to Basic Plots

Important Note: Descriptions of the contents of the dialog boxes now appear ***only*** in Online Help instead of in both the manual and Online Help.

This chapter contains information about the basic plot analyses found in the Plot menu. The chapter presents a brief overview for each of the analyses, describes its tabular and graphical options, and provides an interpretation of each option.

Six different types of plots are available: Scatterplots, Exploratory Plots, Business Charts, Probability Distributions, Response Surfaces Plots, and Custom Charts.

- **Scatterplots**
  Scatterplots are useful for investigating relationships among variables because they allow you to easily see the range of the data (density and shape). The observations for one or more variables are plotted as point symbols along one or more axes, which makes it possible to visually depict the attributes of the data. You can create six types of scatterplots: Univariate Plot, X-Y Plot, X-Y-Z Plot, Multiple X-Y Plot, Multiple X-Y-Z Plot, and Polar Coordinates Plot.

- **Exploratory Plots**
  Exploratory plots are useful for studying symmetry, checking distributional assumptions, and detecting outliers. You can create eight types of exploratory plots: Box-and-Whisker Plot, Multiple Box-and-Whisker Plot, Probability Plot, Frequency Histogram, Dot Diagram, Multiple Dot Diagram, Bubble Chart, and Radar/Spider Plot.

- **Business Charts**
  Business charts are useful for presenting relative quantities of data in an easily understood visual format. These types of charts allow you to compare the size of each group or class of data. You can create five different business charts: Barchart, Multiple Barchart, Piechart, Component Line Chart, and High-Low-Close Plot.

- **Probability Distributions**
  There are twenty-four different distribution choices you can use to generate and save random numbers, calculate tail areas and critical values, and view the probability distribution using a variety of distribution plots. The 24 distributions are: Bernoulli, Binomial, Discrete Uniform, Geometric, Hypergeometric, Negative Binomial, Poisson, Beta, Cauchy, Chi-Square, Erlang, Exponential, Extreme Value, F (Variance Ratio), Gamma, Laplace, Logistic, Lognormal, Normal, Pareto, Student's *t*, Triangular, Uniform, and Weibull.

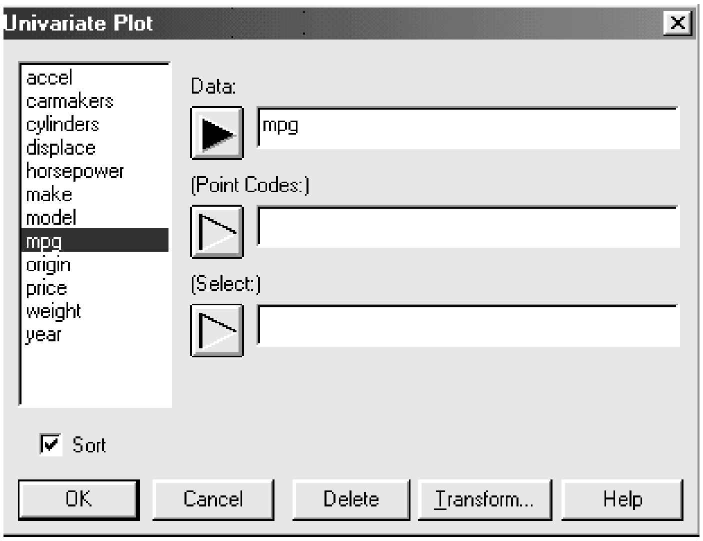
- **Response Surface Plots**
  There are two types of plots: Response Surface and Contour. Both types of plots are helpful in determining the influence of factors on the response variable and in locating optimal regions. Response Surface plots allow you to visualize the relationship among the factors and the response variable. Surface plots depict solid surfaces in three-dimensional space, while Contour plots depict solid surfaces in two-dimensional space with contours that correspond to a particular height of a response surface.

- **Custom Charts**
  Custom Charts are useful when circumstances require that you use a customized chart. For example, when ensuring that everyone who collects experimental data, charts the data the same way.

Using the Univariate Plot Analysis

The Univariate Plot Analysis lets you create a scatterplot for a single numeric variable. It applies to a method of grouping data that involves one characteristic of the members of a population or sample, otherwise known as univariate data. For example, you might examine the age of welfare recipients but not examine their gender or ethnicity. You can enter point codes to show the levels of the classification factor.

To access this analysis, from the menus, choose: PLOT... SCATTERPLOTS... UNIVARIATE PLOT... to display the Univariate Plot Analysis dialog box (see Figure 8-1).



Data: mpg

(Point Codes:)

(Select:)

☑ Sort

Figure 8-1. An Example of An Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates the results of an analysis for a single column of data. The summary includes the name of the selected variable, the number of values in the variable, and the range of the values.

Graphical Options

Scatterplot

The Scatterplot option creates a plot that is useful for identifying individual data values that represent regions of highest density, and for determining if outliers are present (see Figure 8-2). A horizontal axis covers the range of the data, while the vertical axis has no real meaning unless it is *jittered (see below)*. The vertical jittering for this analysis is set, by default, to the maximum.

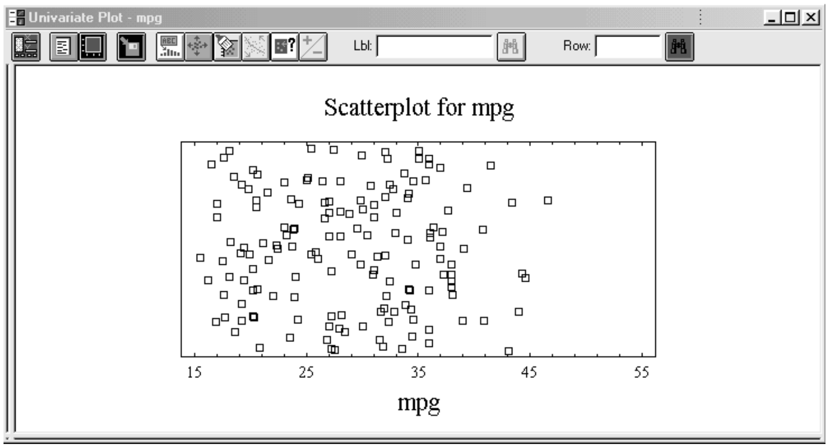


Figure 8-2. Scatterplot

Use this plot when you want to investigate the distribution and range of observations in a variable. Although you can easily identify the range of the data, it may be difficult to identify individual points if they overlap. To reduce the amount of overlapping, you can *jitter* the points using the Jittering button on the Analysis toolbar.

Time Sequence Plot

The Time Sequence Plot option creates a plot that is useful for detecting trends that occur over time, or for determining patterns in the data. Use it only when the data are collected and stored in time order; the observations are plotted in the order in which they were entered into the DataSheet (see Figure 8-3).

Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot. To clearly see the pattern in the plot, choose a smoothing option from the Scatterplot Smoothing Options dialog box, which is accessed by clicking the Smooth/Rotate button on the Analysis toolbar.

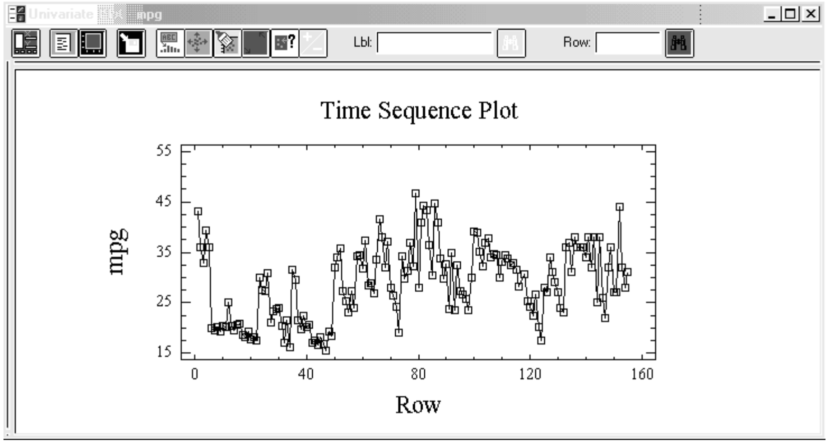


Figure 8-3. Time Sequence Plot

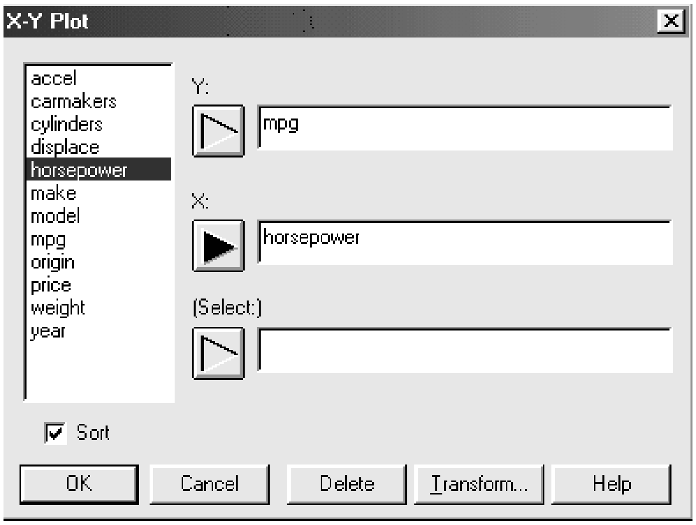
Using the X-Y Plot Analysis

The X-Y Plot Analysis creates a two-dimensional scatterplot for one variable versus another, which you use to examine their relationships. You can produce lineplots, scatterplots, connected scatterplots, coded scatterplots, and plots with standard error bars.

An X-Y lineplot displays connecting lines without points, while an X-Y scatterplot shows points only (no connecting lines). A connected X-Y scatterplot connects lines with points; a coded X-Y scatterplot is more informative than the basic X-Y scatterplot because it uses coded points to show the levels of a classification factor. You can designate the point codes.

When the points in a scatterplot represent means rather than individual values, you may find it useful to illustrate the uncertainty surrounding the points. You can plot standard error bars for the X- or Y- variables, or both. Standard error bars are lines that extend from the means, plus and minus one standard error.

To access this analysis, from the menus, choose: PLOT... SCATTERPLOTS... X-Y PLOT... to display the X-Y Plot Analysis dialog box (see Figure 8-4).



Y: mpg

X: horsepower

(Select:)

Figure 8-4. The X-Y Plot Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the name of the selected variable, and the total number of paired values.

Graphical Options

Scatterplot

The Scatterplot option creates a scatterplot of one variable versus a second (see Figure 8-5). The pattern of the points indicates the strength and direction of the correlation between the two values.

The following guidelines are helpful when you interpret a two-dimensional scatterplot:

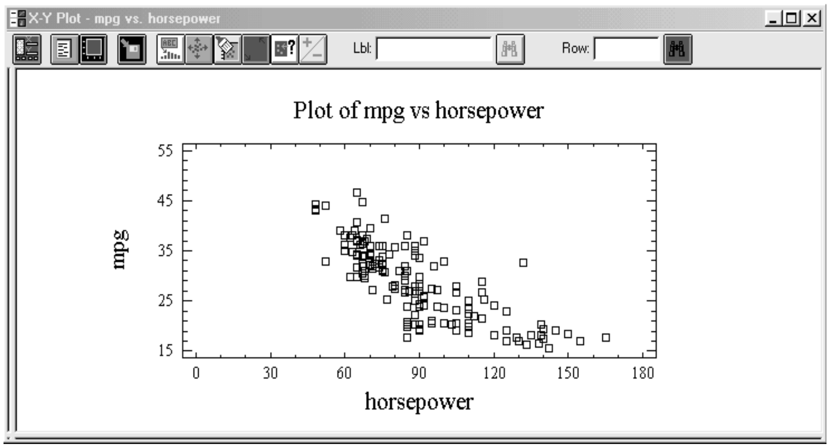


Figure 8-5. Scatterplot

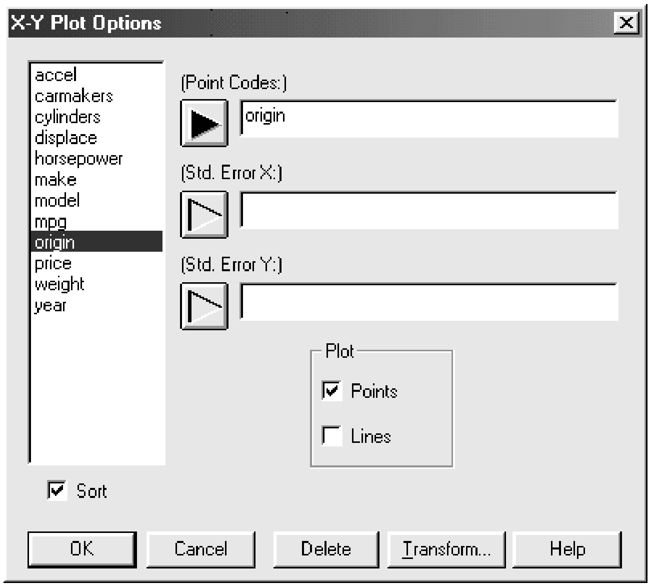
- the more the points tend to cluster around a straight line, the stronger the linear relationship between two variables (the higher the correlation)
- if the line around which the points tend to cluster runs from lower left to upper right, the relationship between the two variables is positive
- if the line around which the points tend to cluster runs from upper left to lower right, the relationship between the two variables is negative
- if there exists a random scatter of points, there is no relationship between the two variables (very low or zero correlation).

Use the *X-Y Plot Options* dialog box to customize the scatterplot (see Figure 8-6). You can enter a code for the points; enter values for the standard error of X and Y; indicate if points or lines will appear on the plot; and indicate if you want to use sorted variables.

To clearly see the pattern in the plot, choose a smoothing option from the Scatterplot Smoothing Options dialog box, which is accessed by clicking the Smooth/Rotate button on the Analysis toolbar.

References

Box, G. E. P. and Jenkins, G. M. 1976. *Time Series Analysis, Forecasting and Control*, second edition. San Francisco: Holden-Day.



X-Y Plot Options

accel
carmakers
cylinders
displace
horsepower
make
model
mpg
origin
price
weight
year

(Point Codes:) origin

(Std. Error X:)

(Std. Error Y:)

Plot
☑ Points
□ Lines

☑ Sort

Figure 8-6. The X-Y Plot Options Dialog Box

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. The Wadsworth Statistics/Probability Series, edited by P. J. Bickel, W. S. Cleveland, and R. M. Dudley. Co-publishing project of Wadsworth International Group and Duxbury Press, divisions of Wadsworth, Inc. Belmont, CA: Wadsworth International Group.

Cleveland, W. S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of American Statistical Association*, **74**:829-836.

Cleveland, W. S. 1981. "Lowess: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression, *The American Statistician*, **35**:54.

Lapin, L. L. 1987. *Statistics for Modern Business Decisions.* Orlando, Florida: Harcourt Brace Jovanovich, Inc.

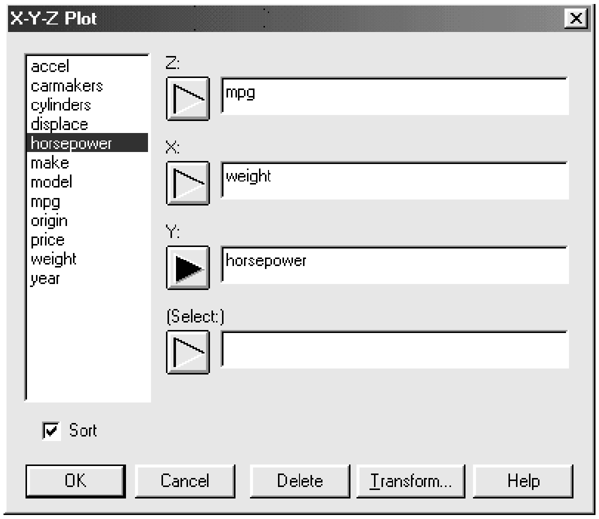
Using the X-Y-Z Plot Analysis

The X-Y-Z Plot Analysis plots three variables instead of two and produces three-dimensional graphs with the X-axis plotted horizontally along the bottom of the screen, the Y-axis extending back into the screen, and the Z-axis aligned vertically. The plot provides options for lineplots, scatterplots, connected scatterplots, and coded scatterplots. You can also plot standard error bars.

An X-Y-Z lineplot displays connecting lines without points, while an X-Y-Z scatterplot shows points only (no connecting lines). A connected X-Y-Z scatterplot connects points with lines. A coded X-Y-Z scatterplot is more informative than the basic X-Y-Z scatterplot because it uses coded points to show the levels for a classification factor. You can designate the point codes.

Because large numbers of points and lines can blend together and make it difficult to read the graph, use this analysis only with small datasets. You can plot more values if you use a high-resolution graphics adapter.

To access the analysis, from the menus, choose: PLOT... SCATTERPLOTS... X-Y-Z PLOT... to display the X-Y-Z Plot Analysis dialog box (see Figure 8-7).



X-Y-Z Plot

Z: mpg

X: weight

Y: horsepower

(Select:)

☑ Sort

Figure 8-7.   The X-Y-Z Plot Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the names of the selected variables, and the total number of grouped values.

Graphical Options

X-Y-Z Scatterplot

The X-Y-Z Scatterplot option creates a three-dimensional scatterplot of the data with only the points plotted (see Figure 8-8). The pattern of the points indicates the strength and direction of the correlation among the values. The more the points tend to cluster around the straight line, the stronger the relationship (the higher the correlation). *See the description of the X-Y Scatterplot for guidelines on interpreting this type of plot.*

Use the *X-Y-Z Plot Options* dialog box to customize the plot (see Figure 8-9).

You can enter a code for the points; enter values for the standard error of X, Y, and Z; indicate if points or lines will appear on the plot; and indicate if reference lines will appear on the plot and, if so, where they will be drawn. You can also indicate if the variables should be sorted.

You can use the Smooth/Rotate button on the Analysis toolbar to change the angle from which you view this plot.

Draftsman's Plot

The Draftsman's Plot option creates a series of two-variable plots for all combinations of the three selected variables (see Figure 8-10). The plot shows the top, front, and side views of the data. Use the plot to detect positive, negative, or zero correlations; outliers; and curvature.

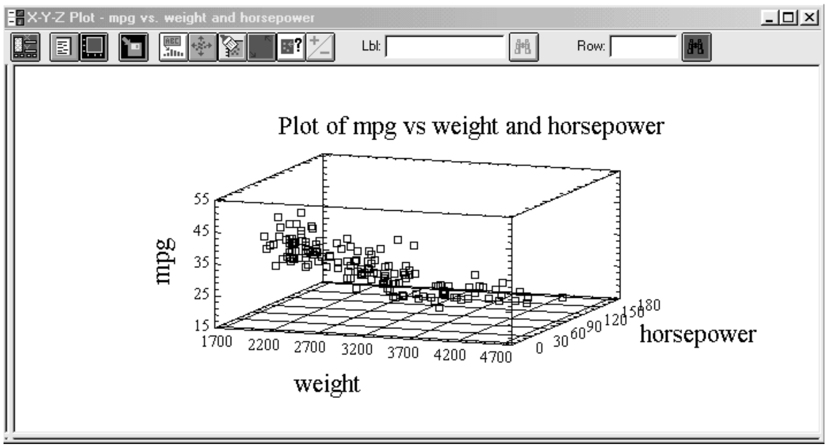
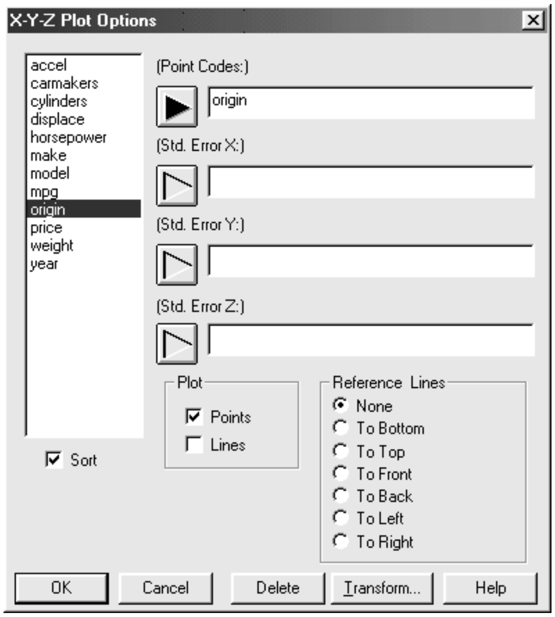


Figure 8-8. Scatterplot



(Point Codes:) origin

(Std. Error X:)

(Std. Error Y:)

(Std. Error Z:)

☑ Sort

Plot
☑ Points
☐ Lines

Reference Lines
● None
○ To Bottom
○ To Top
○ To Front
○ To Back
○ To Left
○ To Right

Figure 8-9. X-Y-Z Plot Options Dialog Box

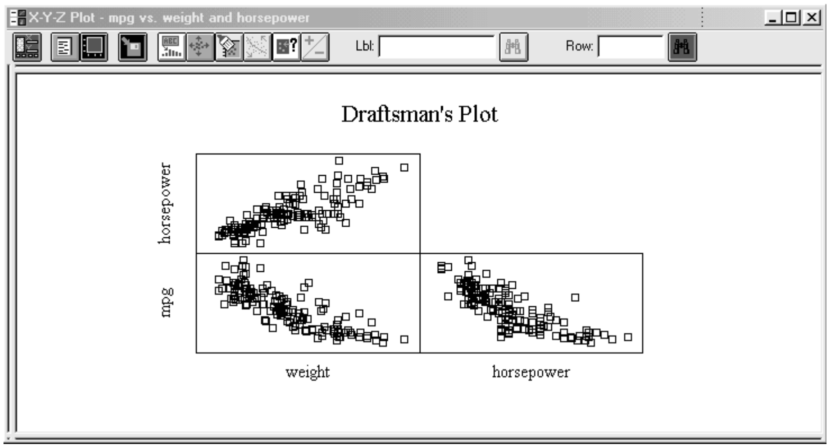


Figure 8-10. Draftsman's Plot

Casement Plot

The Casement Plot option creates values for the X and Y variables in groups determined by the Z variable (see Figure 8-11). The scatterplot contains only those points that fall within the regions covered by the Z variable.

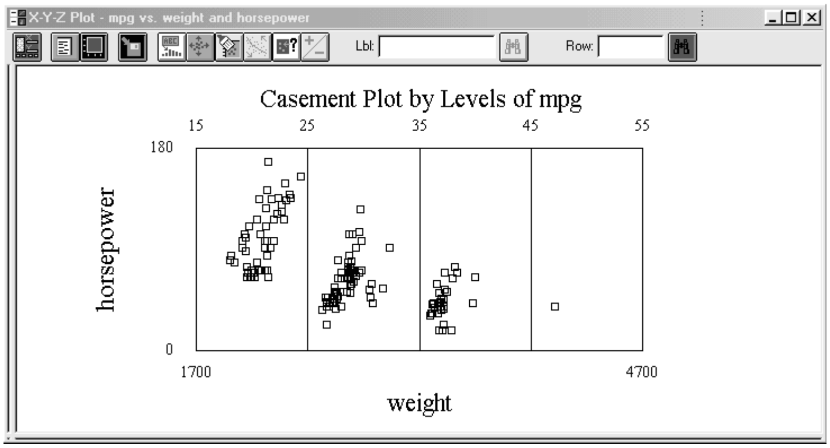

X-Y-Z Plot - mpg vs. weight and horsepower
Lbl:
Row:
Casement Plot by Levels of mpg
15
25
35
45
55
180
0
horsepower
1700
4700
weight

Figure 8-11. Casement Plot

References

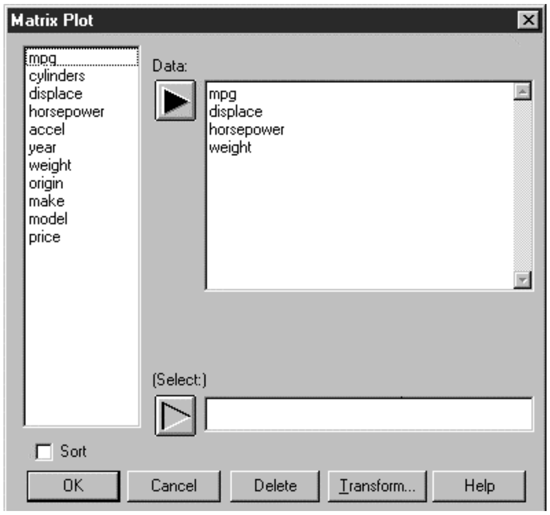
Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. Belmont, California: Wadsworth International Group.

Lapin, L. L. 1987. *Statistics for Modern Business Decisions*. Orlando, Florida: Harcourt Brace Jovanovich, Inc.

Using the Matrix Plot Analysis

The Matrix Plot Analysis produces a scatterplot matrix for three or more numeric variables with a Box-and-Whisker Plot along the diagonal.

To access the analysis, from the menus, choose: PLOT... SCATTERPLOTS...MATRIX PLOT...to display the Matrix Plot Analysis dialog box (see Figure 8-12).



mpg
cylinders
displace
horsepower
accel
year
weight
origin
make
model
price

Data:
mpg
displace
horsepower
weight

(Select:)

□ Sort

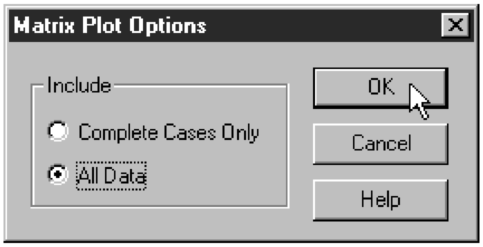
Figure 8-12. Matrix Plot Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option displays the names of the selected variables and the number of complete cases.

Use the *Matrix Plot Options* dialog box to indicate if you want to include complete cases only, or all the data in the analysis (see Figure 8-13). The default is Complete Cases Only.



Include

○ Complete Cases Only

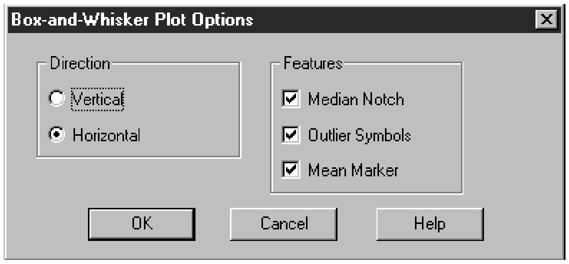
● All Data

Figure 8-13. Matrix Plot Options Dialog Box

Graphical Options

Scatterplot

The output is a matrix of plots with box-and-whisker plots on the diagonal and two-variable scatterplots off the diagonal (see Figure 8-14).



Direction

○ Vertical

● Horizontal

Features

☑ Median Notch

☑ Outlier Symbols

☑ Mean Marker

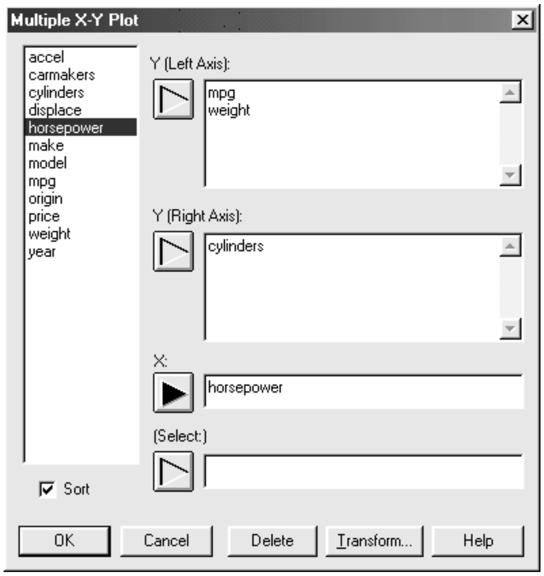
Figure 8-14. Box-and-Whisker Plot Options Dialog Box

The variable indicated on the diagonal forms the vertical axis for every scatterplot in its row and the horizontal axis for every plot in its column.

Using the Multiple X-Y Plot Analysis

The Multiple X-Y Plot Analysis produces scatterplots with one variable on the X-axis and one or more variables on the Y-axis. You can plot Y-axis variables with the scale on the left or the right. The size of the scales is automatically adjusted to reflect the range of values. The scatterplots can contain lines, points, or both.

To access the analysis, from the menus, choose: PLOT... SCATTERPLOTS...MULTIPLE X-Y PLOT...to display the Multiple X-Y Plot Analysis dialog box (see Figure 8-15).



Y (Left Axis): mpg, weight

Y (Right Axis): cylinders

X: horsepower

(Select:)

☑ Sort

Figure 8-15. The Multiple X-Y Plot Options Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the names of the selected variables, and the total number of plotted points.

Graphical Options

Scatterplot

The Scatterplot option creates a plot with one variable on the X-axis and one or more variables on the Y-axis (see Figure 8-16). The pattern of the points indicates the strength and direction of the correlation among the values. The more the points tend to cluster around the straight line, the stronger the relationship (the higher the correlation). *See the description of the X-Y Scatterplot for guidelines on interpreting this type of plot.*

Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot.

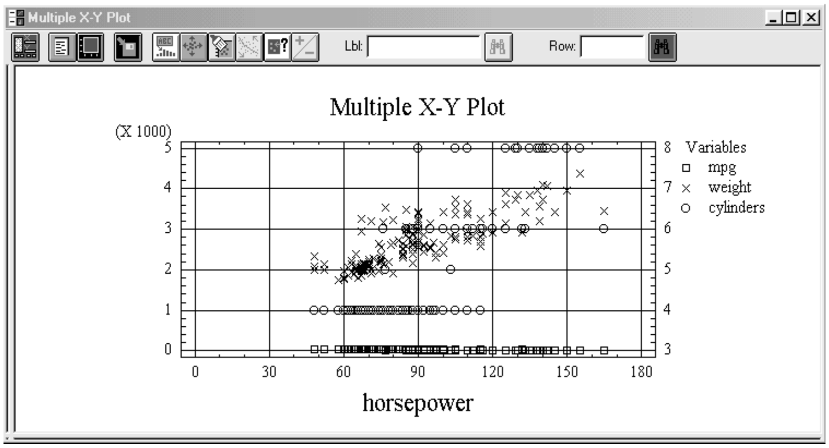
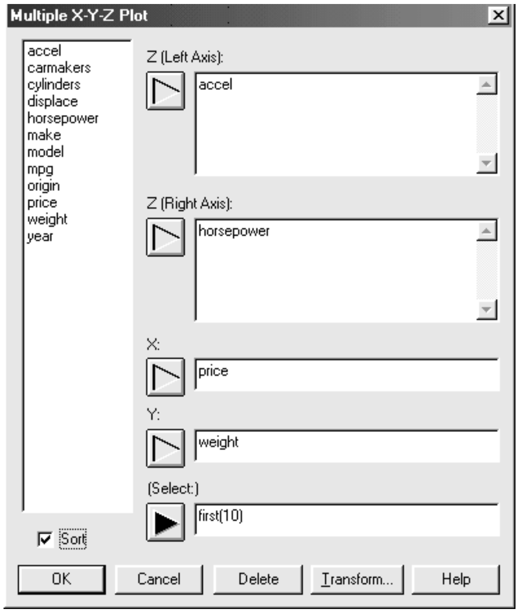


Figure 8-16. Scatterplot

Using the Multiple X-Y-Z Plot Analysis

The Multiple X-Y-Z Plot Analysis produces a standard three-dimensional scatterplot for three or more variables. One variable each is plotted on the X- and Y-axes, while one or more is plotted on the Z-axis. The scale for the Z-axis can appear on the left or the right. The size of the scales is automatically adjusted to reflect the range of the values.

To access the analysis, from the menus, choose: PLOT... SCATTERPLOTS... MULTIPLE X-Y-Z PLOT... to display the Multiple X-Y-Z Plot Analysis dialog box (see Figure 8-17).



Z (Left Axis): accel

Z (Right Axis): horsepower

X: price

Y: weight

(Select:) first(10)

☑ Sort

Figure 8-17. The Multiple X-Y-Z Plot Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the names of the selected variables, and the total number of plotted points.

Graphical Options

Scatterplot

The Scatterplot option creates a scatterplot for three or more variables (see Figure 8-18). One variable each is plotted on the X- and Y-axes, while one or more is plotted on the Z-axis. The scale for the Z-axis can appear on the left or the right. The size of the scales is adjusted to reflect the range of the values. *See the description of the X-Y Scatterplot for guidelines on interpreting this type of plot.*

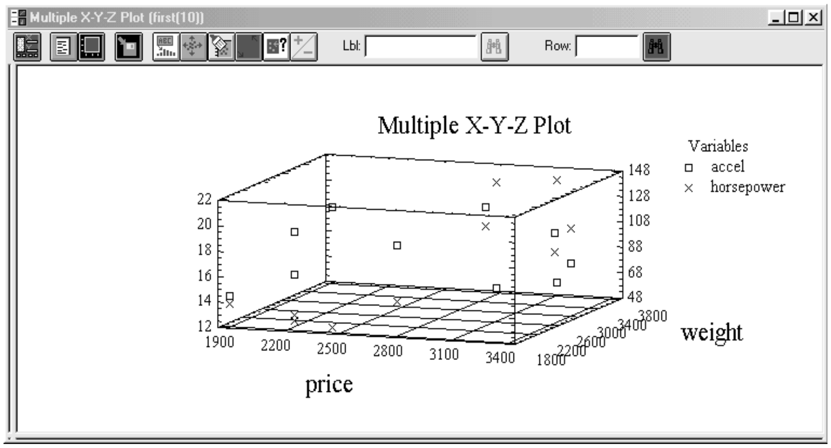


Figure 8-18. Scatterplot

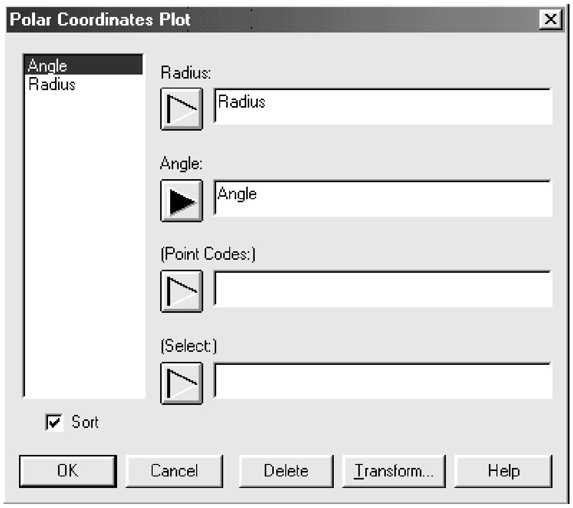
Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot. You can also use the Smooth/Rotate button on the Analysis toolbar to change the angle from which you view this plot.

Using the Polar Coordinates Plot Analysis

The Polar Coordinates Plot Analysis creates a two-dimensional scatterplot or line plot for pairs of points that are defined by radius and angle positions. In other words, it provides information about the location of a point on a graph; for example, finding (x,y) in polar coordinates. This type of plot is useful in manufacturing environments where a polar coordinate system rather than, or in addition to, a rectangular coordinate system might be used to label points on the plane.

A polar coordinate system differs from a rectangular coordinate system in the way it represents points on the plane. In a rectangular coordinate system, a point, (x,y), is defined as the point derived from moving from the origin horizontally *x* units and vertically *y* units. In a polar coordinates system, a point, (x,y), is defined by beginning at the origin, finding the ray that comes from the origin that is an angle of *y* with the positive X-axis, and goes out a distance of *y*, on this ray (The Math Forum, 1997).

To access the analysis, from the menus, choose: PLOT... SCATTERPLOTS... POLAR COORDINATES PLOT... to display the Polar Coordinates Plot Analysis dialog box (see Figure 8-19).



Radius: Radius

Angle: Angle

☑ Sort

Figure 8-19. The Polar Coordinates Plot Analysis Dialog Box

Tabular Options

Analysis Summary

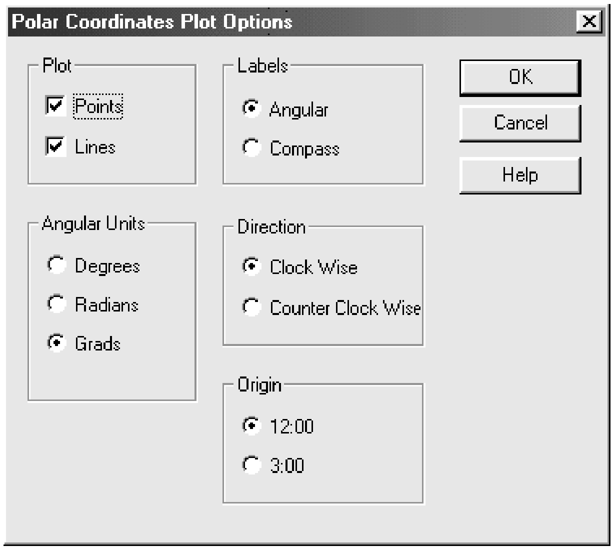
The Analysis Summary option creates a summary of the analysis that displays the names of the variables that contain the values for the radius and angle; it also displays the number of observations.

Use the *Polar Coordinates Plot Options* dialog box to indicate if points or lines will appear on the plot; if the angles will be in degrees, radians, or grads; how the angular locations will be labeled; if the angles should increase clock wise or counter-clock wise; if 0 degrees will be at 12:00 or 3:00 (see Figure 8-20).

Graphical Options

Polar Coordinates Plot

The Polar Coordinates Plot option creates a two-dimensional scatterplot or line plot for pairs of points that are defined by radius and angle positions (see Figure 8-21). The values for the radius variable are plotted against the values for the angle variable.



Polar Coordinates Plot Options

Plot

☑ Points

☑ Lines

Labels

● Angular

○ Compass

Angular Units

○ Degrees

○ Radians

● Grads

Direction

● Clock Wise

○ Counter Clock Wise

Origin

● 12:00

○ 3:00

Figure 8-20. Polar Coordinates Plot Analysis Options Dialog Box

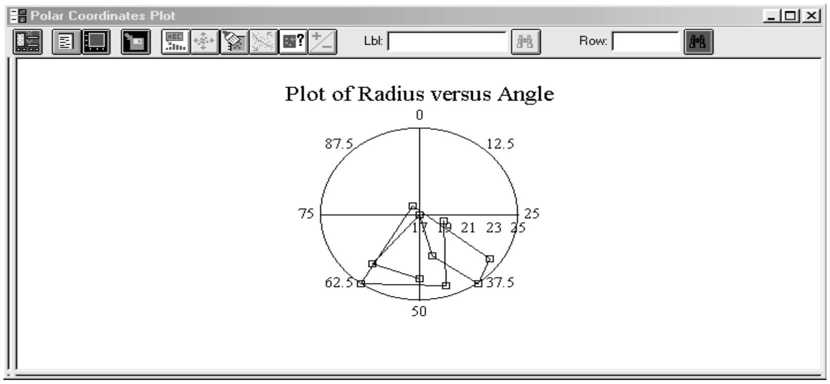


Figure 8-21. Polar Coordinates Plot

References

The Math Forum. 1998. “Ask Dr. Math,” Forum SmartPage Web Tool.

Using the Box-and-Whisker Plot Analysis

The Box-and-Whisker Plot Analysis is a way of summarizing a set of univariate data measured on an interval scale. It is often used in exploratory data analysis to illustrate the major features of the distribution of the data and to compare means and ranges and show outliers and the shape of the distribution. It is particularly useful when large numbers of observations are involved and when you are comparing two or more datasets.

The data are divided into four areas of equal frequency. A box encloses the middle 50 percent, where the median is represented as a vertical line inside the box. The mean may be plotted as a point.

Horizontal lines, called whiskers, extend from each end of the box. The lower (left) whisker is drawn from the lower quartile to the smallest point within 1.5 interquartile ranges from the lower quartile. The other whisker is drawn from the upper quartile to the largest point within 1.5 interquartile ranges from the upper quartile. Values that fall beyond the whiskers, but within 3 interquartile ranges (suspect outliers), are plotted as individual points.

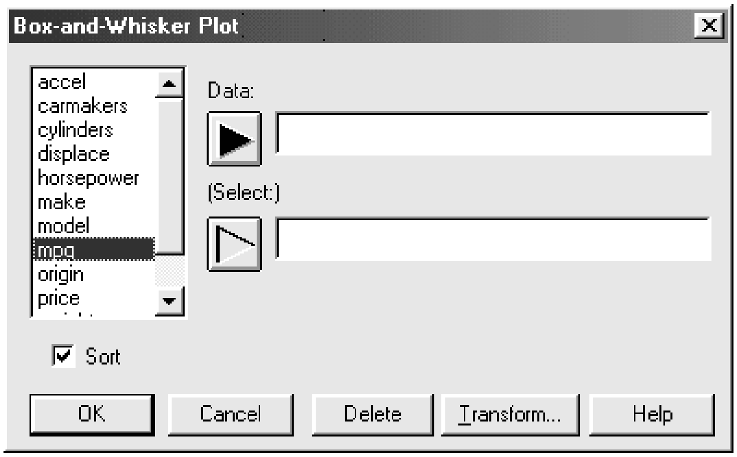
Far outside points (outliers) are distinguished by a special character (a point with a + through it). Far outside points are points more than 3 interquartile ranges below the lower quartile or above the upper quartile.

To access the analysis, from the menus, choose: PLOT... EXPLORATORY PLOTS... BOX-AND-WHISKER PLOT... to display the Box-and-Whisker Plot Analysis dialog box (see Figure 8-22).

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the name of the selected variable, the number of values in the variable, and the range of the values.



☑ Sort

Figure 8-22. The Box-and-Whisker Plot Analysis Dialog Box

Graphical Options

Box-and-Whisker Plot

The Box-and-Whisker Plot option creates a Box-and-Whisker Plot, which is a graphical summary of the presence of outliers in the data (see Figure 8-23). The length of the box represents the interquartile range of the data, which is a measure of variability. The wider the box, the more variability exists in the data. This information is helpful when you are comparing two or more samples of data. The length of the whiskers is also important. If one whisker is clearly longer than the other, the data distribution is probably skewed in the direction of the longest whisker.

Use the *Box-and-Whisker Plot Options* dialog box to indicate if the plot will appear in a vertical or horizontal direction; and to choose features for the plot such as median notch, outlier symbols, and mean marker (see Figure 8-24).

References

Frigge, M., Hoagland, D. C., and Iglewicz, B. 1989. “Some Implementations of the Boxplot,” *American Statistician*, **43**:50-54.

McGill, R., Tukey, J. W., and Larsen, W. A. 1978. “Variation of Box Plots,” *American Statistician*, **32**:12-16.

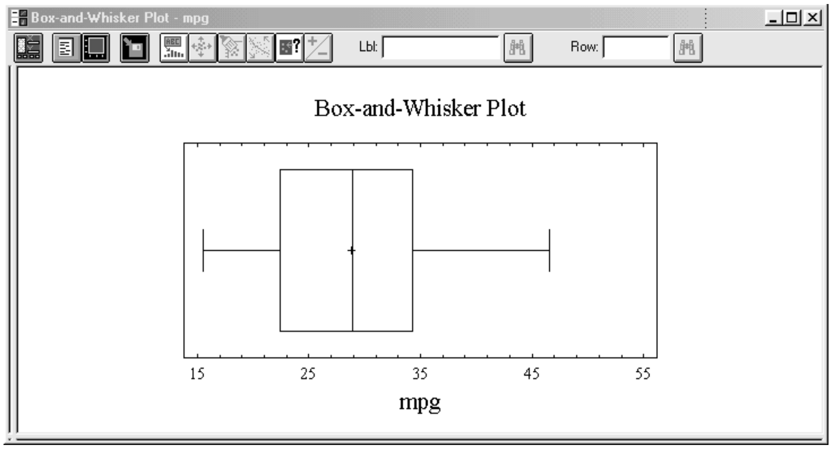
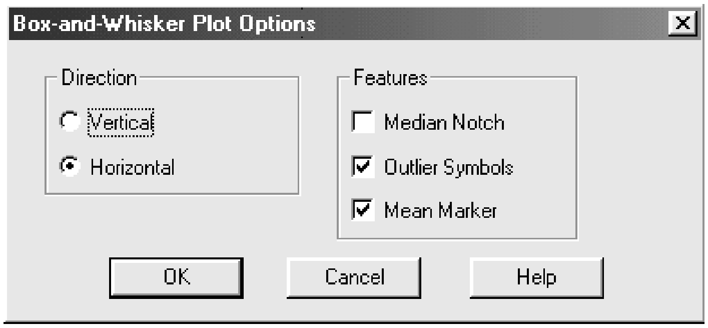


Figure 8-23. Box-and-Whisker Plot


Box-and-Whisker Plot Options

Direction

○ Vertical

● Horizontal

Features

□ Median Notch

☑ Outlier Symbols

☑ Mean Marker

Figure 8-24. Box-and-Whisker Plot Options Dialog Box

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Belmont, California: Duxbury Press.

Using the Multiple Box-and-Whisker Plot Analysis

The Multiple Box-and-Whisker Plot Analysis is simply an extension of the Box-and-Whisker Plot analysis. If you can subdivide the data into groups, this analysis lets you produce separate Box-and-Whisker plots for each group on one chart. This facilitates comparisons of the median, range, and extreme value for each group.

Each of the plots is a statistical summary of a set of univariate observations. It is an exploratory data-analysis tool that is useful for studying symmetry and distributional assumptions, and for detecting outliers.

The data are shown in four areas of equal frequency. A box encloses the middle 50 percent, where the median is represented as a horizontal line inside the box. Vertical lines, called whiskers, extend from each end of the box. The lower whisker is drawn from the first quartile to the smallest point within 1.5 interquartile ranges from the first quartile. The other whisker is drawn from the third quartile to the largest point within 1.5 interquartile ranges from the third quartile.

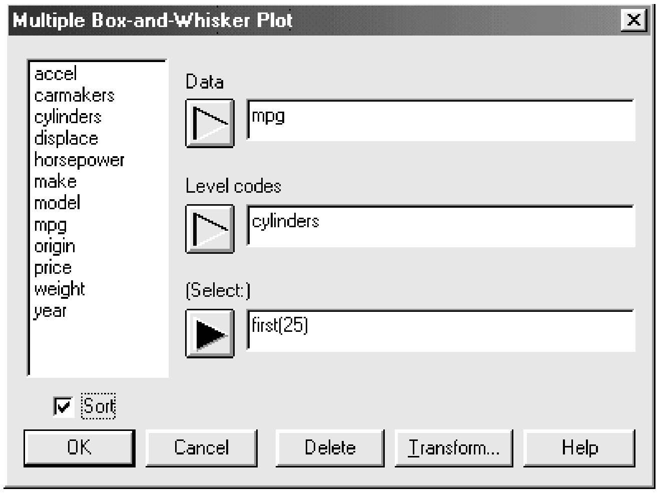
Individual points are plotted beyond the whiskers. Far outliers (points that lie more than 3 interquartile ranges below the first quartile or above the third quartile), are plotted using a special character (+) that makes the points easier to identify.

To access the analysis, from the menus, choose: PLOT... EXPLORATORY PLOTS... MULTIPLE BOX-AND-WHISKER PLOT... to display the Multiple Box-and-Whisker Plot Analysis (see Figure 8-25).

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the name of the dependent variable, the name of the factor, the number of observations, and the number of levels.



Data: mpg

Level codes: cylinders

(Select:) first(25)

☑ Sort

Figure 8-25. The Box-and-Whisker Plot Analysis Dialog Box

Graphical Options

Multiple Box-and-Whisker Plot

The Multiple Box-and-Whisker Plot option creates a graphical summary of the outliers in the data (see Figure 8-26). The length of the box represents the interquartile range of the data, which is a measure of variability. The wider the box, the more variability exists in the data. This information is helpful when you are comparing two or more samples of data. The length of the whiskers is also important. If one whisker is clearly longer than the other, the data distribution is probably skewed in the direction of the longest whisker.

Use the *Multiple Box-and-Whisker Plot Options* dialog box to indicate if the plot will appear in a vertical or horizontal direction; and to choose features for the plot such as median notch, outlier symbols, and mean marker (see Figure 8-27).

References

Frigge, M., Hoagland, D. C., and Iglewicz, B. 1989. “Some Implementations of the Boxplot,” *American Statistician*, **43**:50-54.

McGill, R., Tukey, J. W., and Larsen, W. A. 1978. “Variation of Box Plots,” *American Statistician*, **32**:12-16.

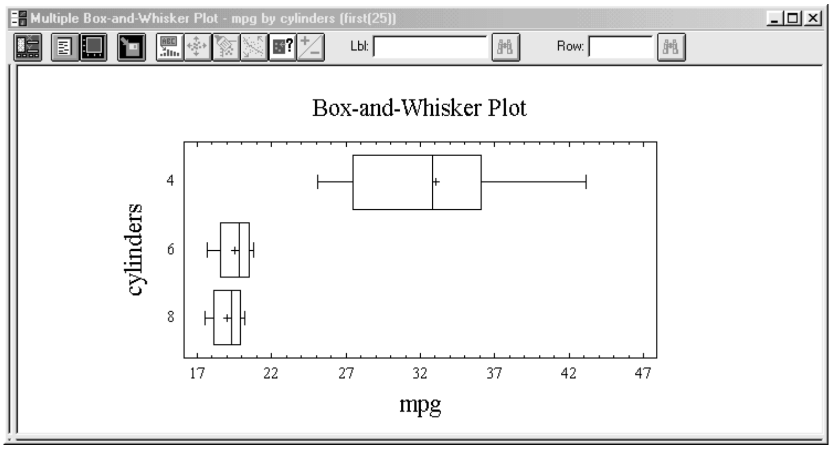
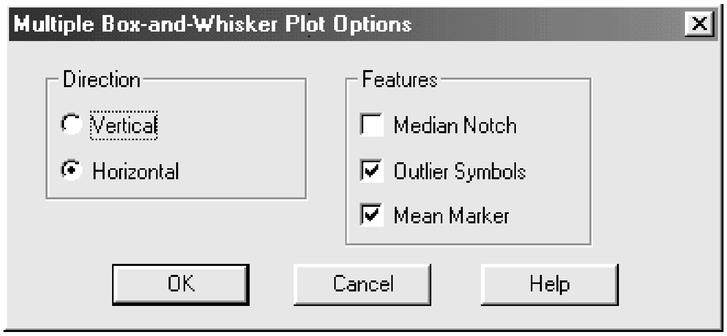

Multiple Box-and-Whisker Plot - mpg by cylinders (first(25))

Figure 8-26. Multiple Box-and-Whisker Plot


Multiple Box-and-Whisker Plot Options

Direction
○ Vertical
● Horizontal

Features
□ Median Notch
☑ Outlier Symbols
☑ Mean Marker

Figure 8-27. Multiple Box-and-Whisker Plot Options Dialog Box

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

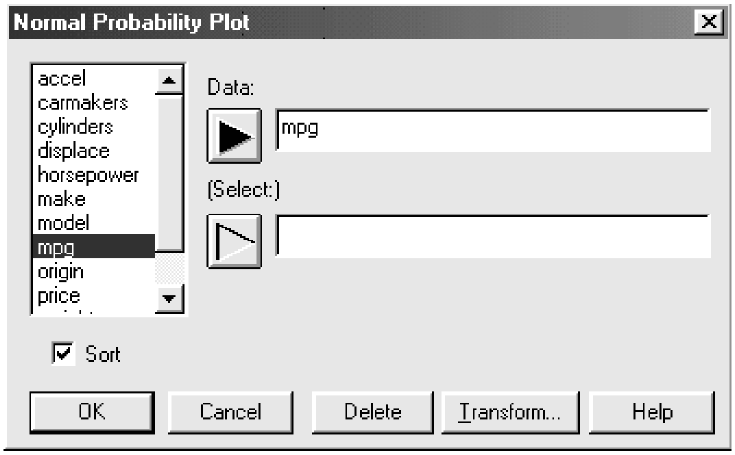
Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Belmont, California: Duxbury Press.

Using the Normal Probability Plot Analysis

The Normal Probability Plot Analysis is used to test a single sample to determine whether or not it is likely to be a sample taken from a normally distributed population. The purpose of the test is to produce a graph and to determine if the shape of the plotted points form a linear pattern (that is, if the points roughly follow a straight line).

The plot also provides an easy way to identify outliers. Outliers clearly highlight potential performance bottlenecks. The points (effects) are arranged in ascending order of their value. The plot consists of an arithmetic (interval) horizontal axis and a vertical axis scaled so the cumulative distribution function (cdf) plots as a straight line.

To access the analysis, from the menus, choose: PLOT... EXPLORATORY PLOTS... NORMAL PROBABILITY PLOT... to display the Normal Probability Plot Analysis dialog box (see Figure 8-28).


Normal Probability Plot

Data: mpg

(Select:)

☑ Sort

Figure 8-28. The Normal Probability Plot Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the name of the selected variable, and the range of the values.

Graphical Options

Normal Probability Plot

The Normal Probability Plot option creates a plot using values that have been sorted from smallest to largest (see Figure 8-29). If the data come from a normal distribution, the points should fall approximately along a straight line. To help determine the closeness of the points to a straight line, a reference line is superimposed on the plot. The reference line passes through the median with slope determined by the interquartile range. Points showing significant curvature indicate skewness in the data.

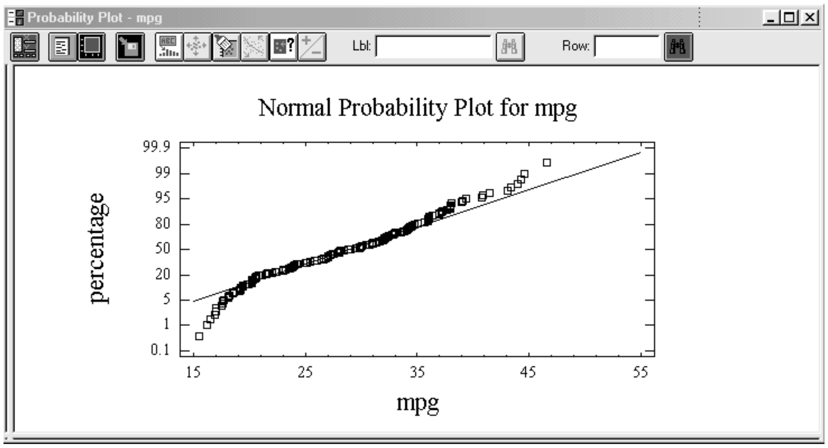


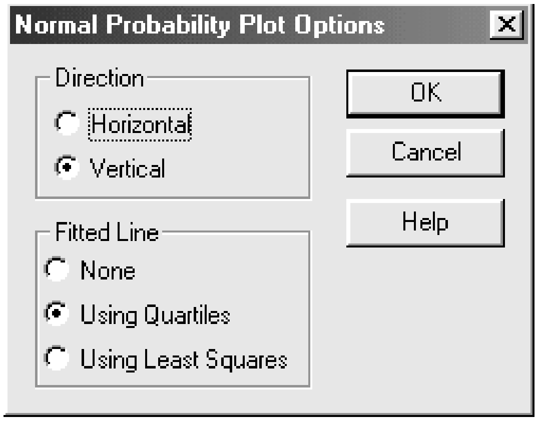
Figure 8-29. Normal Probability Plot

Use the *Normal Probability Plot Options* dialog box to indicate if the plot will display in a horizontal or vertical direction; and to indicate if a fitted line will appear on the plot; if so, whether quartiles or least squares will be the method used to fit the line (see Figure 8-30).

References

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. Belmont, California: Duxbury Press.

Law, A. M. and Kelton, W. D. 1982. *Simulation Modeling and Analysis*. New York: McGraw-Hill.



Normal Probability Plot Options

Direction

○ Horizontal

● Vertical

Fitted Line

○ None

● Using Quartiles

○ Using Least Squares

Figure 8-30. Normal Probability Plot Options Dialog Box

Using the Frequency Histogram Analysis

The Frequency Histogram Analysis provides a way to summarize data that are measured on a discrete or continuous interval scale. It is used to illustrate the major features of the distribution of the data, such as the number of school-age children in each of *n* families (Weiss and Hassett, 1991).

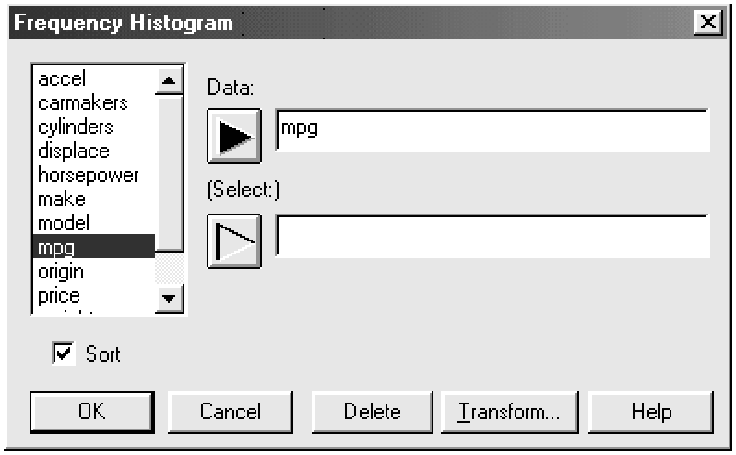
The range of possible values are divided into classes or groups. For each group, a rectangle is drawn with a base height equal to the frequency of the values in that specific group, and an area proportional to the number of observations that fall into that group.

To access the analysis, from the menus, choose: PLOT... EXPLORATORY PLOTS... FREQUENCY HISTOGRAM... to display the Frequency Histogram Analysis dialog box (see Figure 8-31).

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the name of the selected variable, the number of observations in the variable, and the range of the values.



Data: mpg

(Select:)

☑ Sort

Figure 8-31. The Frequency Histogram Analysis Dialog Box

Graphical Options

Frequency Histogram

The Frequency Histogram option creates a histogram of the values in the variables (see Figure 8-32). The graph is plotted with classes on the horizontal axis and frequencies on the vertical, where the frequency of each class is represented by vertical bars.

Use the *Frequency Plot Options* dialog box to enter values for the number of classes into which the data will be grouped, as well as for the lower limit for the first class and the upper limit for the last class (see Figure 8-33).

You can also indicate if the scale for the Y-axis will be relative and/or cumulative, if the current scaling will be retained if you make changes on the dialog box, and if you want to create a histogram or a polygon.

References

Lapin, L. L. 1987. *Statistics for Modern Business Decisions*. Orlando, Florida: Harcourt Brace Jovanovich, Inc.

Weiss, N. A. and M. J. Hassett. 1991. *Introductory Statistics*, third edition. Chicago: Addison-Wesley Publishing Company, Inc.

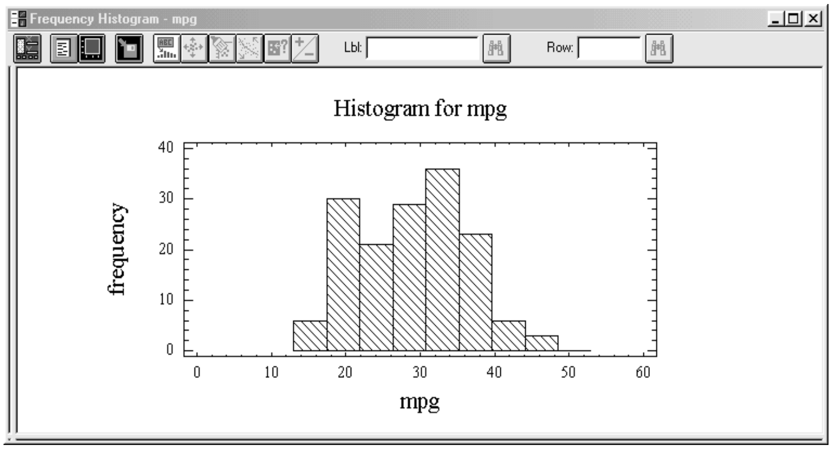
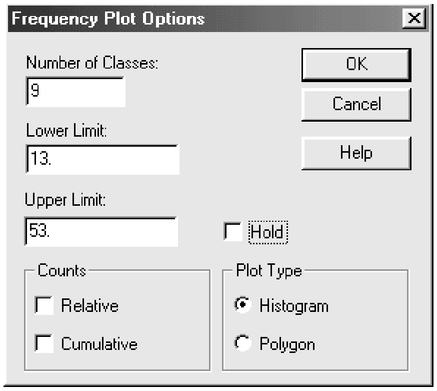


Figure 8-32.   Frequency Histogram



Number of Classes: 9

Lower Limit: 13.

Upper Limit: 53.

☐ Hold

Counts

☐ Relative

☐ Cumulative

Plot Type

● Histogram

○ Polygon

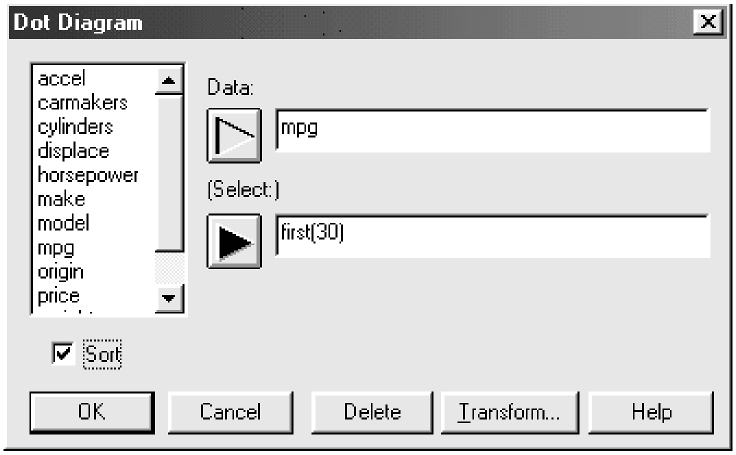
Figure 8-33.   Frequency Plot Options Dialog Box

Using the Dot Diagram Analysis

The Dot Diagram Analysis is way of summarizing data in an exploratory data analysis, to show the primary features of the distribution of the data in a convenient form. The graph provides another type of graphical display for numeric data, which is particularly useful for showing the relative positions of data in a dataset, or for comparing two or more datasets. The plot is also helpful in detecting unusual outliers or gaps in the data.

The plot is similar to a barchart or a histogram, with the bars or rectangles replaced by a series of dots. Each dot represents a fixed number of individuals.

To access the analysis, from the menus, choose: PLOT... EXPLORATORY PLOTS... DOT DIAGRAM... to display the Dot Diagram Analysis dialog box (see Figure 8-34).



Data: mpg

(Select:) first(30)

☑ Sort

Figure 8-34. The Dot Diagram Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the name of the selected variable, the number of observations in the variable, the mean, and the standard deviation.

Graphical Options

Dot Diagram

The Dot Diagram option creates a plot that is useful for showing the relative positions of data in a dataset where every value in a numeric column is shown as a small square (see Figure 8-35).

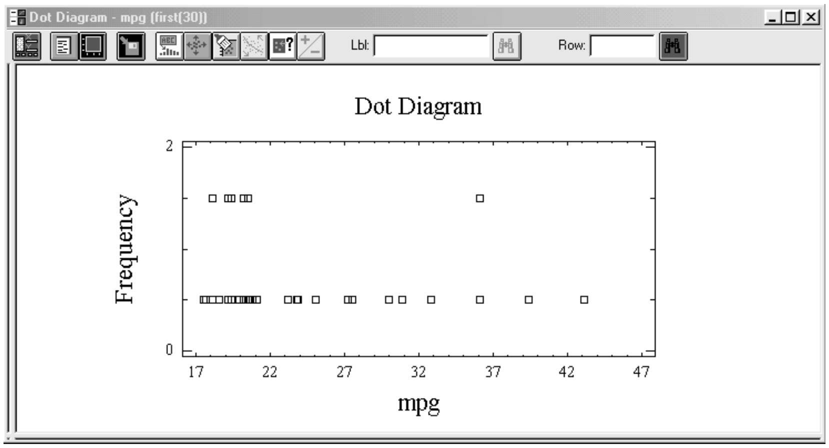
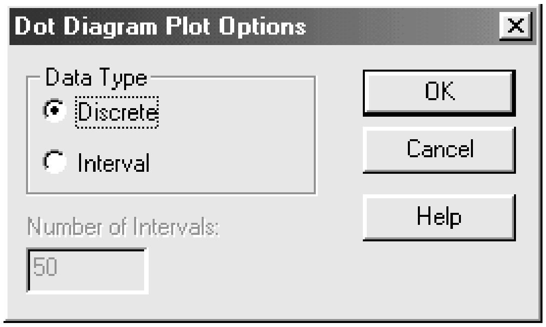


Figure 8-35.   Dot Diagram

Use the *Dot Diagram Plot Options* dialog box to indicate if discrete or interval data will appear on the plot, and to enter a number for the increments on the X-axis (see Figure 8-36).



Data Type
● Discrete
○ Interval

Number of Intervals: 50

Figure 8-36.   Dot Diagram Plot Options Dialog Box

References

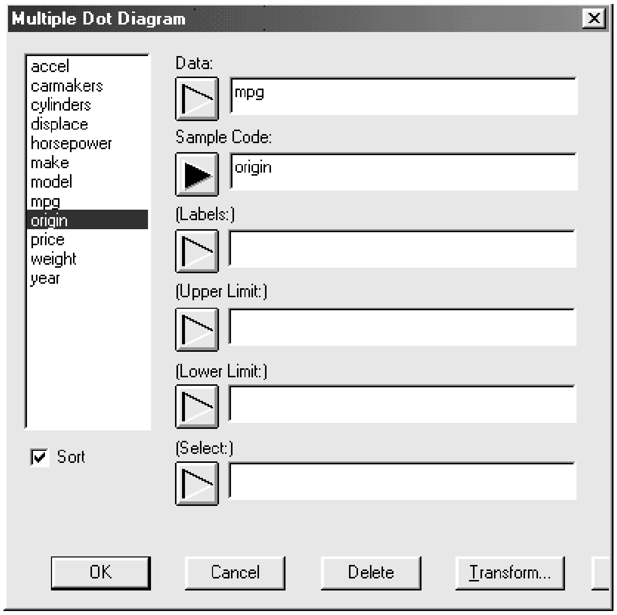
Weiss, N. A. and Hassett, M. J. 1991. *Introductory Statistics*, third edition. New York: Addison-Wesley Publishing Company.

Using the Multiple Dot Diagram Analysis

The Multiple Dot Diagram Analysis creates a dot diagram for data that are divided into more than one group. It displays summary statistics and confidence intervals for each group.

This plot is a way to summarize data that are often used in exploratory data analysis to illustrate the major features of the distribution of data in a convenient form. A Multiple Dot Diagram can also help detect any unusual observations (outliers), or gaps in the dataset.

To access the analysis, from the menus, choose: PLOT... EXPLORATORY PLOTS... MULTIPLE DOT DIAGRAM... to display the Multiple Dot Diagram Analysis dialog box (see Figure 8-37).



Multiple Dot Diagram

Data: mpg

Sample Code: origin

☑ Sort

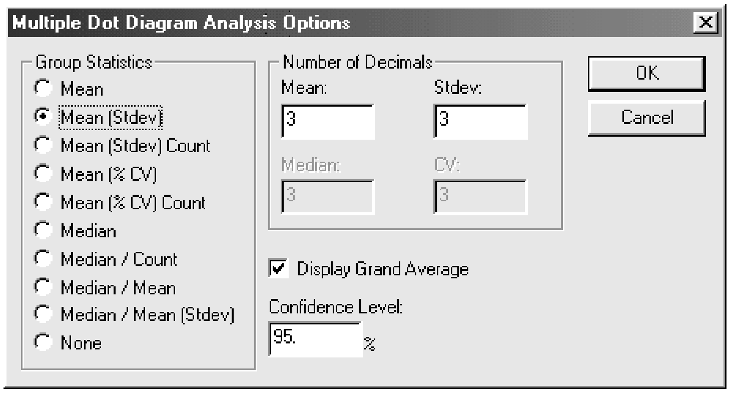
Figure 8-37.   The Multiple Dot Diagram Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary that includes the names of the variables that contain the data and the level codes. It also displays the number of levels and observations in the analysis, as well as the selected values for the upper and lower limits.

Use the *Multiple Dot Diagram Analysis Options* dialog box to indicate the type of mean or median statistics, if any, that will appear on the plot; to enter values for the mean, median, standard deviation, and coefficient of variation; to indicate if a line representing the Grand Average should appear on the plot; and to enter a value for the confidence level that will be used to calculate the intervals around each group mean (see Figure 8-38).



Multiple Dot Diagram Analysis Options

Group Statistics

- ○ Mean
- ● Mean (Stdev)
- ○ Mean (Stdev) Count
- ○ Mean (% CV)
- ○ Mean (% CV) Count
- ○ Median
- ○ Median / Count
- ○ Median / Mean
- ○ Median / Mean (Stdev)
- ○ None

Number of Decimals

Mean: 3

Stdev: 3

Median: 3

CV: 3

☑ Display Grand Average

Confidence Level: 95. %

Figure 8-38. Multiple Dot Diagram Analysis Options Dialog Box

Graphical Options

Multiple Dot Diagram

The Multiple Dot Diagram option creates a plot that shows the location for each level mean, as well as the confidence intervals for the means (see Figure 8-39). The X-axis shows the mean and standard deviation for each level.

A vertical line is drawn at the grand mean of the total observations. The plot displays the grand mean and the standard deviation at the top of the vertical line.

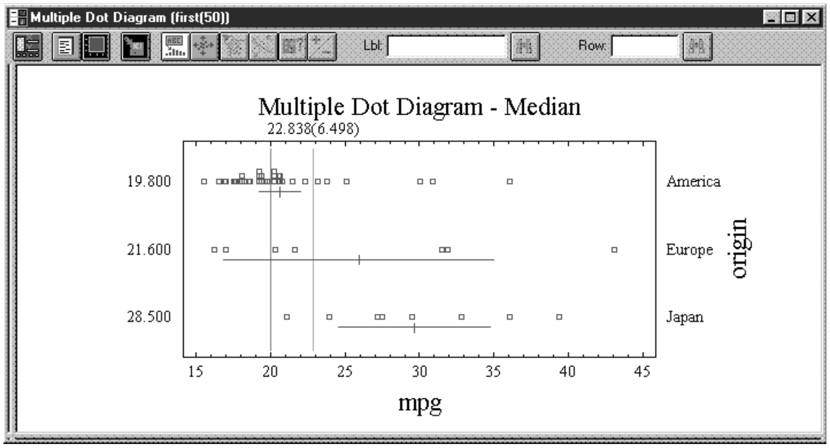


Figure 8-39. Multiple Dot Diagram

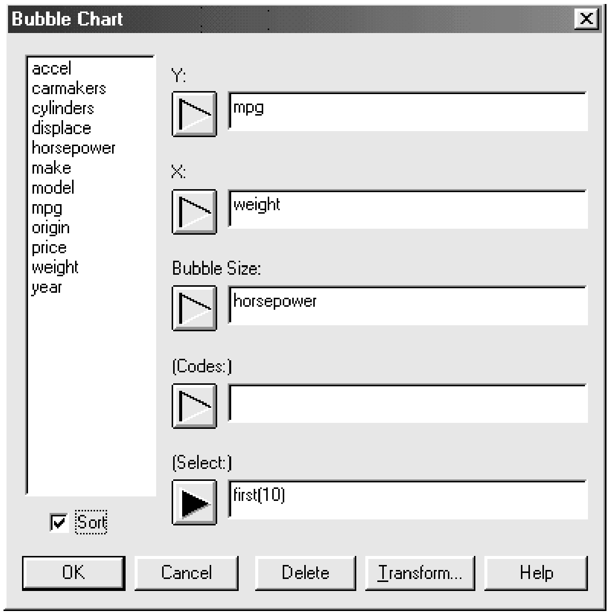
Using the Bubble Chart Analysis

The Bubble Chart Analysis allows you to create an X-Y scatterplot consisting of circles, which is another way of plotting three parameters on two axes — two variables are represented by the X- and Y-coordinates of the plotted points, and a third variable is portrayed by a plotted circle or "bubble." The size of the bubble depends on the numeric value of the third variable. The larger the value, the larger the bubble. This approach is a basic graphics tool that is helpful when you are looking for a fairly direct way to view data in higher dimensions.

For example, suppose you are analyzing a product's competitive standing; that is, top-rated products and lower-rated products. Each product in the study is represented by a bubble that moves in relation to its market position. Each bubble represents a point, allowing you to quickly spot important shifts in market position.

The chart shows the location of X and Y as well as their relative sizes; the size of each bubble is based on the size of the third numeric variable. You can use different colors for different bubbles, indicate if you want filled bubbles, and enter the maximum size for the largest bubbles.

To access the analysis, from the menus, choose: PLOT... EXPLORATORY PLOTS... BUBBLE CHART... to display the Bubble Chart Analysis dialog box (see Figure 8-40).



Y: mpg

X: weight

Bubble Size: horsepower

(Codes:)

(Select:) first(10)

☑ Sort

Figure 8-40. The Bubble Chart Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which displays the names of the variables that will appear on the X- and Y-axes as well as the name of the variable that contains the bubble size. The number of observations is also shown.

Graphical Options

Bubble Chart

The Bubble Chart option creates a symbolic X-Y scatterplot (see Figure 8-41). The size of each bubble on the chart is based on the magnitude of a third numeric variable.

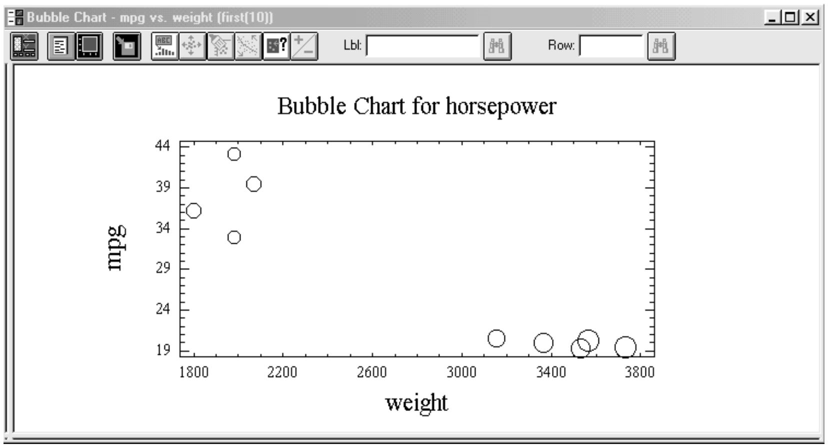
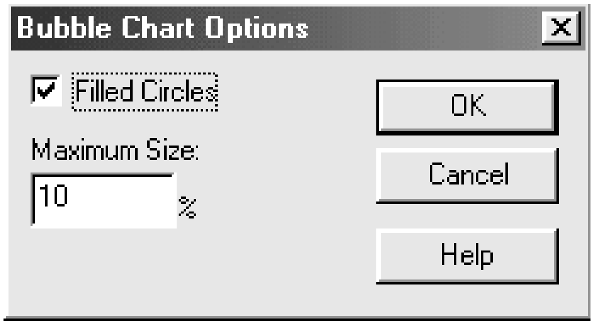


Figure 8-41. *Bubble Chart*

Use the *Bubble Chart Options* dialog box to indicate if filled bubbles (colored bubbles) should be used, and to enter a number that will determine the maximum size of the largest bubble (see Figure 8-42).


Bubble Chart Options

☑ Filled Circles

Maximum Size: 10 %

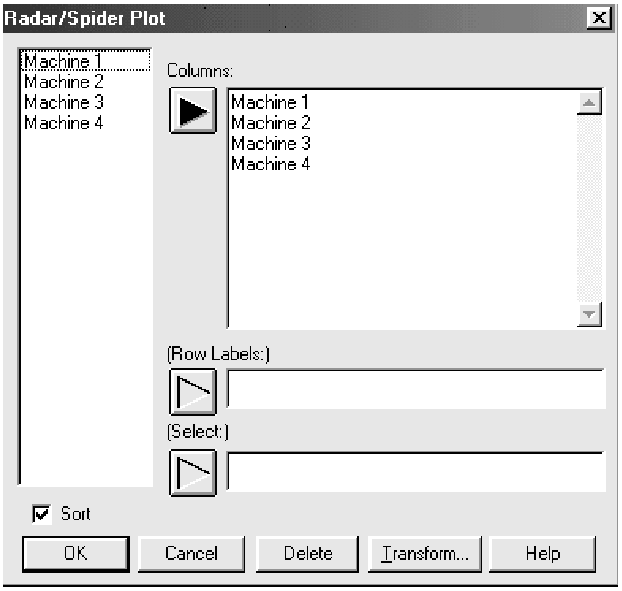
Figure 8-42. *Bubble Chart Options Dialog Box*

Using the Radar/Spider Plot Analysis

The Radar/Spider Plot Analysis is a simple visualization technique often considered valuable in meeting the demands of multivariate data because of its ability to simultaneously portray numerous aspects of the data. This technique plots multiple measurements on equally spaced radii that are linked to create a star-like form. The radii extend from the center of a circle and give an overall impression of changing values across subjects.

A Radar plot is much like a Star plot. In a Radar plot, the value of the measurement is also represented by radii stretching out from the center of a circle; however, each radius stands for a subject instead of a variable. Points of different colors represent the response for each variable. When there are too many variables and subjects, the pattern of the data is concealed.

To access the analysis, from the menus, choose: PLOT... EXPLORATORY PLOTS... RADAR/SPIDER PLOT... to display the Radar/Spider Plot Analysis dialog box (see Figure 8-43).



Radar/Spider Plot

Machine 1
Machine 2
Machine 3
Machine 4

Columns:
Machine 1
Machine 2
Machine 3
Machine 4

(Row Labels:)

(Select:)

☑ Sort

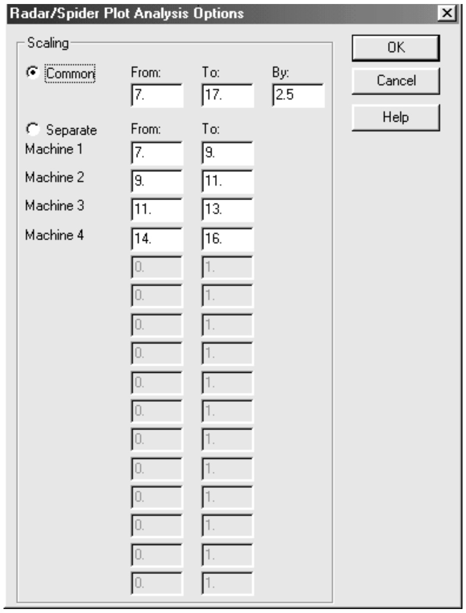
Figure 8-43. The Radar/Spider Chart Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that displays the names of the data variables.

Use the *Radar/Spider Plot Analysis Options* dialog box to make changes to the analysis (see Figure 8-44). You can indicate type of scale and grid that will be used for a plot, if spokes will be drawn, and if filled polygons will be used.



Scaling

● Common

From:	To:	By:
7.	17.	2.5

○ Separate

	From:	To:
Machine 1	7.	9.
Machine 2	9.	11.
Machine 3	11.	13.
Machine 4	14.	16.
	0.	1.
	0.	1.
	0.	1.
	0.	1.
	0.	1.
	0.	1.
	0.	1.
	0.	1.
	0.	1.
	0.	1.
	0.	1.
	0.	1.

Figure 8-44. Radar/Spider Plot Analysis Options Dialog Box

Graphical Options

Overlay Plot

The Overlay Plot option creates a plot that shows the data in each row, with the size of each of the variables plotted along one of the spokes (see Figure 8-45). You can plot up to 16 cases on a single diagram, which is helpful in comparing cases from a multivariate perspective.

Using the Barchart Analysis

The Barchart Analysis is a way of summarizing a set of categorical data, often a frequency table. It is used in exploratory data analysis to illustrate the major features of the distribution of data in a convenient form. It displays the data using a number of rectangles, of the same width, each of which represents a particular category. The length (area) of each rectangle is proportional to the number of cases in the category it represents.

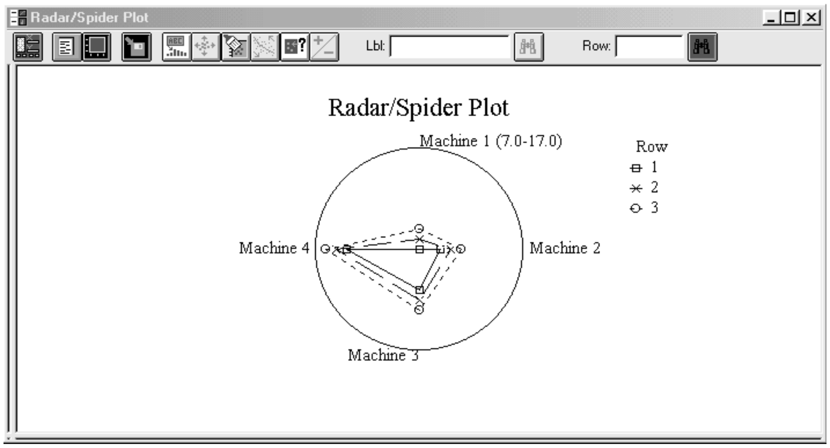


Figure 8-45. Radar/Spider Plot

For example, you might use barcharts to present the Republican affiliation of members of various religious categories: 31 percent of Protestants surveyed reported affiliation with the Republicans, roughly 17 percent of Catholics did, and so on.

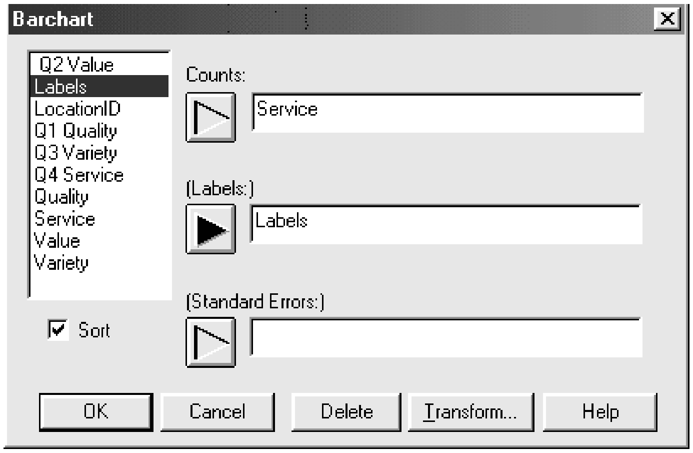
You can display the barcharts either horizontally or vertically; they can include horizontally or vertically stacked bars, clustered bars, or percentages. The variable you use should contain tabulated data (counts) — a count of the number of observations in each class. If the data contain more than 20 classes, an error message displays.

To access the analysis, from the menus, choose: PLOT... BUSINESS CHARTS... BARCHART... to display the Barchart Analysis dialog box (see Figure 8-46).

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the name of the selected variable, the number of values in the variable, and the sum of the values.



Counts: Service

(Labels:) Labels

(Standard Errors:)

☑ Sort

Figure 8-46. The Barchart Analysis Dialog Box

Graphical Options

Barchart

The Barchart option creates a plot similar to a histogram except that its bars are separated (see Figure 8-47). Classes are displayed on the horizontal axis; relative frequency on the vertical. The height of the vertical bar is equal to the relative frequency of the class.

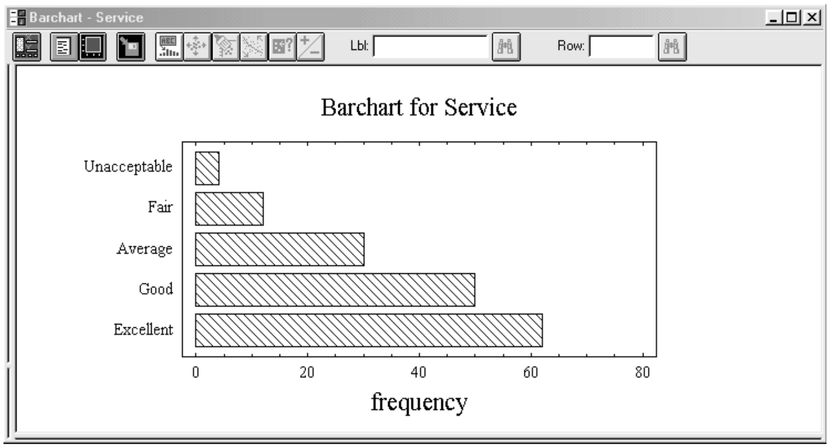
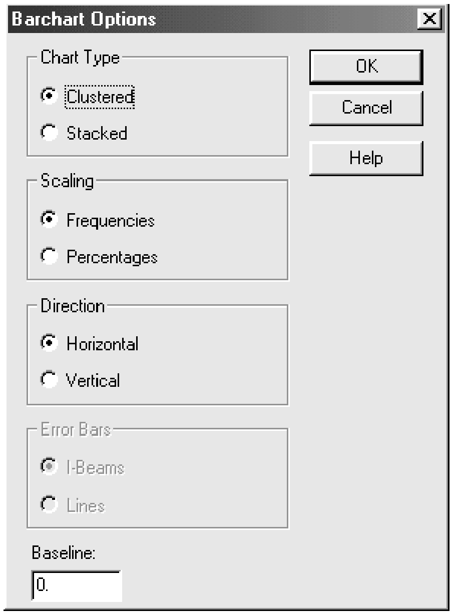


Figure 8-47. Barchart

Use the *Barchart Options* dialog box to choose a format for the bars on the chart (clustered or stacked); indicate if you want to plot frequencies or percentages; indicate if the plot will be displayed horizontally or vertically; indicate if I-Beams or Lines will be used for the bars; and enter values for the starting point for the bars (see Figure 8-48).



Barchart Options

Chart Type
● Clustered
○ Stacked

Scaling
● Frequencies
○ Percentages

Direction
● Horizontal
○ Vertical

Error Bars
● I-Beams
○ Lines

Baseline:
0.

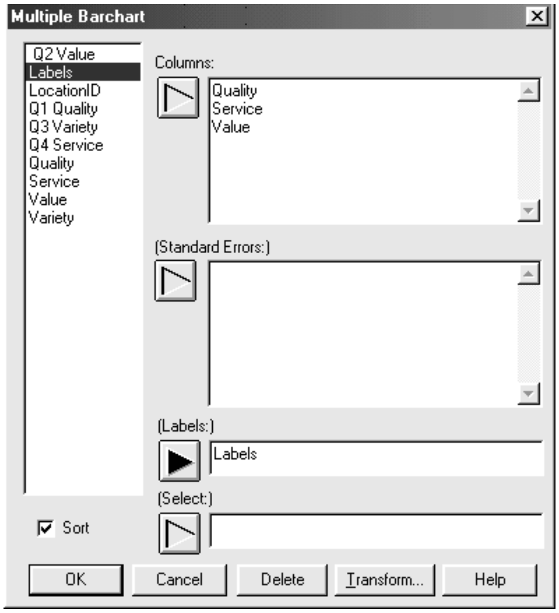
Figure 8-48. Barchart Options Dialog Box

Using the Multiple Barchart Analysis

The Multiple Barchart Analysis creates a plot with one or more frequency bars for each classification factor (row). The analysis is helpful when you need to graphically represent a secondary classification factor within a primary classification factor.

The variable you use should contain tabulated data (counts) — a count of the number of observations in each class. If the data contain more than 20 classes, an error message displays.

To access the analysis, from the menus, choose: PLOT... BUSINESS CHARTS... MULTIPLE BARCHART... to display the Multiple Barchart Analysis dialog box (see Figure 8-49).



Q2 Value
Labels
LocationID
Q1 Quality
Q3 Variety
Q4 Service
Quality
Service
Value
Variety

Columns:
Quality
Service
Value

(Standard Errors:)

(Labels:)
Labels

(Select:)

☑ Sort

Figure 8-49. The Multiple Barchart Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which contains the names of the row and column variables, the number of variables you selected (if you used this option), and the number of observations, rows, and columns.

Graphical Options

Multiple Barchart

The Multiple Barchart option creates a plot with a column of numeric values for each of the variables (see Figure 8-50). A single bar is placed for each row of the column.

Use the *Multiple Barchart Options* dialog box to choose a format for the bars on the chart; indicate if you want to plot frequencies or percentages; indicate

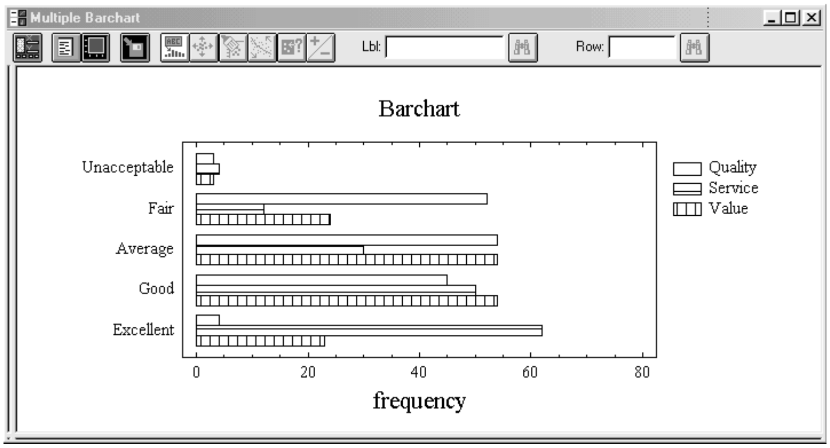


Figure 8-50. The Multiple Barchart

if the plot will be displayed horizontally or vertically; indicate if I-Beams or Lines will be used for the bars; and enter values for the starting point for the bars (see Figure 8-51).

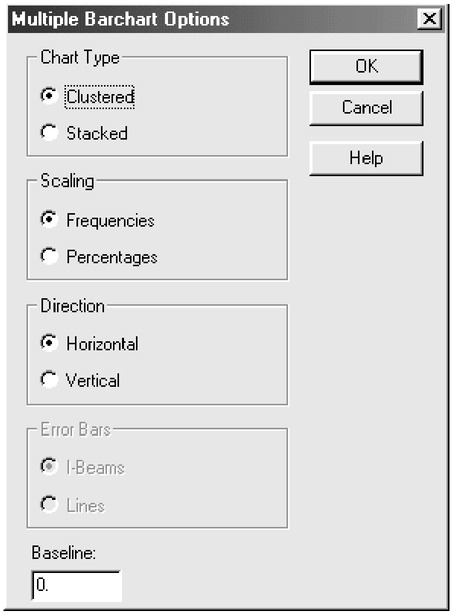

Multiple Barchart Options

Chart Type
- ● Clustered
- ○ Stacked

Scaling
- ● Frequencies
- ○ Percentages

Direction
- ● Horizontal
- ○ Vertical

Error Bars
- ● I-Beams
- ○ Lines

Baseline: 0.

Figure 8-51. Multiple Barchart Options Dialog Box

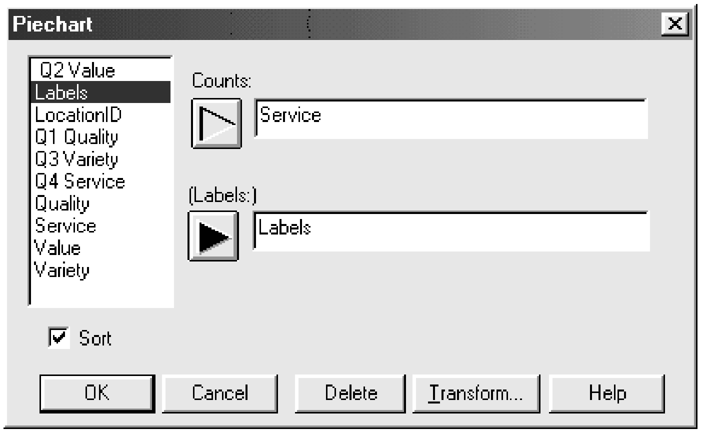
Using the Piechart Analysis

The Piechart Analysis creates a plot that is a way of summarizing a set of categorical data. The plot contains a circle divided into segments where each segment is proportional to the number of cases in that category. You can offset a segment of the chart to display it more prominently.

The variable you use should contain tabulated data (counts) — a count of the number of observations in each class. If the data contain more than 20 classes, an error message displays.

A Piechart is useful for displaying breakdowns of percentages. For example, you could use a Piechart to show how an athletic-shoe manufacturer spent its advertising budget of 6 million dollars: 3 million for television advertisements, 2 million for sponsorships, and 1 million for newspaper advertisements.

To access the analysis, from the menus, choose: PLOT... BUSINESS CHARTS... PIECHART... to display the Piechart Analysis dialog box (see Figure 8-52).



Counts: Service

(Labels:) Labels

☑ Sort

Figure 8-52.   The Piechart Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the name of the selected variable, the number of observations the variable represents, and the sum of the values.

Graphical Options

Piechart

The Piechart option creates a circular graph made up of segments whose circumference represent the frequency that corresponds to one row in a frequency table (see Figure 8-53).

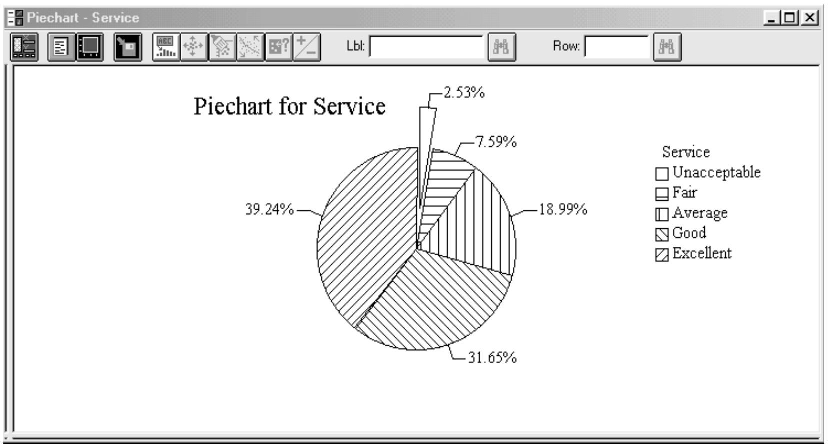
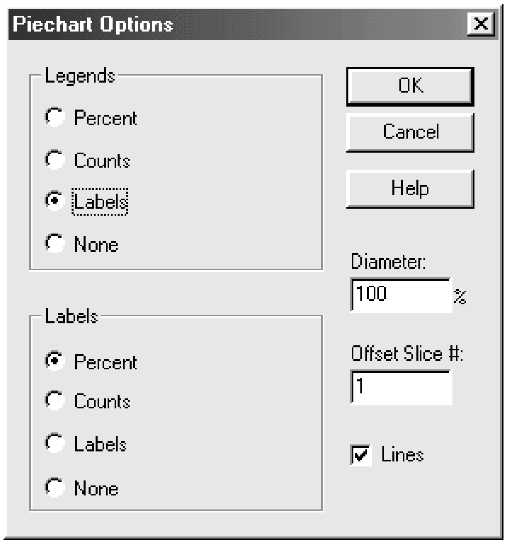


Figure 8-53.   The Piechart

Use the *Piechart Options* dialog box to choose labels for the legends and the segments; to enter a size for the circle; to enter the number of the segment that will be offset; and to indicate if lines will appear from the label to the segments on the Piechart (see Figure 8-54).



Piechart Options

Legends

○ Percent

○ Counts

● Labels

○ None

Labels

● Percent

○ Counts

○ Labels

○ None

Diameter: 100 %

Offset Slice #: 1

☑ Lines

Figure 8-54. *Piechart Options Dialog Box*

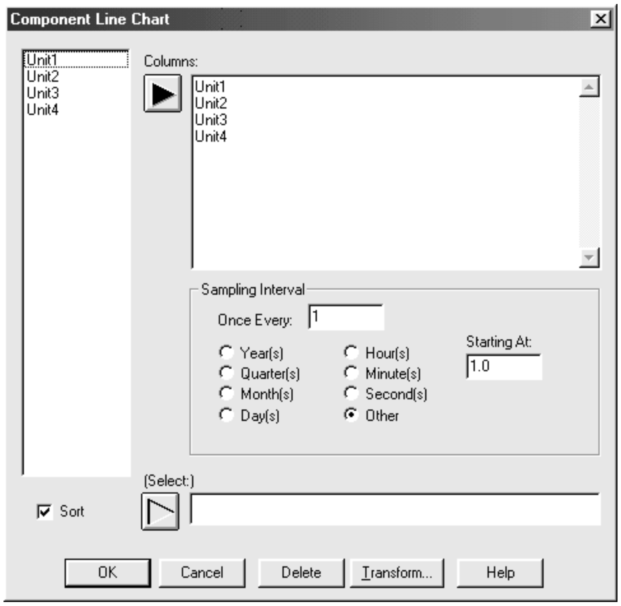
References

Lapin, L. L. 1987. *Statistics for Modern Business Decisions*, fourth edition. New York: Harcourt Brace Jovanovich.

Using the Component Line Chart Analysis

The Component Line Chart Analysis is useful for displaying one or more sets of time-series data, which are useful for investigating trends, patterns, or other nonrandom behavior in the time series; for example stock prices. You can plot original values individually or cumulatively.

To access the analysis, from the menus, choose: PLOT... BUSINESS CHARTS... COMPONENT LINE CHART... to display the Component Line Chart Analysis dialog box (see Figure 8-55).



Component Line Chart

Unit1
Unit2
Unit3
Unit4

Columns:
Unit1
Unit2
Unit3
Unit4

Sampling Interval
Once Every: 1
○ Year(s) ○ Hour(s)
○ Quarter(s) ○ Minute(s)
○ Month(s) ○ Second(s)
○ Day(s) ● Other
Starting At: 1.0

(Select:)

☑ Sort

Figure 8-55. The Component Line Chart Analysis Options Dialog Box

Tabular Options

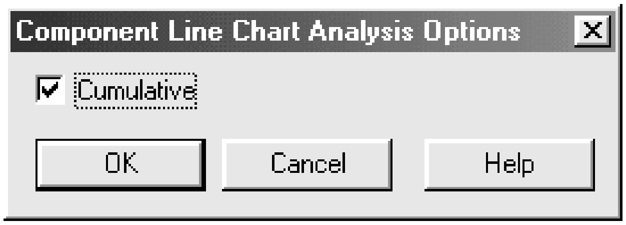
Analysis Summary

The Analysis Summary option creates a summary of the analysis that displays the names of the variables in the columns, the start index, and the sampling interval. It also displays the number of periods that contain data.

Use the *Component Line Chart Analysis Options* dialog box to indicate if the variables should be plotted cumulatively, which is the default (see Figure 8-56).

If you choose Cumulative, the patterns for the variables appear vertically stacked, one on top of the other. In plots that contain positive data, the values are added at each time point. In plots that contain negative data, the values are subtracted from the values for the previous variable.

If you deselect the default (remove the checkmark), the patterns for each variable appear as overlays and points in one variable may hide the points for



☑ Cumulative

Figure 8-56. Component Line Chart Analysis Options Dialog Box

variables that are plotted later. To minimize this problem, first enter the variable with the smallest values; then the variable with the second smallest values; and so on. The program plots the last variable first, the second-to-last on top of the last variable, and so on until all the variables appear on the plot.

Graphical Options

Component Line Chart

The Component Line Chart option creates a filled polygon with the values plotted against time (see Figure 8-57). The values appear on the plot in the order in which they appear in the DataSheet. The area under any specific time series is filled with a unique color. You can create a cumulative or noncumulative plot.

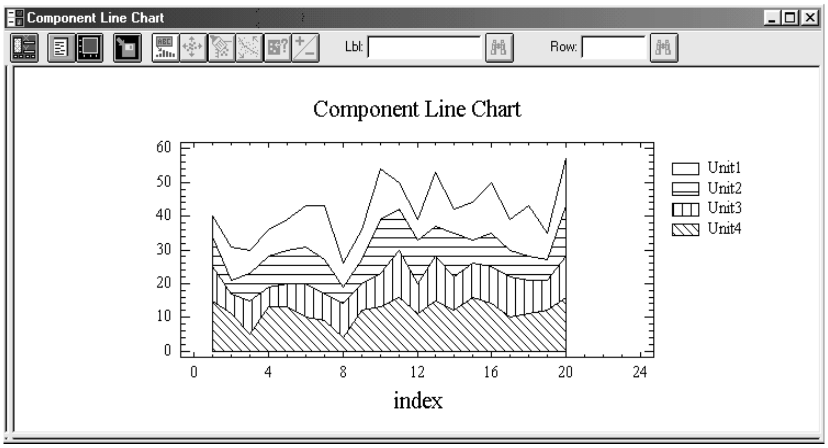
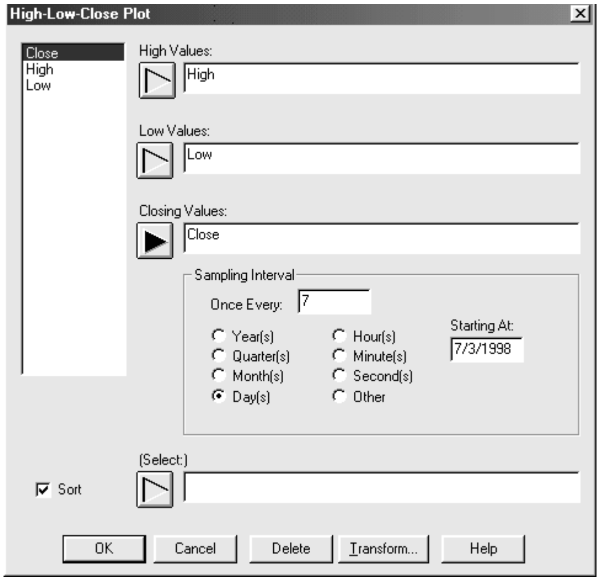


Figure 8-57. Component Line Chart

Using the High-Low-Close Plot Analysis

The High-Low-Close Plot Analysis is valuable when you want to display stock, commodity, currency, and other market data that fluctuate greatly from hour-to-hour, day-to-day, or week-to-week. To convey a sense of short-term change while also viewing long-term change requires that each category of data show a range of values.

To access the analysis, from the menus, choose: PLOTS... BUSINESS CHARTS... HIGH-LOW-CLOSE PLOT... to display the High-Low-Close Plot Analysis dialog box (see Figure 8-58).



High Values: High

Low Values: Low

Closing Values: Close

Sampling Interval

Once Every: 7

○ Year(s)
○ Quarter(s)
○ Month(s)
● Day(s)
○ Hour(s)
○ Minute(s)
○ Second(s)
○ Other

Starting At: 7/3/1998

(Select:)

☑ Sort

Figure 8-58.   The High-Low-Close Plot Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that displays the names of the variables for High, Low, and Close values; and the values for the Starting Index and Sampling Interval.

It also displays the number of periods that contain data; and values for the Maximum High, Minimum Low, and Average Close.

Graphical Options

High-Low-Close Plot

The High-Low-Close Plot option creates a plot with a vertical line that extends from each low value to its corresponding high value (see Figure 8-59). A horizontal line is drawn at each of the closing values.

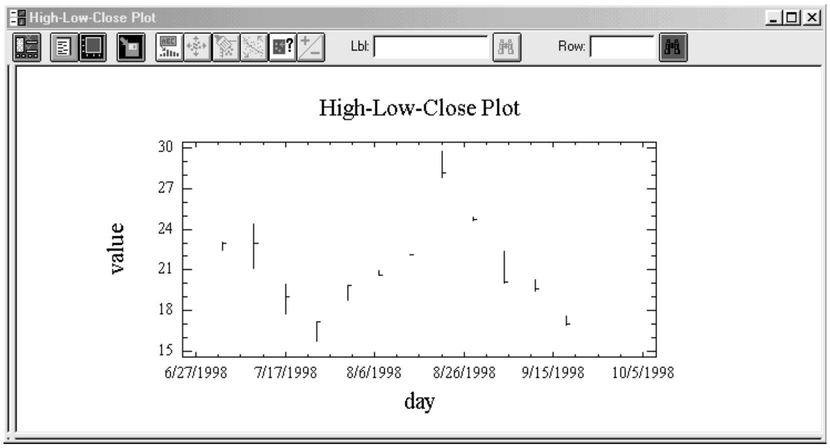


Figure 8-59. The High-Low-Close Plot

Using the Probability Distribution Analysis

The Probability Distribution Analysis contains functions that allow you to perform three basic operations for each of 24 different probability distributions:

- calculate probabilities
- create plots of the probability and cumulative distributions
- generate random numbers.

The 24 distributions are

- **Bernoulli**
  A distribution whose outcome has only two possibilities: Success or failure; for example, heads or tails, good or bad; defective or nondefective. Data fit to this distribution should have values of only 0 or 1.

- **Binomial**
  A distribution that fits data that follow a Binomial distribution. The distribution gives the probability of observing successes in a fixed number of independent or Bernoulli trials. When you use this distribution you must choose the number of trials (experiments). Data fit to this distribution should be integers greater than or equal to 0.

- **Discrete Uniform**
  A distribution that allocates equal probabilities to all integer values between a lower and an upper limit. Data fit to this distribution should be integers.

- **Geometric**
  A distribution that characterizes the number of failures that occur before the first success in a series of Bernoulli trials; a special case of Negative Binomial distribution, where $k = 1$. Data fit to this distribution should be integers.

- **Hypergeometric**
  A distribution that arises when a random selection is made between objects of two distinct types (success,fail). The sampling occurs without replacement; that is, each time an item is drawn and studied, it is not placed back into the population. The distribution gives the probability for the number of successes. Data fit to this distribution should be integers greater than or equal to 0.

- **Negative Binomial**
  A distribution that characterizes the number of failures before the *k*th success in a series of Bernoulli trials. When you use this distribution you must declare the number of successes. Data fit to this distribution should be integers greater than 0.

- **Poisson**
  A distribution that expresses probabilities that concern the number of events per unit time; for example, the number of times a computer malfunctions per year. Data fit to this distribution should be integers greater than or equal to 0.

- **Beta**
  A distribution that is useful for random variables that are constrained to lie between 0 and 1; characterized by two parameters: Shape 1 and Shape 2.

- **Cauchy**
  A distribution that fits data that follow a Cauchy distribution. The distribution's probability density function has no mean and an infinite variance. It is characterized by two parameters: Mode and Scale. Data fit to this distribution should be continuous data with a Mode between -infinity to +infinity and a Scale greater than 0.

- **Chi-Square**
  A distribution that is useful for random variables that are constrained to be greater than 0; characterized by one parameter: Degrees of Freedom. This distribution is used most often as the sampling distribution for various statistical tests.

- **Erlang**
  A distribution useful for random variables that are constrained to be greater than 0, such as the time required to complete a task; characterized by two parameters: Shape and Scale. This distribution is a special case of the Gamma distribution, which requires that the Shape parameter be an integer.

- **Exponential**
  A distribution that fits time-series data, such as arrival times, where arrivals are expected at a constant rate; useful for random variables that are constrained to be greater than 0. This distribution is a special case of both the Gamma and the Weibull distributions.

- **Extreme Value**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by two parameters: Mode and Scale. Also known as a Gumbels' distribution.

- **F (Variance Ratio)**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by two parameters: Numerator Degrees of Freedom and Denominator Degrees of Freedom. It is often used as the distribution for test statistics that are created as variance ratios.

- **Gamma**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by two parameters: Shape and Scale. This

distribution is often used to model data that are positively skewed, such as the time required to complete a task.

- **Laplace**
  A distribution useful for random variables from a distribution that is more peaked than a Normal distribution; characterized by two parameters: Mean and Scale. This distribution is sometimes called the “double exponential” distribution because it looks like an exponential distribution with a mirror image.

- **Logistic**
  A distribution useful for random variables that are not constrained to be greater than or equal to 0; characterized by two parameters: Mean and Standard Deviation.

- **Lognormal**
  A distribution that is useful for processes in which the value is a random proportion of the previous value, such as personal incomes or particle sizes from breakage processes. The log of data that follow the lognormal distribution are normally distributed. The distribution is positively skewed and can take on various shapes. Data fit to this distribution should be values greater than 0; characterized by two parameters: Mean and Standard Deviation.

- **Normal**
  A distribution that is useful in instances when you plot a Frequency Histogram of the data and the bars form a common, bell-shaped curve. This option is the default.

- **Pareto**
  A distribution with a decreasing density function. One parameter, Shape, is necessary to specify the distribution. Data fit to this distribution should be values greater than 0.

- **Student’s** *t*
  A distribution useful in forming confidence intervals when the variance is unknown, testing to determine if two sample means are significantly different, or testing to determine the significance of coefficients in a regression. The distribution is similar in shape to a Normal distribution. The mean of the *t* distribution is always equal to 0, while the standard deviation is usually slightly greater than 1. One parameter, Degrees of Freedom, is necessary to completely specify the distribution.

- **Triangular**
  A distribution useful for random variables that are constrained to lie between two fixed limits. Unlike the Uniform distribution, in which all the values between the limits are equally likely, the Triangular distribution peaks at some value between the two limits. This distribution is characterized by three parameters: Lower Limit, Central Value (Mode), and Upper Limit.

- **Uniform**
  A distribution useful for characterizing data that range over an interval of values, each of which is equally likely. The distribution is completely determined by the smallest possible value, *a*, and the largest possible value, *b*. The mean equals (*a* + *b*)/2, while the variance equals ((*b* - *a*)^2)/12. For discrete data, there is a related discrete uniform distribution.

- **Weibull**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by two parameters: Shape and Scale. Because its failure-rate curves can take various shapes, it is an appropriate model for product failures. The distribution is a generalization of an Exponential distribution.

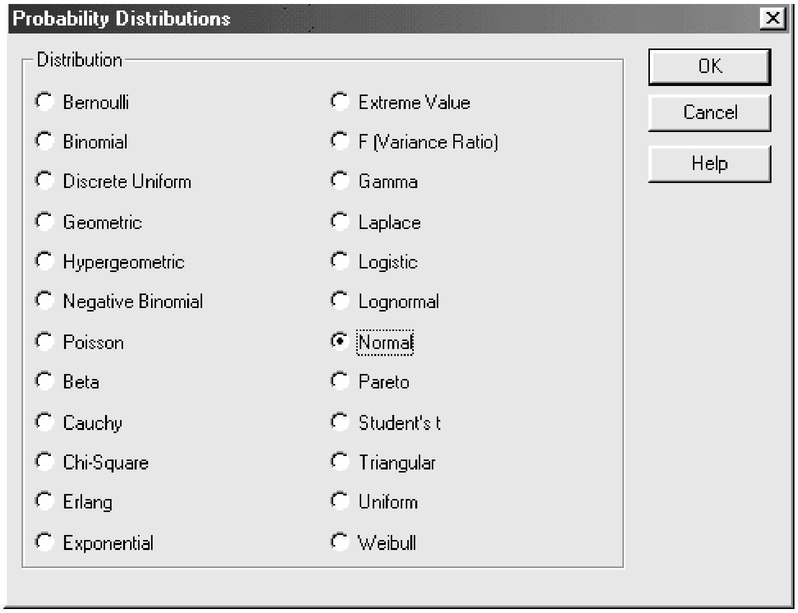
The first six distributions are appropriate for discrete data; that is, data that can take only integer values. For discrete variables, the probabilities are relative frequencies. The other distributions are appropriate for continuous variables that can take any value over a continuous interval.

To access the analysis, from the menus, choose: PLOT... PROBABILITY DISTRIBUTIONS... to display the Probability Distributions Analysis dialog box (see Figure 8-60).

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the distribution you select; the contents of this pane will vary depending on the distribution. The summary also lists the parameters for the distribution, and shows the values for the mean and standard deviation.



Distribution

○ Bernoulli
○ Binomial
○ Discrete Uniform
○ Geometric
○ Hypergeometric
○ Negative Binomial
○ Poisson
○ Beta
○ Cauchy
○ Chi-Square
○ Erlang
○ Exponential
○ Extreme Value
○ F (Variance Ratio)
○ Gamma
○ Laplace
○ Logistic
○ Lognormal
● Normal
○ Pareto
○ Student's t
○ Triangular
○ Uniform
○ Weibull

Figure 8-60. The Probability Distributions Analysis Dialog Box

Cumulative Distribution

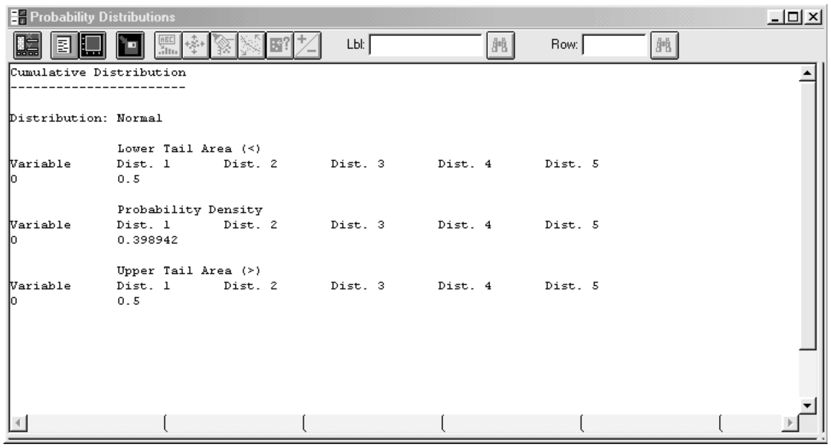
The Cumulative Distribution option creates a summary of the evaluation for the cumulative distribution you selected (see Figure 8-61). The summary includes the tail areas for up to five critical values of the distribution. It also displays a value for the height of the probability density function at a given value.

Use the *Cumulative Distribution Options* dialog box to enter values for the random variables.

Inverse CDF

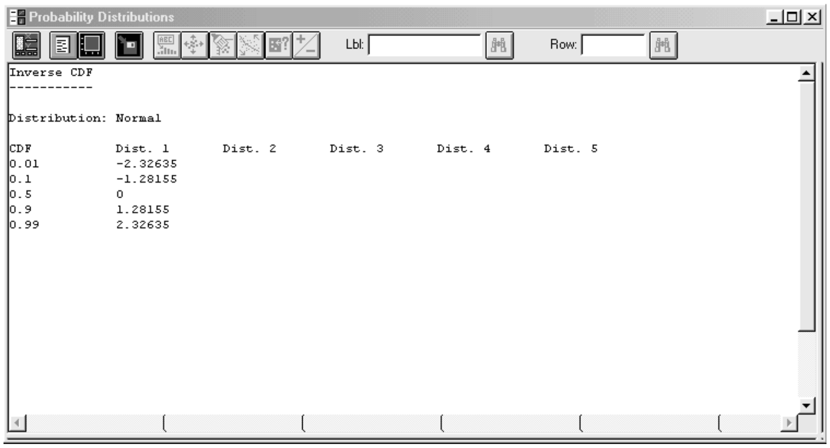
The Inverse CDF option creates a summary of the critical values for the selected distribution (see Figure 8-62). The summary includes the tail areas for up to five critical values of the distribution.

Use the *Inverse CDF Options* dialog box to enter values for the tail areas.



```
Cumulative Distribution
-----------------------

Distribution: Normal

              Lower Tail Area (<)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
0             0.5

              Probability Density
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
0             0.398942

              Upper Tail Area (>)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
0             0.5
```

Figure 8-61.   Cumulative Distribution for Normal Distribution

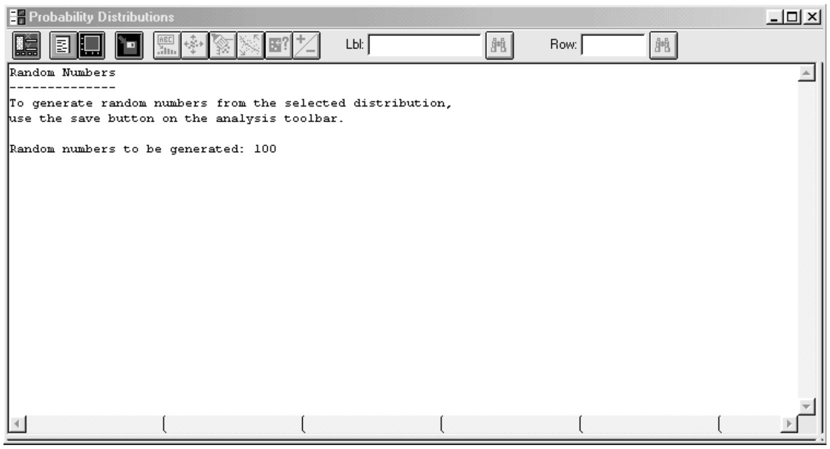


```
Inverse CDF
-----------

Distribution: Normal

CDF           Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
0.01          -2.32635
0.1           -1.28155
0.5           0
0.9           1.28155
0.99          2.32635
```

Figure 8-62.   Inverse CDF for Normal Distribution

Random Numbers

The Random Numbers option creates a summary of the random numbers from a distribution (see Figure 8-63). You can save the values for the random samples for future use; for example, each time you save the results using the Save Results dialog box, the program generates a new random sample.



```
Random Numbers
--------------
To generate random numbers from the selected distribution,
use the save button on the analysis toolbar.

Random numbers to be generated: 100
```

Figure 8-63. Random Numbers for Normal Distribution

Use the *Random Numbers Options* dialog box to enter the number of observations that will be included in the random sample.

Graphical Options

Density/Mass Function Plot

The Density/Mass Function Plot option creates a plot of the probability density function for the distribution you are evaluating (see Figure 8-64). The height of the function indicates the probability of obtaining various values from the distribution you selected.

CDF Plot

The CDF Plot option creates a plot of the cumulative probability distribution for the distribution you are evaluating (see Figure 8-65).

Survivor Function Plot

The Survivor Function Plot option creates a plot of the survival probability function for the distribution you are evaluating (see Figure 8-66). The function indicates the probability of obtaining a value greater than or equal to the values on the X-axis.

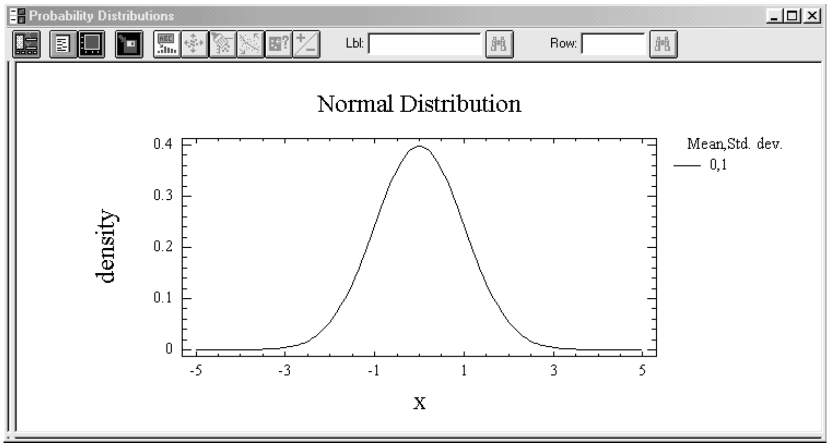


Figure 8-64.   Density/Mass Function Plot

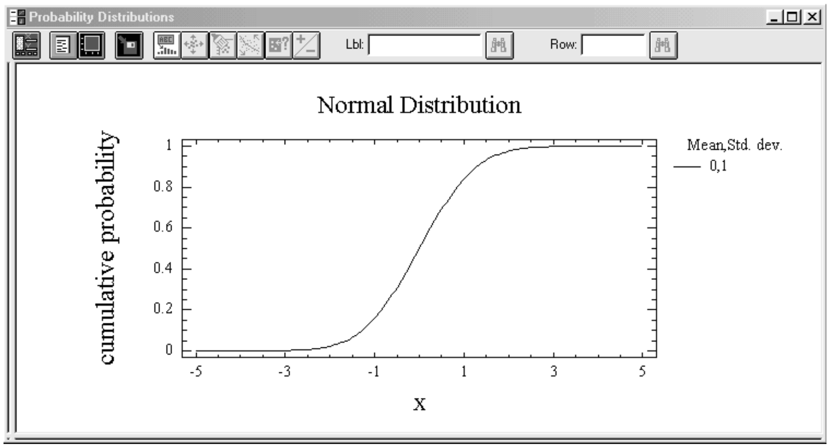


Figure 8-65.   CDF Plot

Log Survivor Function Plot

The Log Survivor Function option creates a plot of the log survival probability function for the distribution you are evaluating (see Figure 8-67). The function indicates the probability of obtaining a value greater than or equal to the values on the X-axis.

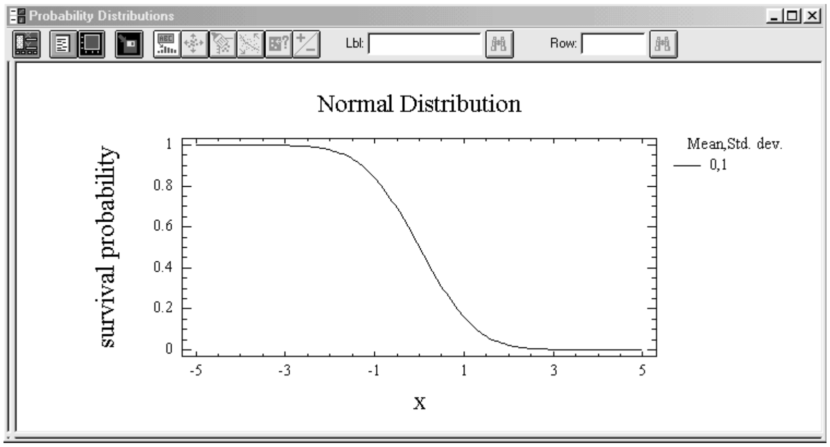


Figure 8-66. Survivor Function Plot

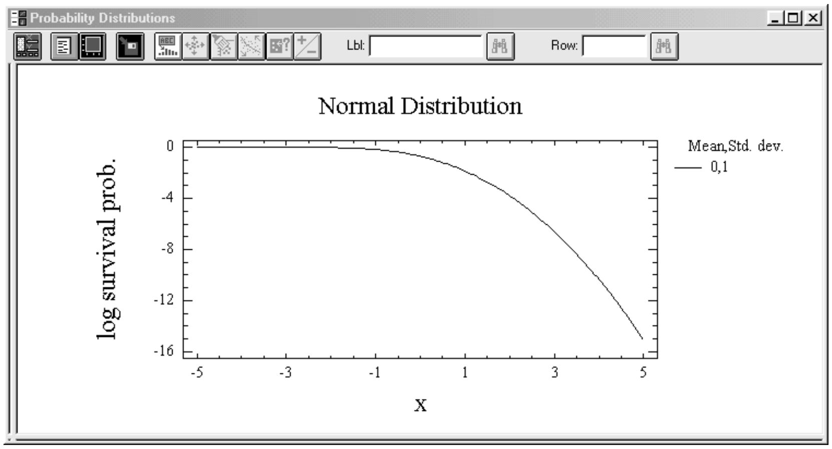


Figure 8-67. Log Survivor Function Plot

Hazard Function Plot

The Hazard Function option creates a plot of the hazard function for the distribution you are evaluating (see Figure 8-68). The hazard function is equal to the probability density function divided by the survival function. When you are modeling lifetime data, the hazard function represents the instantaneous failure rate.

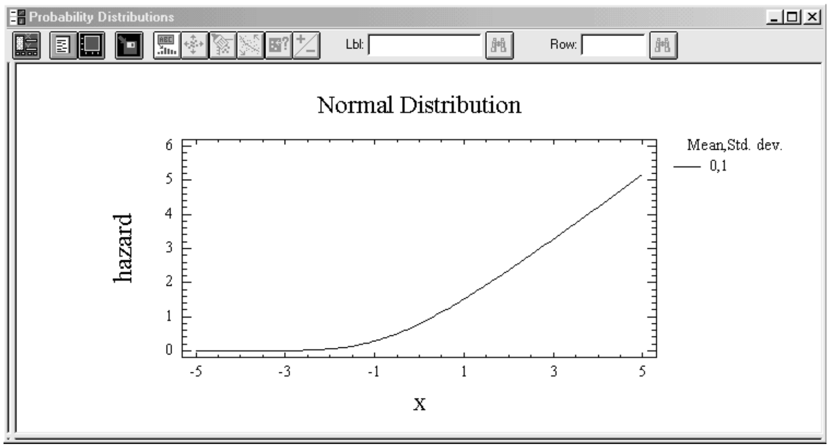


Figure 8-68. Hazard Function Plot

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There is one selection: Random Numbers for Dist. 1.

You can also use the Target Variables text box to enter the name of the variable in which you want to save the value generated during the analysis. You can enter a new name or accept the default.

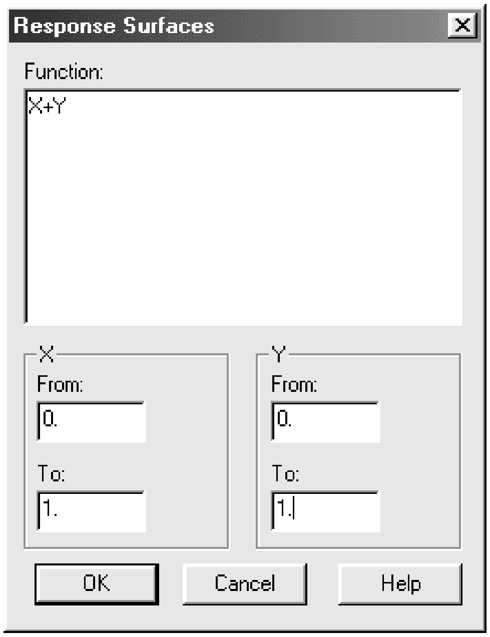
Using the Response Surface Plots Analysis

The Response Surfaces Plot Analysis allows you to create Surface and Contour plots for functions that you specify. You must write the functions involving X and Y; for example,

10+2*X+3*Y-10*X*Y

The plots are helpful when you need to visualize the shape of a response surface and when you need to locate optimal settings.

To access the analysis, from the menus, choose: PLOT... RESPONSE SURFACES... to display the Response Surfaces Analysis dialog box, which allows you to enter a function (see Figure 8-69).



Response Surfaces

Function: X+Y

X

From: 0.

To: 1.

Y

From: 0.

To: 1.

Figure 8-69. The Response Surfaces Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option displays the function that will be plotted.

Graphical Options

Surface Plot

The Surface Plot option creates a three-dimensional plot of the function (see Figure 8-70).

Use the *Surface Plot Options* dialog box to enter the number of horizontal and vertical lines that will appear on the plot, to determine the type of plot, to enter a number for the density values that will be calculated, to indicate if a Contour Plot should be displayed below the Surface Plot, to determine the values for the attributes of the contours, and to determine if lines will connect the points and if the regions of the contours will be painted (see Figure 8-71).

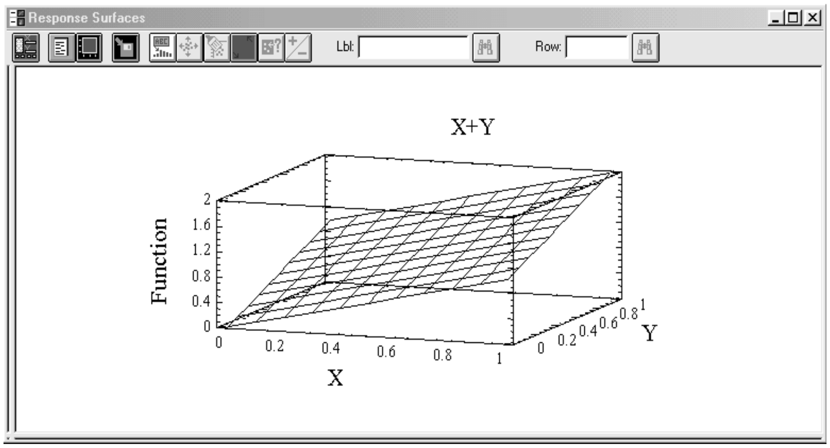
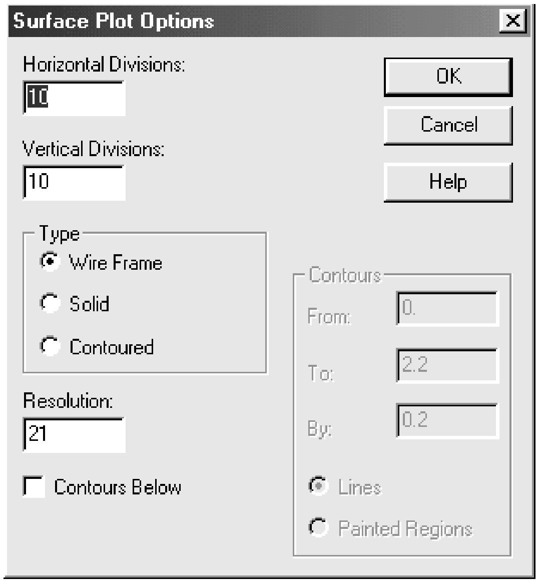


Figure 8-70. Surface Plot



Surface Plot Options

Horizontal Divisions: 10

Vertical Divisions: 10

Type
- ● Wire Frame
- ○ Solid
- ○ Contoured

Resolution: 21

□ Contours Below

Contours
- From: 0.
- To: 2.2
- By: 0.2
- ● Lines
- ○ Painted Regions

Figure 8-71. Surface Plot Options Dialog Box

You can also use the Smooth/Rotate button on the Analysis toolbar to change the angle from which you view this plot.

Contour Plot

The Contour Plot option creates a two-dimensional plot of the function (see Figure 8-72).

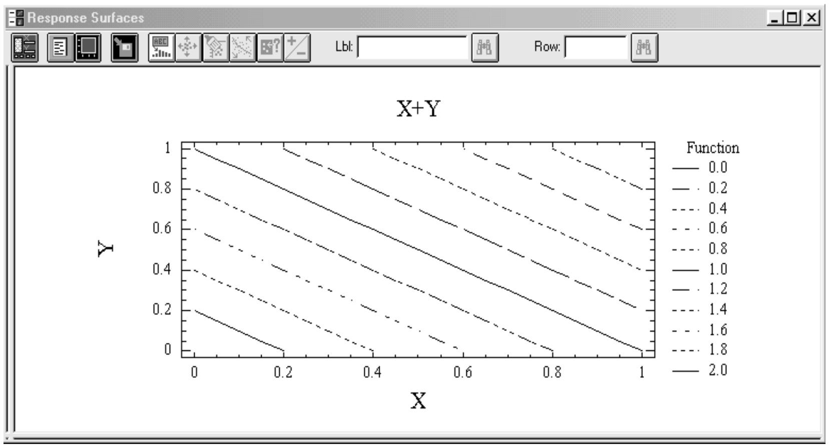
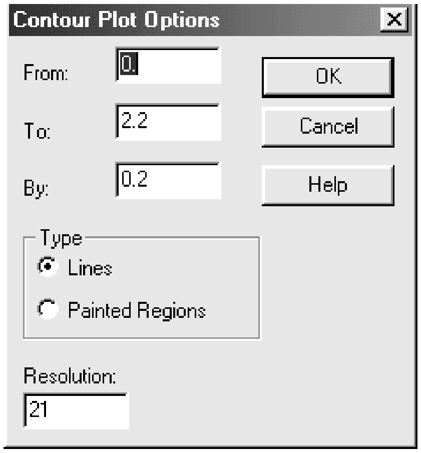


Figure 8-72. Contour Plot

Use the *Contour Plot Options* dialog box to enter values for the first and last contour lines, to enter a value for the spacing of the contour lines, to indicate if lines or painted regions will appear on the plot, and to enter a value that designates the number of density values that will be calculated (see Figure 8-73).



From: 0.

To: 2.2

By: 0.2

Type

● Lines

○ Painted Regions

Resolution: 21

Figure 8-73. Contour Plot Options Dialog Box

References

Belsley, D. A., Kuh, E., and Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity*. New York: Wiley.

Box, G. E. P. and Draper, N. R. 1987. *Empirical Model-Building and Response Surfaces*. New York: Wiley.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. New York: Wiley.

Cornell, J. A. 1973. "Experiments with Mixtures: A Review," *Technometrics*, **15**:437-455.

Cornell, J. A. 1990. *Experiments with Mixtures*, second edition. New York: Wiley & Sons.

Cornell, J. A. and Piepel, G. F. 1993. *Design and Analysis of Mixture Experiments*. Computer Associates.

Haaland, P. 1989. *Experimental Design in Biotechnology*. New York: Marcel Dekker.

Montgomery, D. C. 1991. *Design and Analysis of Experiments*, third edition. New York: John Wiley & Sons.

Using the Custom Charts Analysis

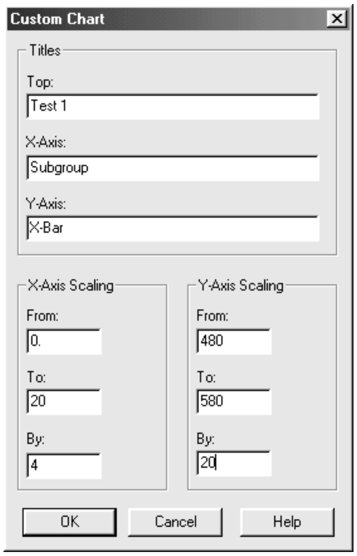
The Custom Charts Analysis allows you to create charts with custom titles, scaling, and horizontal and vertical lines. Its primary purpose is to let you create charts you can use to overlay on other plots when you use the StatGallery. A control chart is an example of a customized chart you could create.

To access the analysis, from the menus, choose: PLOT... CUSTOM CHARTS... to display the Custom Chart Analysis dialog box, which allows you to enter information about the chart you are creating (see Figure 8-74).

Tabular Options

Analysis Summary

The Analysis Summary option creates instructions for customizing the chart.



Top: Test 1

X-Axis: Subgroup

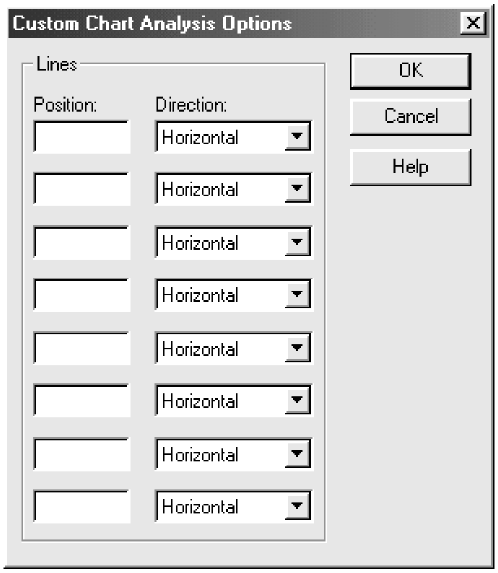
Y-Axis: X-Bar

X-Axis Scaling — From: 0. To: 20 By: 4

Y-Axis Scaling — From: 480 To: 580 By: 20

Figure 8-74. Custom Chart Analysis Dialog Box

Use the *Custom Chart Analysis Options* dialog box to choose the position and direction of up to eight lines for a customized chart (see Figure 8-75).



Position:	Direction:
	Horizontal
	Horizontal
	Horizontal
	Horizontal
	Horizontal
	Horizontal
	Horizontal
	Horizontal

Figure 8-75. The Custom Chart Analysis Options Dialog Box

Graphical Options

Custom Chart

The Custom Chart option creates a customized chart that contains the selected titles and axis scaling (see Figure 8-76).

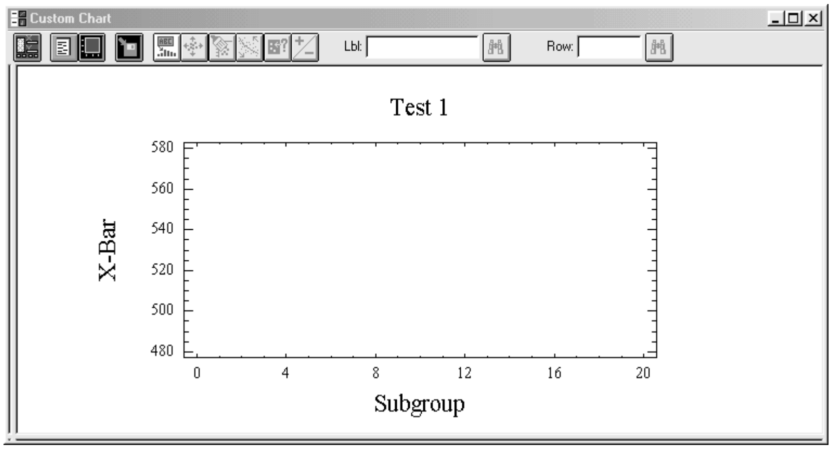


Figure 8-76. Custom Chart

References

Johnson, N. L. and Kotz, S. 1970. *Continuous Univariate Distributions - 1. New York: Wiley.*

9 Describing Numeric Data

This chapter contains information about the analyses and statistical methods you use to describe and summarize a single set of data: One-Variable Analysis, Multiple-Variable Analysis, Subset Analysis, Row-Wise Statistics, Power Transformations, Statistical Tolerance Limits, and Outlier Identification. The analyses include numeric statistics such as means, standard deviations, and frequency tabulations, as well as Box-Cox transformations and tolerance limits. Graphical summaries include types of histograms, barcharts, and probability plots.

Using the One-Variable Analysis

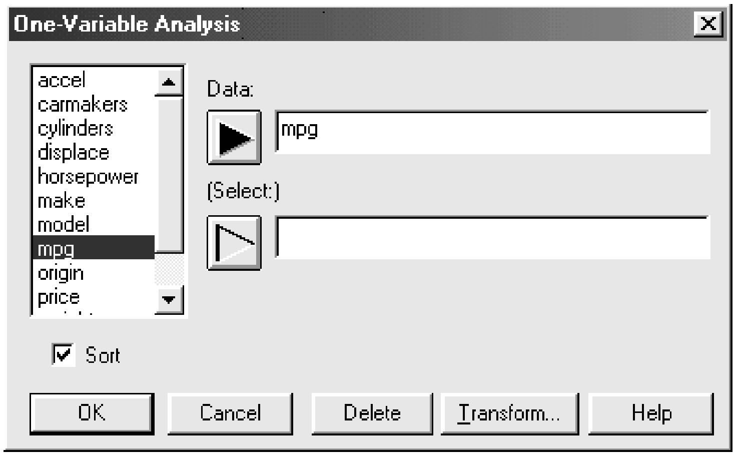
The One-Variable Analysis allows you to summarize a single column of numeric data. It calculates statistics, performs hypothesis tests, and constructs a variety of graphs. For example, options in this analysis allow you to calculate specific summary statistics, such as standard deviation and average, percentiles, and confidence intervals. You can plot the data in a scatterplot, histogram, probability plot, or symmetry plot, among others.

To access the analysis, from the menus, choose: DESCRIBE... NUMERIC DATA... ONE-VARIABLE ANALYSIS... to display the Analysis dialog box (see Figure 9-1).

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which displays the name of the selected variable, and the range of the values.



Data: mpg

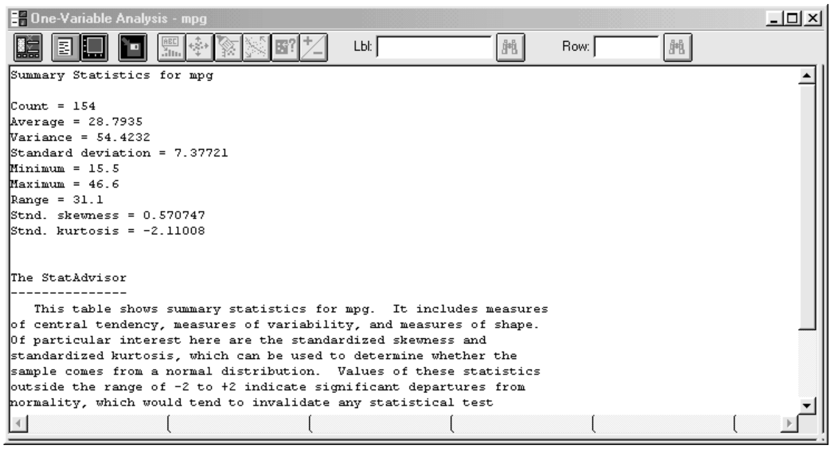
(Select:)

☑ Sort

Figure 9-1. *The One-Variable Analysis Dialog Box*

Summary Statistics

The Summary Statistics option calculates summary statistics for the variable that include measures of the center, spread, and shape of the data, such as the number of values (count), average, variance, standard deviation, minimum and maximum values, range, standardized skewness and standardized kurtosis (see Figure 9-2).

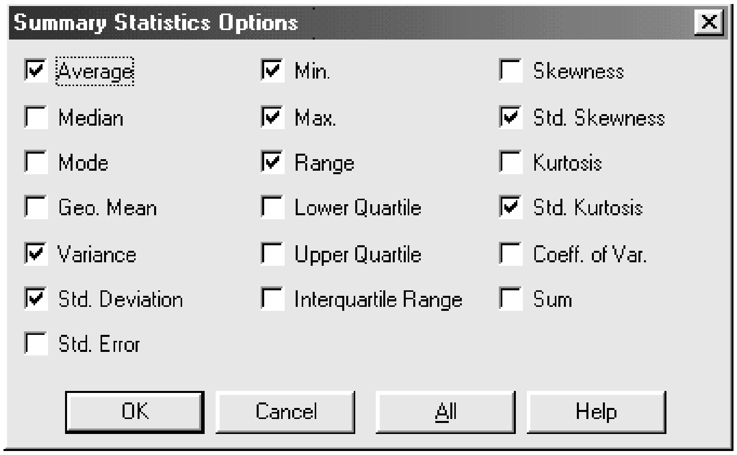


```
Summary Statistics for mpg

Count = 154
Average = 28.7935
Variance = 54.4232
Standard deviation = 7.37721
Minimum = 15.5
Maximum = 46.6
Range = 31.1
Stnd. skewness = 0.570747
Stnd. kurtosis = -2.11008


The StatAdvisor
---------------
   This table shows summary statistics for mpg.  It includes measures
of central tendency, measures of variability, and measures of shape.
Of particular interest here are the standardized skewness and
standardized kurtosis, which can be used to determine whether the
sample comes from a normal distribution.  Values of these statistics
outside the range of -2 to +2 indicate significant departures from
normality, which would tend to invalidate any statistical test
```

Figure 9-2. *Summary Statistics*

This information is helpful when you need to determine if other statistical analyses might be more appropriate to use with the data, or when you need to determine if you should transform the data.

Use the *Summary Statistics Options* dialog box to choose the statistics you want calculated (see Figure 9-3); the figure shows the defaults.



☑ Average
☐ Median
☐ Mode
☐ Geo. Mean
☑ Variance
☑ Std. Deviation
☐ Std. Error
☑ Min.
☑ Max.
☑ Range
☐ Lower Quartile
☐ Upper Quartile
☐ Interquartile Range
☐ Skewness
☑ Std. Skewness
☐ Kurtosis
☑ Std. Kurtosis
☐ Coeff. of Var.
☐ Sum

Figure 9-3. Summary Statistics Options Dialog Box

Percentiles

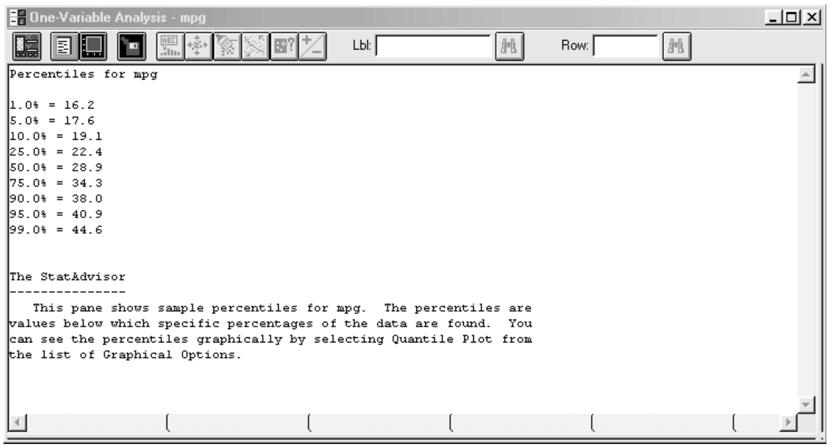
The Percentiles option calculates percentiles for the data (see Figure 9-4). The results show the percent of the values that are equal to or less than a given value.

Use the *Percentiles Options* dialog box to enter values for other percentiles (see Figure 9-5); the figure shows the defaults.

Frequency Tabulation

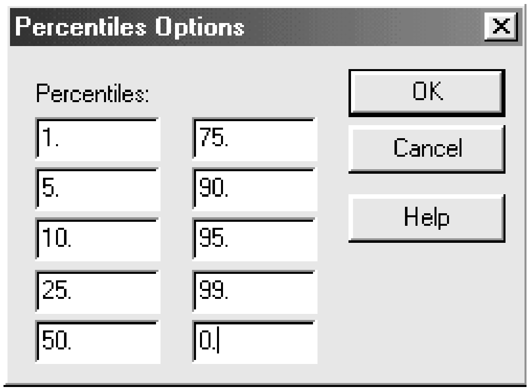
The Frequency Tabulation option groups the data into a number of class intervals, and produces a table of frequencies and cumulative frequencies that summarizes the distribution of the data (see Figure 9-6).

Use the *Frequency Tabulation Options* dialog box to choose options for the report (see Figure 9-7). You can enter the number of classes into which the



```
Percentiles for mpg

1.0% = 16.2
5.0% = 17.6
10.0% = 19.1
25.0% = 22.4
50.0% = 28.9
75.0% = 34.3
90.0% = 38.0
95.0% = 40.9
99.0% = 44.6


The StatAdvisor
---------------
   This pane shows sample percentiles for mpg.  The percentiles are
values below which specific percentages of the data are found.  You
can see the percentiles graphically by selecting Quantile Plot from
the list of Graphical Options.
```

Figure 9-4.   Percentiles

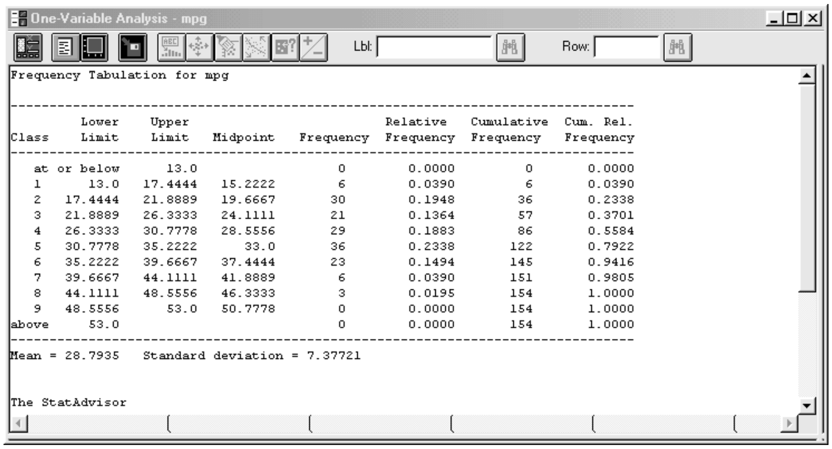


Percentiles:

1.	75.
5.	90.
10.	95.
25.	99.
50.	0.

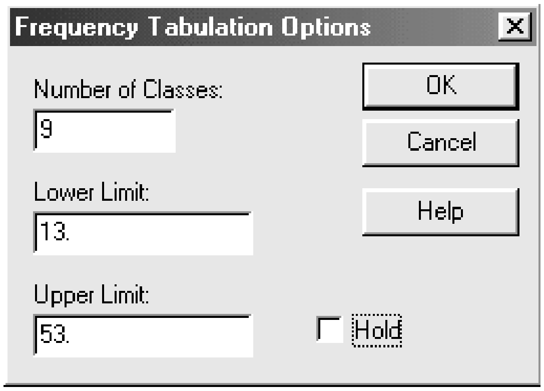
Figure 9-5.   Percentiles Options Dialog Box

data will be grouped, enter values for the Lower and Upper limits, and indicate if the current scaling should be retained.



```
Frequency Tabulation for mpg

----------------------------------------------------------------------------------
         Lower    Upper                           Relative   Cumulative  Cum. Rel.
Class    Limit    Limit   Midpoint    Frequency  Frequency  Frequency    Frequency
----------------------------------------------------------------------------------
   at or below      13.0                   0        0.0000         0        0.0000
   1      13.0   17.4444   15.2222         6        0.0390         6        0.0390
   2   17.4444   21.8889   19.6667        30        0.1948        36        0.2338
   3   21.8889   26.3333   24.1111        21        0.1364        57        0.3701
   4   26.3333   30.7778   28.5556        29        0.1883        86        0.5584
   5   30.7778   35.2222      33.0        36        0.2338       122        0.7922
   6   35.2222   39.6667   37.4444        23        0.1494       145        0.9416
   7   39.6667   44.1111   41.8889         6        0.0390       151        0.9805
   8   44.1111   48.5556   46.3333         3        0.0195       154        1.0000
   9   48.5556      53.0   50.7778         0        0.0000       154        1.0000
above     53.0                             0        0.0000       154        1.0000
----------------------------------------------------------------------------------
Mean = 28.7935   Standard deviation = 7.37721


The StatAdvisor
```

Figure 9-6. Frequency Tabulations



Number of Classes: 9

Lower Limit: 13.

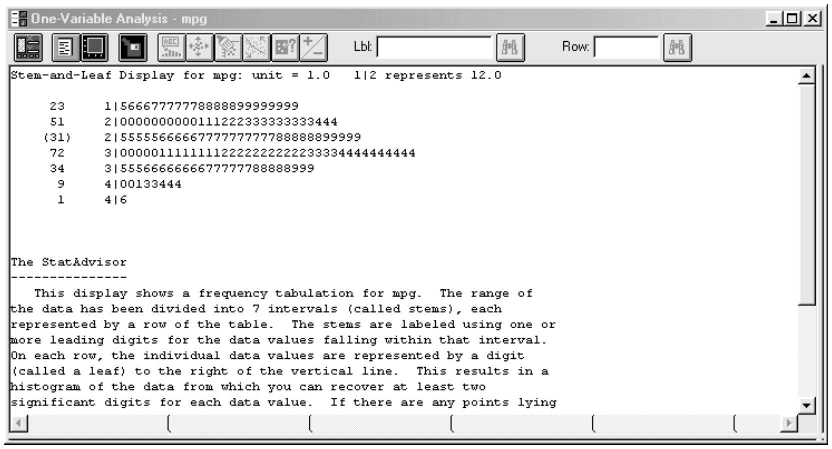
Upper Limit: 53.

□ Hold

Figure 9-7. Frequency Tabulation Options Dialog Box

Stem-and-Leaf Display

The Stem-and-Leaf Display option calculates the range and concentration of the values, the symmetry of the data, and indicates if there are gaps or outliers (see Figure 9-8).



```
Stem-and-Leaf Display for mpg: unit = 1.0   1|2 represents 12.0

     23     1|56667777778888899999999
     51     2|0000000000111222333333333444
    (31)    2|5555566666777777777778888889999
     72     3|00000111111112222222222222333444444444
     34     3|5556666666667777788888999
      9     4|00133444
      1     4|6


The StatAdvisor
---------------
   This display shows a frequency tabulation for mpg.  The range of
the data has been divided into 7 intervals (called stems), each
represented by a row of the table.  The stems are labeled using one or
more leading digits for the data values falling within that interval.
On each row, the individual data values are represented by a digit
(called a leaf) to the right of the vertical line.  This results in a
histogram of the data from which you can recover at least two
significant digits for each data value.  If there are any points lying
```

Figure 9-8. Stem-and-Leaf Display

Use the *Stem-and-Leaf Display Options* dialog box if you do not want the program to flag outliers; the default is Flag Outliers.

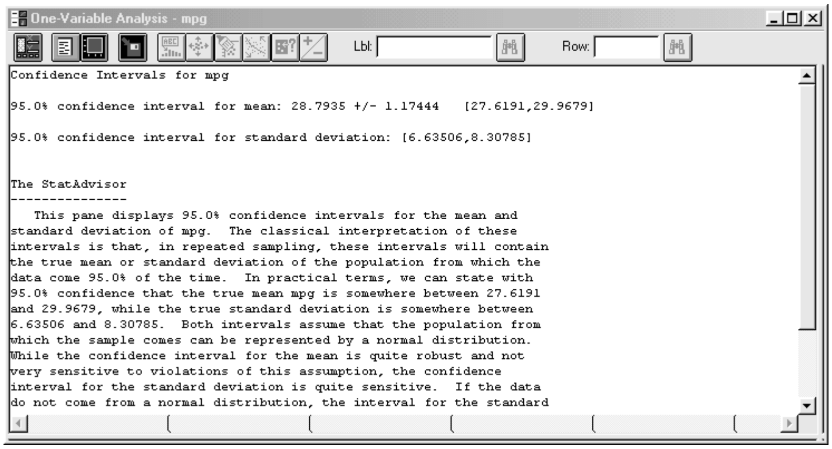
Confidence Intervals

The Confidence Intervals option calculates confidence intervals for the mean and standard deviation of the variable (see Figure 9-9). You can assume, with 95 percent confidence, that the true population mean and population standard deviation lie within these respective intervals.

Use the *Confidence Intervals Options* dialog box to enter other values for the confidence level (see Figure 9-10); the default is 95 percent.

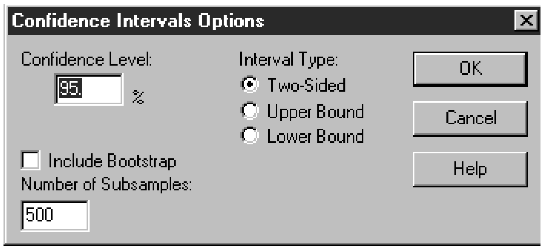
The Confidence Intervals tabular option now allows for both one-sided bounds and two-sided intervals. In addition, bootstrap intervals (or bounds) for the mean, standard deviation, and median have been added as an option.

Bootstrap intervals are formed by selecting k random subsamples of n observations each with replacement for the n data values, calculating each statistic, and then displaying percentiles for the k values of the computed statistics. When standard assumptions are violated, such as normality of the data values, the resulting intervals may be better than those usually computed.



```
Confidence Intervals for mpg

95.0% confidence interval for mean: 28.7935 +/- 1.17444   [27.6191,29.9679]

95.0% confidence interval for standard deviation: [6.63506,8.30785]


The StatAdvisor
---------------
   This pane displays 95.0% confidence intervals for the mean and
standard deviation of mpg.  The classical interpretation of these
intervals is that, in repeated sampling, these intervals will contain
the true mean or standard deviation of the population from which the
data come 95.0% of the time.  In practical terms, we can state with
95.0% confidence that the true mean mpg is somewhere between 27.6191
and 29.9679, while the true standard deviation is somewhere between
6.63506 and 8.30785.  Both intervals assume that the population from
which the sample comes can be represented by a normal distribution.
While the confidence interval for the mean is quite robust and not
very sensitive to violations of this assumption, the confidence
interval for the standard deviation is quite sensitive.  If the data
do not come from a normal distribution, the interval for the standard
```

Figure 9-9. Confidence Intervals



Confidence Level: 95. %

Interval Type:
● Two-Sided
○ Upper Bound
○ Lower Bound

□ Include Bootstrap

Number of Subsamples: 500

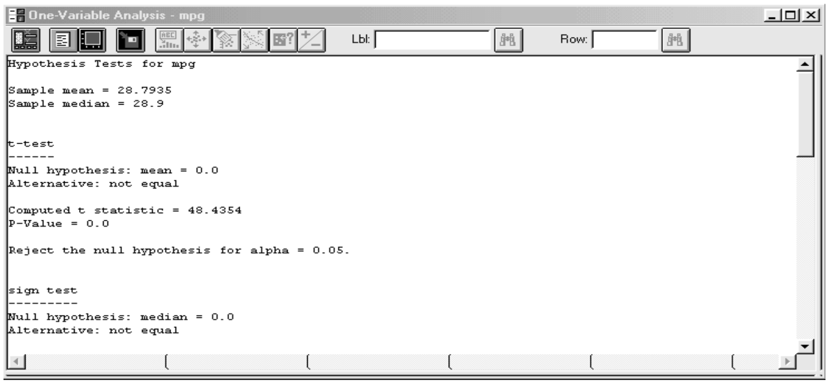
Figure 9-10. Confidence Intervals Options Dialog Box

Hypothesis Tests

The Hypothesis Tests option calculates the results of three tests that are derived from the sample's center population: a *t*-statistic, a sign test, or a signed rank test (see Figure 9-11).

The first test is a *t*-test of the null hypothesis that the mean of the variable equals a given value versus the alternative hypothesis that the mean of the variable is not equal to a given value.

The second test is a sign test of the null hypothesis that the median of the variable equals a given value versus the alternative hypothesis that the median of the variable is not equal to a given value.

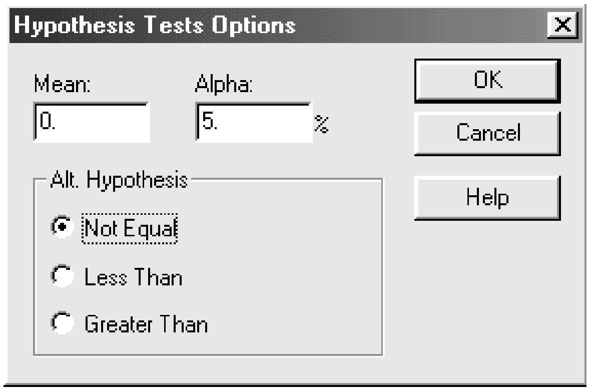


```
Hypothesis Tests for mpg

Sample mean = 28.7935
Sample median = 28.9


t-test
------
Null hypothesis: mean = 0.0
Alternative: not equal

Computed t statistic = 48.4354
P-Value = 0.0

Reject the null hypothesis for alpha = 0.05.


sign test
---------
Null hypothesis: median = 0.0
Alternative: not equal
```

Figure 9-11. Hypothesis Tests

The third test is a signed rank test of the null hypothesis that the median of the variable equals a given value versus the alternative hypothesis that the median of the variable is not equal to a given value. It is calculated by comparing the average ranks of values above and below the hypothesized median. The sign and signed rank tests are less sensitive to the presence of outliers, but are somewhat less effective than the *t*-test if all the data are from a single normal distribution.

Use the *Hypothesis Tests Options* dialog box to modify the null hypothesis and test options (see Figure 9-12); the figure shows the defaults.



Mean: 0.

Alpha: 5. %

Alt. Hypothesis

● Not Equal

○ Less Than

○ Greater Than

Figure 9-12. Hypothesis Tests Options Dialog Box

Graphical Options

Scatterplot

The Scatterplot option creates a univariate scatterplot of the values for the variable along a single axis as point symbols with no connecting lines (see Figure 9-13). *For more information about a scatterplot, see the Scatterplot graphical option for the Univariate Plot Analysis in Chapter 8, Using Basic Plots.*

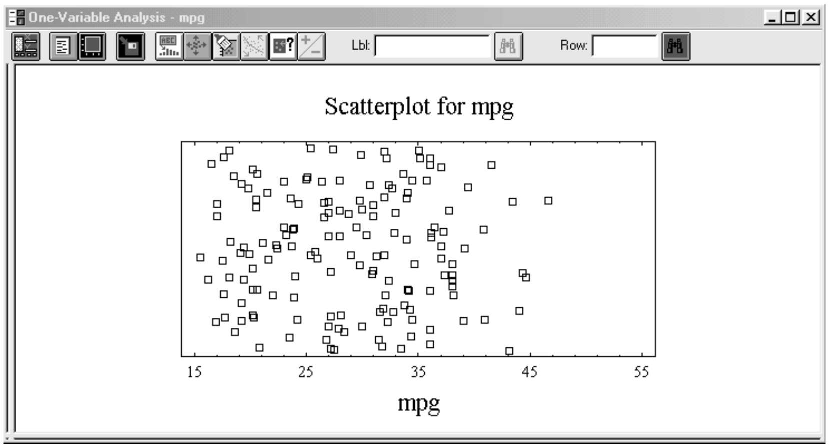


Figure 9-13. Scatterplot

The plot allows you to easily identify the range of the data; however, it may be difficult to identify individual points if they overlap. To reduce the amount of overplotting, you can *jitter* the points using the Jittering button on the Analysis toolbar.

Box-and-Whisker Plot

The Box-and-Whisker Plot option creates a plot of the data, which is divided into four equal areas of frequency (quartiles) (see Figure 9-14) . A box encloses the middle 50 percent, where the median is drawn as a vertical line inside the box.

Horizontal lines, known as whiskers, extend from each end of the box. The left (or lower) whisker is drawn from the lower quartile to the smallest point

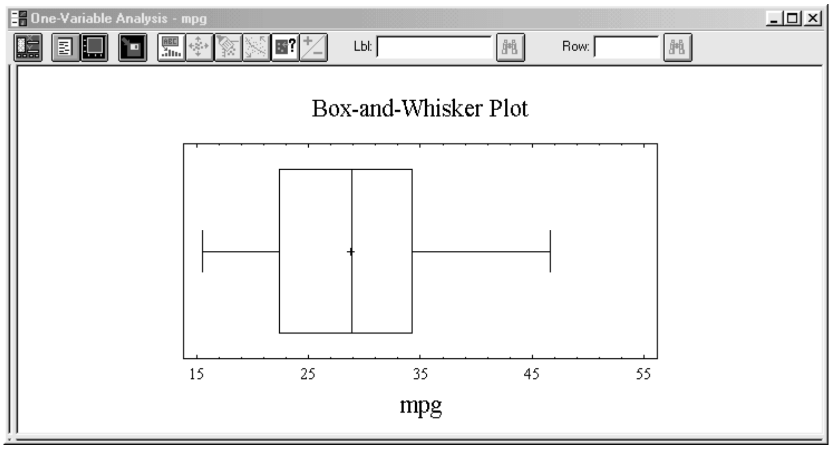


Figure 9-14. Box-and-Whisker Plot

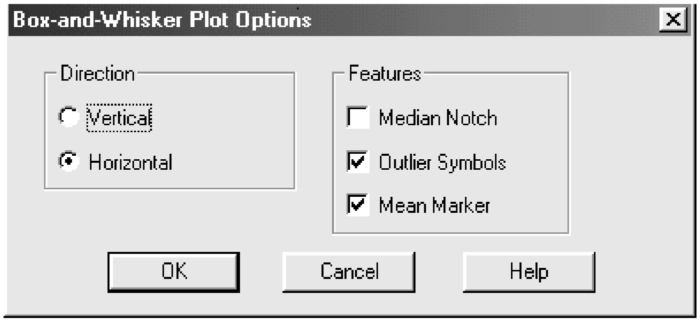
within 1.5 interquartile ranges from the lower quartile. The other whisker is drawn from the upper quartile to the largest point within 1.5 interquartile ranges from the upper quartile.

Use the *Box-and-Whisker Plot Options* dialog box to indicate if the plot will appear in a vertical or horizontal direction, and to choose features for the plot such as median notch, outlier symbols, and mean marker (see Figure 9-15); the figure shows the defaults.

Frequency Histogram

The Frequency Histogram option creates a Frequency Histogram of the data (see Figure 9-16). The program divides the data into sets of nonoverlapping intervals and plots bars for each interval. The height of each bar is proportional to the number of observations that fall within that interval, which allows you to see the distribution of the data.

Use the *Frequency Plot Options* dialog box to enter values for number of classes into which the data will be grouped, as well as for the Lower limit for the first class and the Upper limit for the last class. You can also indicate if the scale for the Y-axis will be relative or cumulative, if the current scaling will be retained, and whether you want to create a histogram or a polygon (see Figure 9-17); the figure shows the defaults.



Direction

○ Vertical

● Horizontal

Features

□ Median Notch

☑ Outlier Symbols

☑ Mean Marker

Figure 9-15. Box-and-Whisker Plot Options Dialog Box

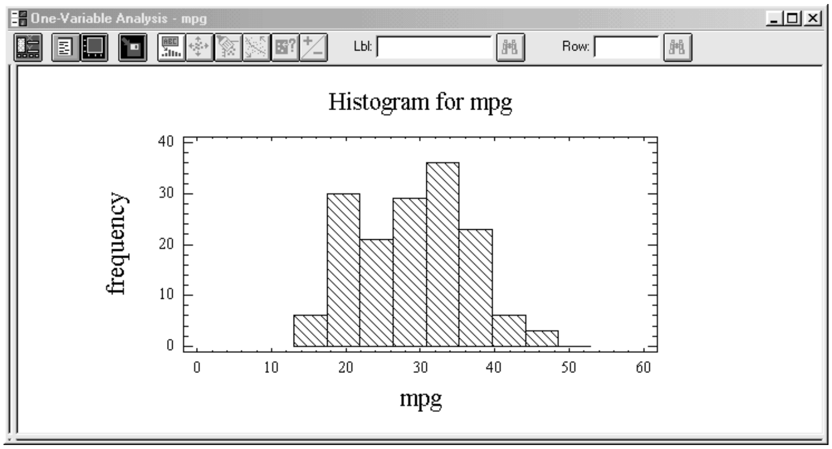
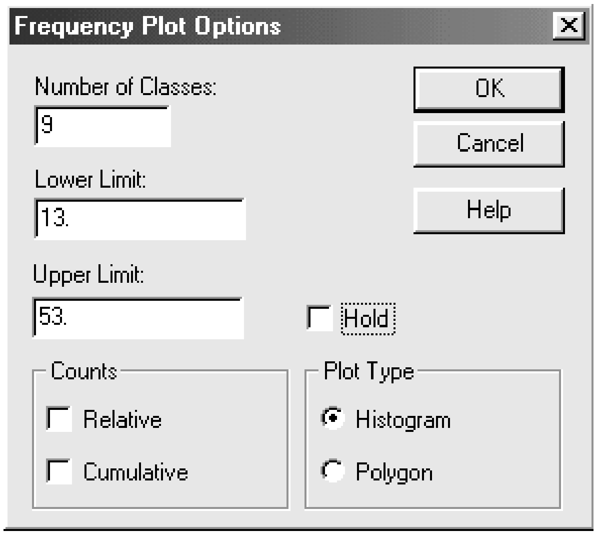


Figure 9-16. Frequency Histogram

Note: When you make changes using this dialog box, the changes also affect the Frequency Tabulation tabular option.

Quantile Plot

The Quantile Plot option creates a plot of the percentiles of the data (see Figure 9-18). The Y-axis represents the proportion of values that are below a particular value.



Frequency Plot Options

Number of Classes: 9

Lower Limit: 13.

Upper Limit: 53.

☐ Hold

Counts

☐ Relative

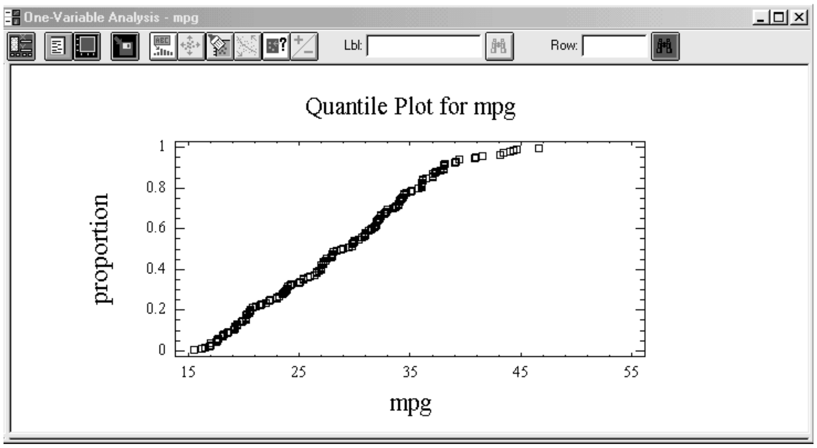
☐ Cumulative

Plot Type

● Histogram

○ Polygon

Figure 9-17. Frequency Plot Options Dialog Box



One-Variable Analysis - mpg

Figure 9-18. Quantile Plot

Normal Probability Plot

The Normal Probability Plot option helps to determine if the data come from a normal distribution (see Figure 9-19). The plot consists of an arithmetic (interval) horizontal axis scaled for the data, and a vertical axis scaled so the cumulative distribution function of a normal distribution plots as a straight line. The closer the data are to being on a straight line, the more likely they follow a normal distribution. Significant curvature of the data indicates skewness.

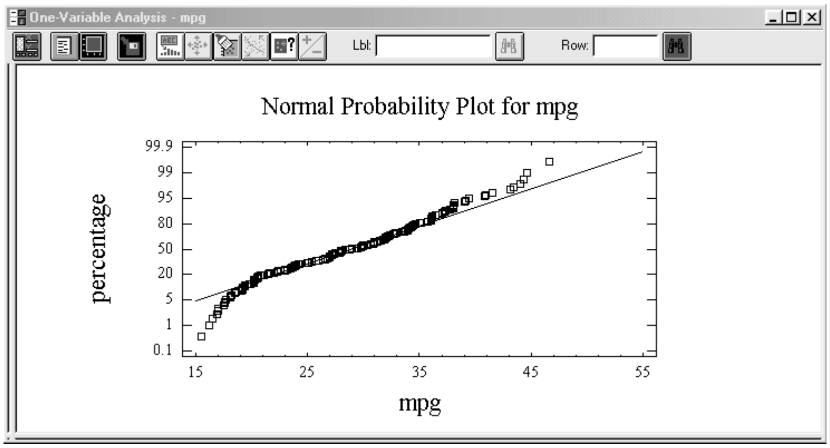
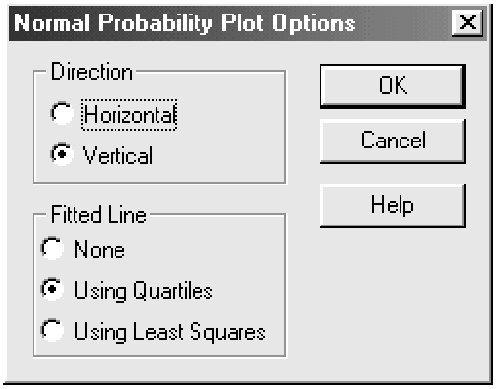


Figure 9-19.   Normal Probability Plot

Use the *Normal Probability Plot Options* dialog box to indicate the direction of the plot, to indicate if you want a fitted line on the plot and, if so, to choose the method that will be used to calculate it (see Figure 9-20); the figure shows the defaults.

Density Trace

The Density Trace option creates a plot you can use to view the shape or distribution of the data, especially the variations in density over the range of the data (see Figure 9-21). Unlike a histogram, the Density Trace uses overlapping intervals and a weight function to smooth the densities, resulting in a continuous line rather than a group of rectangles as in a histogram.



Normal Probability Plot Options

Direction

○ Horizontal

● Vertical

Fitted Line

○ None

● Using Quartiles

○ Using Least Squares

Figure 9-20. Normal Probability Plot Options Dialog Box

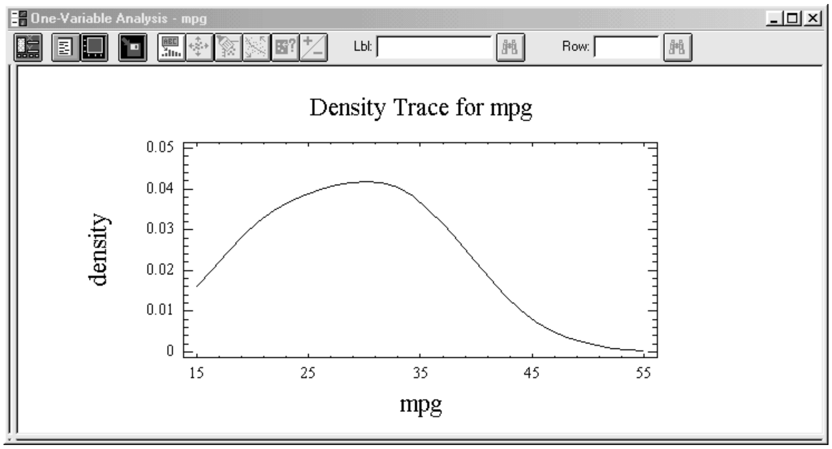
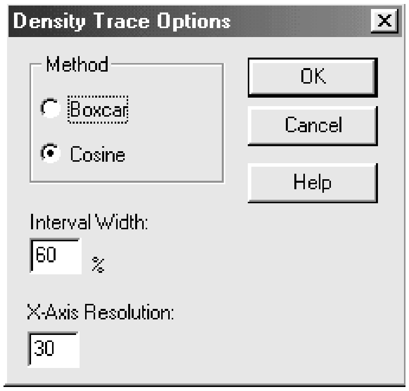


Figure 9-21. Density Trace

Use the *Density Trace Options* dialog box to choose the method that will be used to estimate the density function, to enter a value for the degree of overlap that will be used to compute the density trace, and to enter the number of density values that will be calculated (see Figure 9-22); the figure shows the defaults.



Density Trace Options

Method

○ Boxcar

● Cosine

Interval Width: 60 %

X-Axis Resolution: 30

Figure 9-22.   Density Trace Options Dialog Box

Symmetry Plot

The Symmetry Plot option creates a plot that is helpful for determining the symmetry of the data (see Figure 9-23). Use this plot if you want to use analyses that are best applied to data from a symmetrical distribution. A symmetrical dataset will closely follow the reference line.

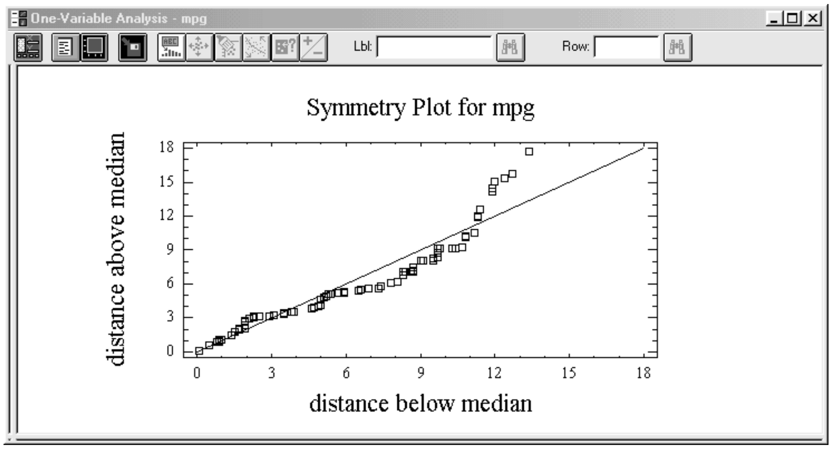


Figure 9-23.   Symmetry Plot

Saving the Results

Use the Save Results Options dialog box to choose the results you want to save. There are six selections: Summary Statistics, Percentiles, Frequencies, Cumulative Frequencies, Relative Frequencies, and Cumulative Relative Frequencies.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. Belmont, California: Wadsworth International Group.

Frigge, M., Hoagland, D. C., and Iglewicz, B. 1989. "Some Implementations of the Boxplot," *American Statistician*, **43**: 50-54.

Efron, Bradley and Tibshirani, Robert J. 1994. *An Introduction to the Bootstrap*. Chapman and Hall.

Escobar, Luis A. And Meeker, William Q. 1998. *Statistical Methods for Reliability Data*. New York: Wiley.

Guttman, I., Wilks, S. S., and Hunter, J. S. 1982. *Introductory Engineering Statistics*, third edition. New York: Wiley.

Lapin, L. L. 1987. *Statistics for Modern Business Decisions*. Orlando, Florida: Harcourt Brace Jovanovich, Inc.

McGill, R., Tukey, J. W., and Larsen, W. A. 1978. "Variation of Box Plots," *American Statistician*, **32**:12-16.

Snedecor, G. W. and Cochran, W. G. 1967. *Statistical Methods*, sixth edition. Ames, Iowa: Iowa State University Press.

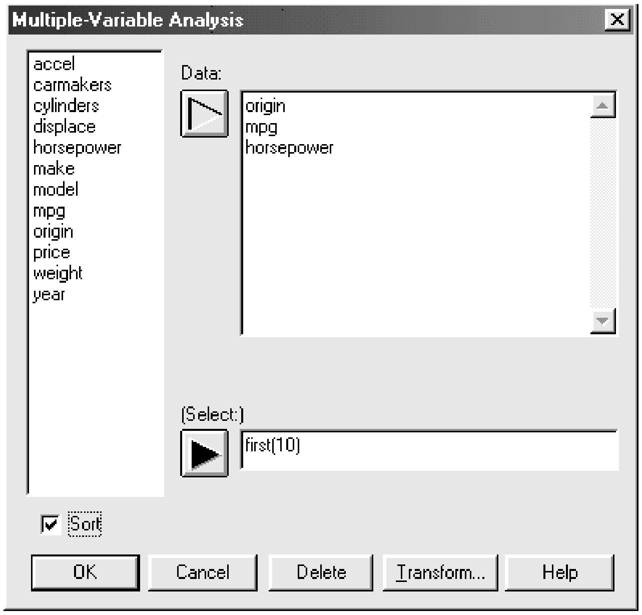
Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Belmont, California: Duxbury Press.

Using the Multiple-Variable Analysis

The Multiple-Variable analysis allows you to summarize several columns of quantitative data by producing descriptive statistics for each variable. For example, the options allow you to calculate summary statistics, correlations, covariances, and partial correlations. The plots in the analysis provide different views of the data; for example, as a Scatterplot Matrix, as a Star Plot or Sunray Plot, or as a Key Glyph. The StatAdvisor provides suggestions for using this analysis to build statistical models.

To access the analysis, from the menus, choose: DESCRIBE... NUMERIC DATA... MULTIPLE-VARIABLE ANALYSIS... to display the Analysis dialog box (see Figure 9-24).



Data:
origin
mpg
horsepower

(Select:) first(10)

☑ Sort

Figure 9-24. The Multiple-Variable Analysis Dialog Box

Tabular Options

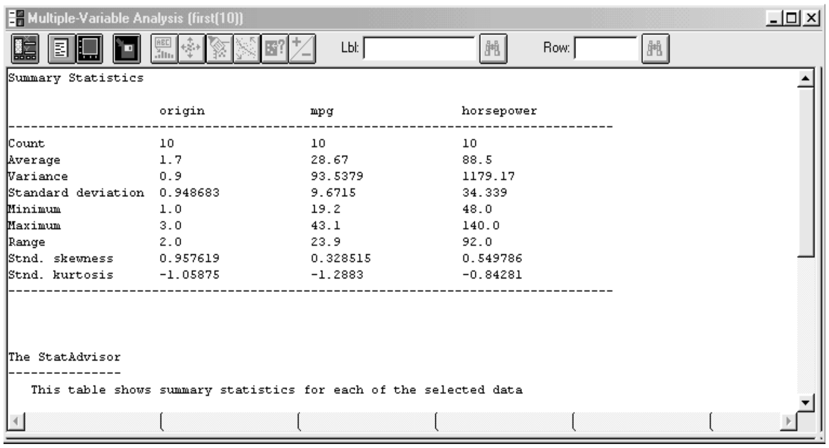
Analysis Summary

The Analysis Summary option displays the names of the selected variables, and the number of complete cases.

Use the *Multiple-Variable Options* dialog box to indicate if you want to include complete cases only or all the data in the analysis. The default is Complete Cases Only.

Summary Statistics

The Summary Statistics option calculates summary statistics for the variable that include measures of the center, spread, and shape of the data, such as the number of values (count), average, variance, standard deviation, minimum and maximum values, range, standardized skewness and standardized kurtosis (see Figure 9-25). This information is helpful when you need to determine if other statistical analyses might be appropriate to use with the data, or when you need to determine if you should transform the data.



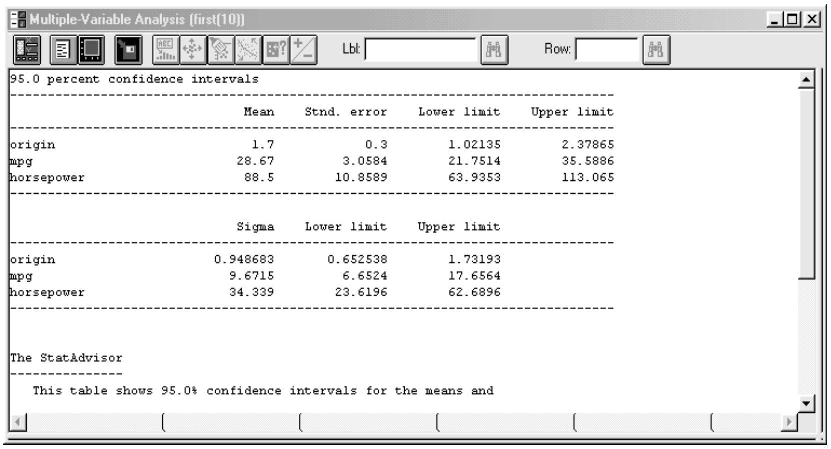
```
Summary Statistics

                    origin              mpg                 horsepower
--------------------------------------------------------------------------------
Count               10                  10                  10
Average             1.7                 28.67               88.5
Variance            0.9                 93.5379             1179.17
Standard deviation  0.948683            9.6715              34.339
Minimum             1.0                 19.2                48.0
Maximum             3.0                 43.1                140.0
Range               2.0                 23.9                92.0
Stnd. skewness      0.957619            0.328515            0.549786
Stnd. kurtosis      -1.05875            -1.2883             -0.84281
--------------------------------------------------------------------------------



The StatAdvisor
---------------
   This table shows summary statistics for each of the selected data
```

Figure 9-25.   Summary Statistics

Use the *Summary Statistics Options* dialog box to choose the statistics you want calculated (see Figure 9-3 for an example of this dialog box). The figure shows the defaults.

Confidence Intervals

The Confidence Intervals option calculates the confidence intervals (in percentages) for the mean and standard deviation of the variables (see Figure 9-26). You can assume, with 95 percent confidence, that the true population mean and population standard deviation lie within these respective intervals.



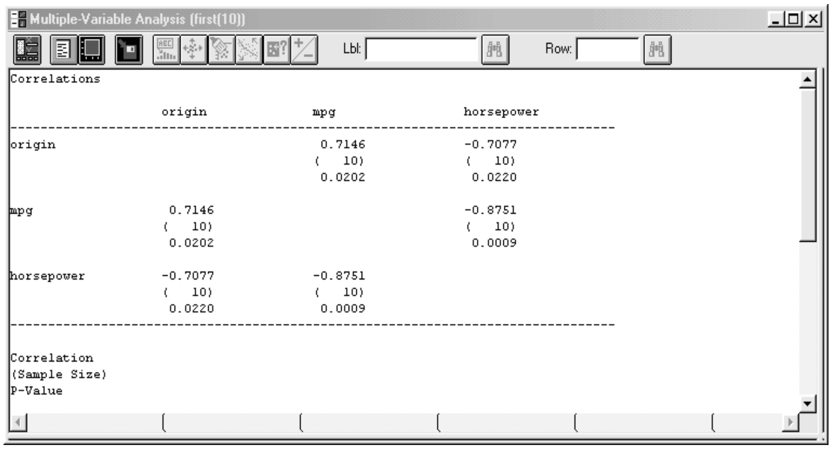
```
95.0 percent confidence intervals
--------------------------------------------------------------------------------
                              Mean    Stnd. error    Lower limit    Upper limit
--------------------------------------------------------------------------------
origin                         1.7            0.3        1.02135        2.37865
mpg                          28.67         3.0584        21.7514        35.5886
horsepower                    88.5        10.8589        63.9353        113.065
--------------------------------------------------------------------------------

                             Sigma    Lower limit    Upper limit
--------------------------------------------------------------------------------
origin                    0.948683       0.652538        1.73193
mpg                         9.6715         6.6524        17.6564
horsepower                  34.339        23.6196        62.6896
--------------------------------------------------------------------------------


The StatAdvisor
---------------
   This table shows 95.0% confidence intervals for the means and
```

Figure 9-26. Confidence Intervals

Use the *Confidence Intervals Options* dialog box to enter other values for the confidence level; the default is 95 percent.

Correlations

The Correlations option calculates a matrix of Pearson Product-Moment correlation coefficients for the observed values (see Figure 9-27). The table provides a preliminary view of the relationships among the selected variables. Each row and column in the table corresponds to a variable in the dataset.



```
Correlations

                    origin              mpg                 horsepower
--------------------------------------------------------------------------------
origin                                   0.7146              -0.7077
                                        (    10)             (    10)
                                         0.0202               0.0220

mpg                  0.7146                                  -0.8751
                    (    10)                                 (    10)
                     0.0202                                   0.0009

horsepower          -0.7077             -0.8751
                    (    10)            (    10)
                     0.0220              0.0009
--------------------------------------------------------------------------------

Correlation
(Sample Size)
P-Value
```

Figure 9-27.   Correlations

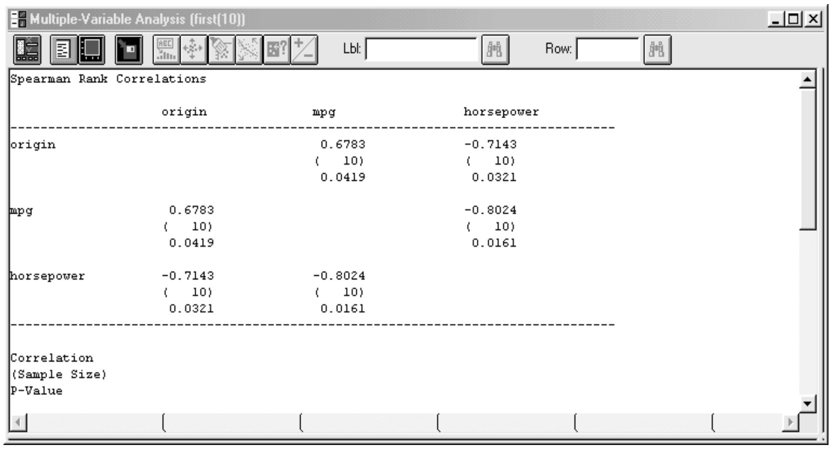
Three numbers appear in each cell of the matrix: the correlation coefficient for the two variables, represented by the cell; the sample size (in parentheses); and the *p*-value for the correlation.

Correlations range from -1 to +1. A positive correlation suggests that two variables vary in the same direction, while a negative correlation suggests that two variables vary in the opposite direction; *p*-values below .05 indicate a statistically significant correlation at the 95 percent confidence level. Statistically independent variables have an expected correlation coefficient of zero.

Rank Correlations

The Rank Correlations options calculate a matrix of Spearman rank correlation coefficients that is similar to the correlation coefficients matrix or Kendall’s Tau (see Figure 9-28) by using the Rank Correlation Options dialog box. The difference is that the ranks of the values are used rather than the values themselves.

Choose this option if you can rank order the data, but you cannot assign values on an interval scale.



```
Spearman Rank Correlations

                    origin              mpg                 horsepower
--------------------------------------------------------------------------------
origin                                   0.6783              -0.7143
                                        (    10)             (    10)
                                         0.0419               0.0321

mpg                  0.6783                                  -0.8024
                    (    10)                                 (    10)
                     0.0419                                   0.0161

horsepower          -0.7143             -0.8024
                    (    10)            (    10)
                     0.0321              0.0161
--------------------------------------------------------------------------------

Correlation
(Sample Size)
P-Value
```

Figure 9-28. Rank Correlations

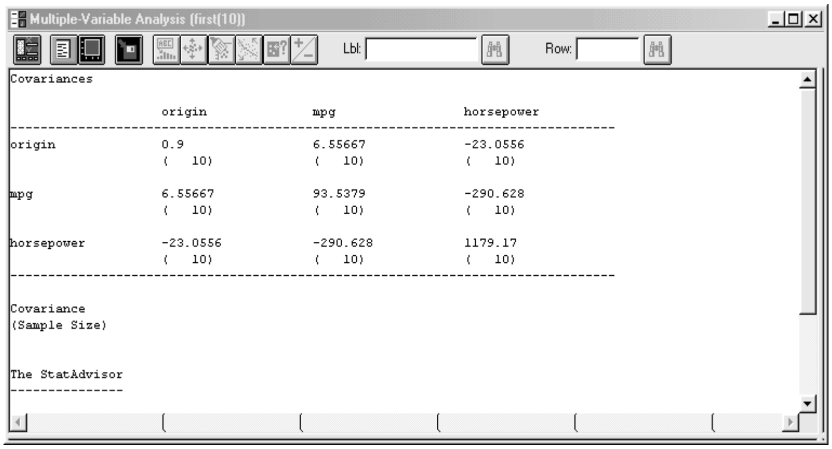
Covariances

The Covariances option calculates a matrix of the estimated covariance for each pair of variables (see Figure 9-29). Covariance measures the linear association between two variables. If the variables tend to fall above or below their means at the same time, the covariance is positive. If one variable is above its mean while the other is below, the covariance is negative. The value of the covariance is sensitive to scaling because it retains the units of the original data.

Partial Correlations

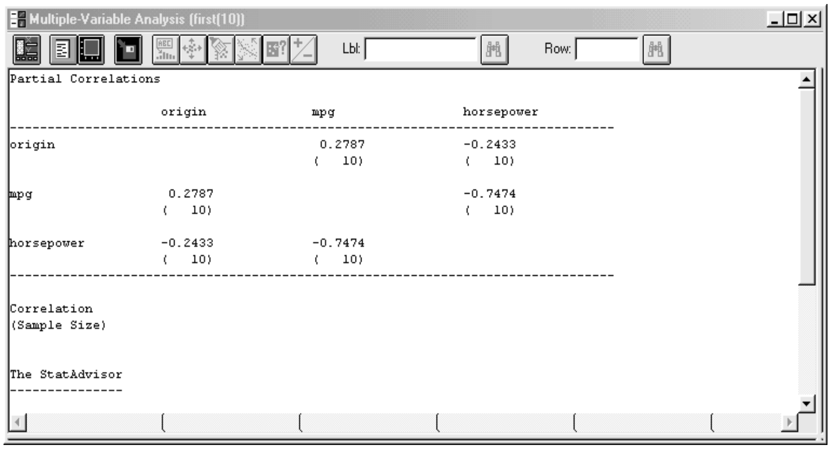
The Partial Correlations option calculates a matrix of estimated partial correlation coefficients for a set of observed values (see Figure 9-30). A partial correlation coefficient measures the relationship between two variables while controlling for possible effects of the other variables.

Removing the linear relationship controls the effects before the calculations between the two variables are made. Partial correlations are useful for uncovering hidden relationships, identifying intervening variables, and detecting false relationships.



```
Covariances

                    origin              mpg                 horsepower
--------------------------------------------------------------------------------
origin              0.9                 6.55667             -23.0556
                    (    10)            (    10)            (    10)

mpg                 6.55667             93.5379             -290.628
                    (    10)            (    10)            (    10)

horsepower          -23.0556            -290.628            1179.17
                    (    10)            (    10)            (    10)
--------------------------------------------------------------------------------

Covariance
(Sample Size)


The StatAdvisor
---------------
```

Figure 9-29. Covariances



```
Partial Correlations

                    origin              mpg                 horsepower
--------------------------------------------------------------------------------
origin                                   0.2787             -0.2433
                                        (    10)            (    10)

mpg                  0.2787                                 -0.7474
                    (    10)                                (    10)

horsepower          -0.2433             -0.7474
                    (    10)            (    10)
--------------------------------------------------------------------------------

Correlation
(Sample Size)


The StatAdvisor
---------------
```

Figure 9-30.   Partial Correlations

Graphical Options

Scatterplot Matrix

The Scatterplot Matrix option creates a matrix plot of all the two-variable scatterplots (see Figure 9-31). The Scatterplot Matrix visually represents the relationships among the pairs of variables and is helpful in identifying strongly correlated pairs.

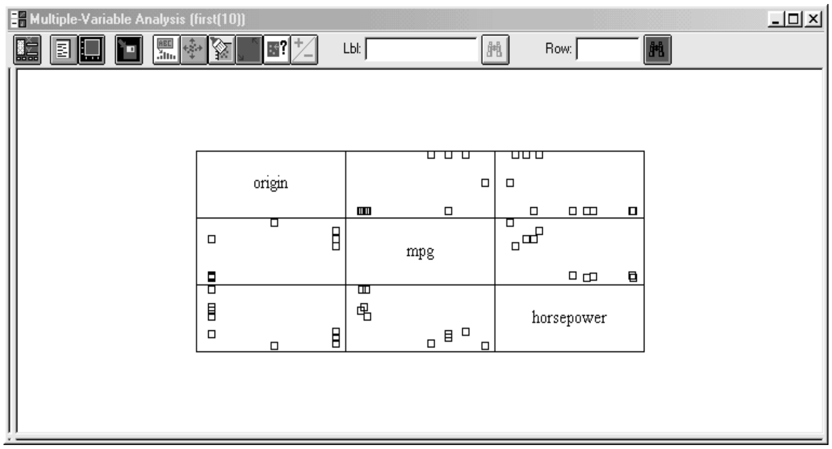


Figure 9-31. Scatterplot Matrix

Star Plot

The Star Plot option creates a graph of multivariate data that allows you to visually compare the different observations, and to see the relative values for all the variables at once (see Figure 9-32). Each star represents one observation in the data. It is an *n*-polygon drawn so the distance of the vertices from the center point represents that observation's values in relation to the other variables.

The first variable in the list is plotted starting at the 12:00 position; then the remaining variables are plotted moving in a clock-wise direction. The program uses the shortest ray to plot the smallest value in each variable and the longest ray to plot the largest value. You can plot 25 glyphs at a time.

Use the *Star Plot/Sunray Plot Options* dialog box to do the following: enter a name for the labels for the observations; choose the type of label that will appear on the plot; enter the number of the starting glyph; and indicate if you want to alphabetically sort the variables (see Figure 9-33). The figure shows the defaults.

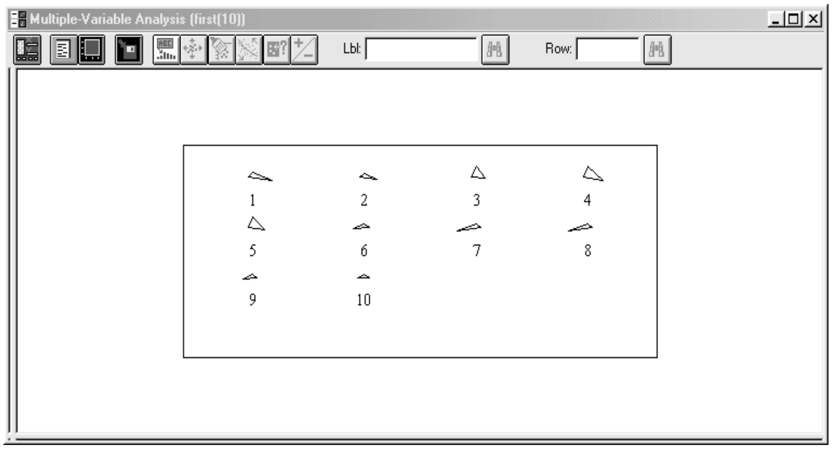
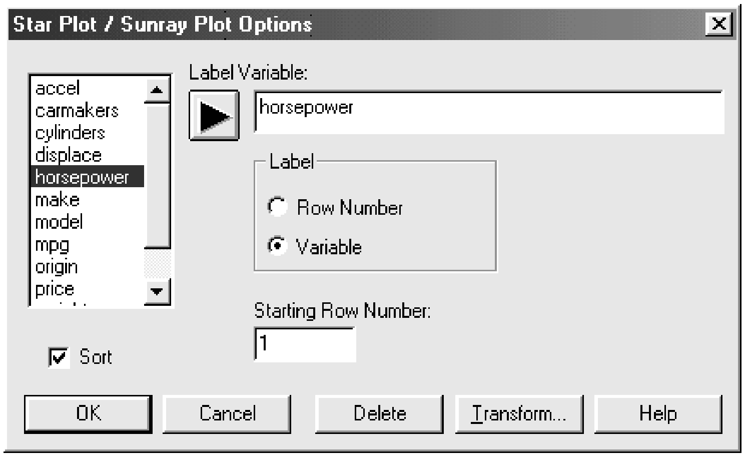


Figure 9-32. Star Plot



Label Variable: horsepower

Label

○ Row Number

● Variable

Starting Row Number: 1

☑ Sort

Figure 9-33. Star Plot/Sunray Plot Options Dialog Box

Sunray Plot

The Sunray Plot option creates a plot that lets you visually compare observations with the means for the variables (see Figure 9-34). A separate glyph is plotted for each observation in the data.

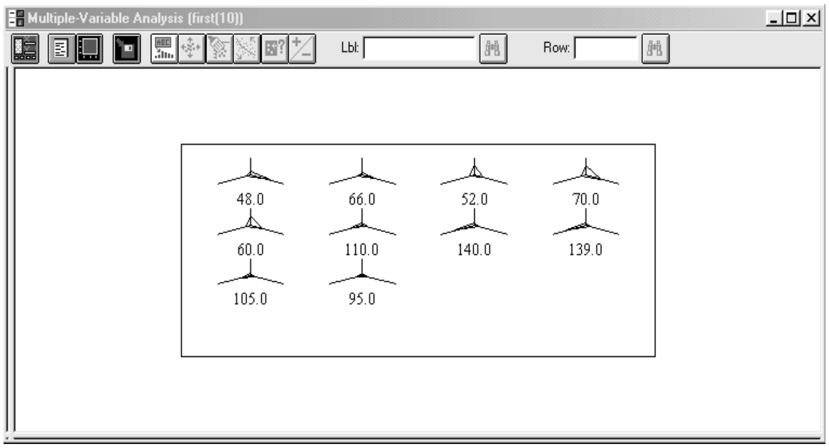


Figure 9-34. Sunray Plot

A glyph consists of a series of rays drawn from a center point and lines that connect the rays that (usually) form a polygon. Each ray is scaled so the middle represents the sample mean of the variable. You can plot 25 glyphs at a time.

Use the *Star Plot/Sunray Plot Options* dialog box to enter a name for the labels for the observations, to choose the type of label that will appear on the plot, to enter the number of the starting glyph, and to indicate if you want to alphabetically sort the variables (see Figure 9-33, above). The figure shows the defaults.

Key Glyph

The Key Glyph option creates a Key Glyph, which is a single glyph with individual rays labeled with the name of its corresponding variable (see Figure 9-35). A Key Glyph is helpful when you need to interpret the glyphs on smaller Sunray and Star plots. The program plots the first variable you enter into the dialog box, starting at the 12:00 position and continuing clockwise.

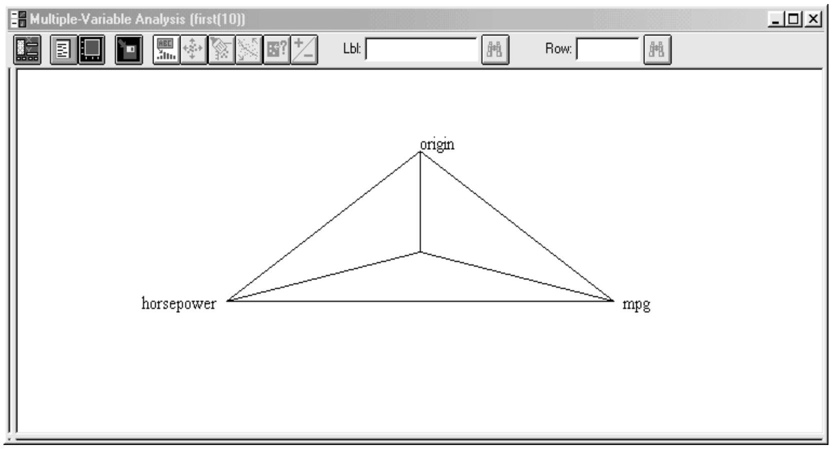


Figure 9-35. Key Glyph

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are four selections: Correlations, Rank Correlations, Covariances, and Partial Correlations.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Anderson, T. W. 1958. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Chambers, J. M., Cleveland, W. S., Kliner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. Boston: Duxbury Press.

Gibbons, Jean D. 1976. *Nonparametric Methods for Quantitative Analysis*. New York: Holt, Rinehart and Winston.

Hollander, M. and Wolfe, D. A. 1973. *Nonparametric Statistical Methods*. New York: John Wiley and Sons, Inc.

Jackson, J. E. 1959. "Quality Control Methods for Several Related Variables," *Technometrics*, **1**:4.

Johnson, R. A. and Wichern, D. W. 1982. *Applied Multivariate Statistical Analysis*. New Jersey: Prentice-Hall.

Morrison, D. F. 1990. *Multivariate Statistical Methods*, third edition. New York: McGraw-Hill Publishing Co.

Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Snedecor, G. W. and Cochran, W. G. 1967. *Statistical Methods*, sixth edition. Ames, Iowa: Iowa State University Press.

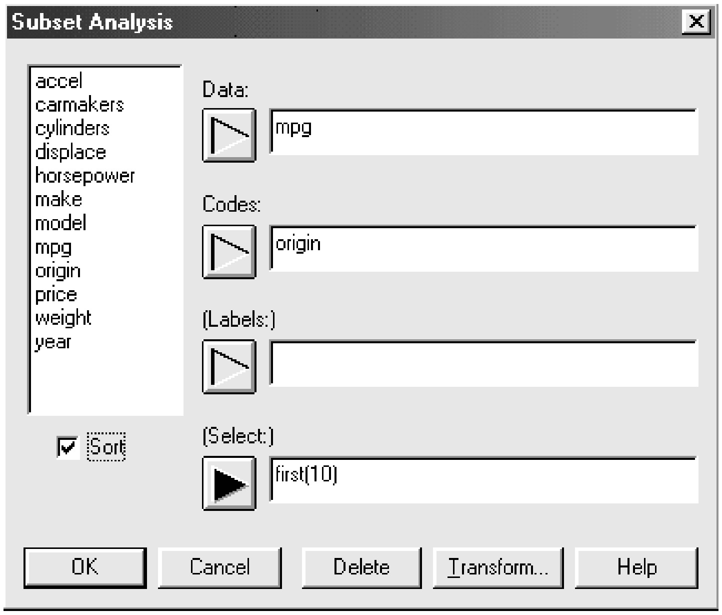
Tatsuoka, M. M. 1971. *Multivariate Analysis*. New York: Wiley.

Using the Subset Analysis

The Subset Analysis provides the ability to calculate statistics for a single column of data at each level of a second code variable. You can calculate separate summary statistics and plot means and confidence intervals for the subsets of data.

You define the subsets using the values in any variable that can partition the observations into subsets; for example, country, fiscal year, age, and sex. Doing this allows you to compare the central tendency, spread, and shape among the groups of data. For example, using the **Cardata** sample dataset, you might want to calculate statistics on miles per gallon ratings for each of three levels of origin. This is easily done using this analysis, which can calculate summary statistics for each level of a variable.

To access the analysis, from the menus, choose: DESCRIBE... NUMERIC DATA... SUBSET ANALYSIS... to display the Subset Analysis dialog box (see Figure 9-36).



Data: mpg

Codes: origin

(Labels:)

(Select:) first(10)

☑ Sort

Figure 9-36. Subset Analysis Dialog Box

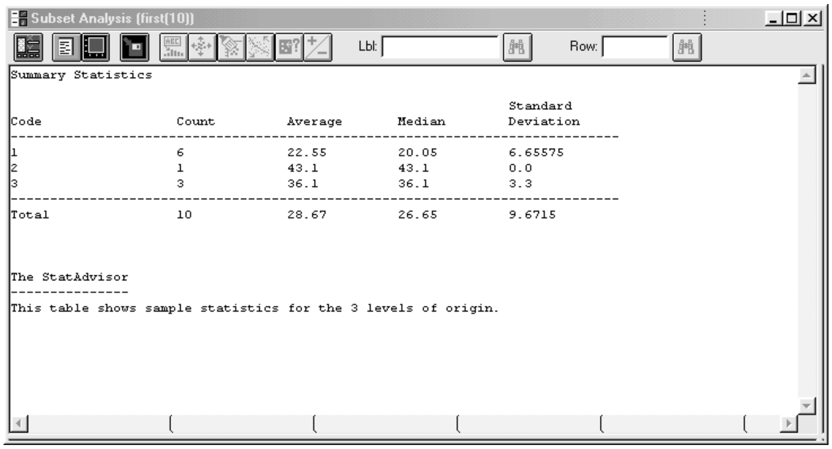
Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which includes the names of the selected variables, the number of observations, and the number of levels.

Summary Statistics

The Summary Statistics option calculates summary statistics for the variable that include measures of the center, spread, and shape of the data, such as the number of values (count), average, variance, standard deviation, minimum and maximum values, range, standardized skewness and standardized kurtosis (see Figure 9-37).



```
Summary Statistics

                                                               Standard
Code                 Count         Average       Median        Deviation
------------------------------------------------------------------------------
1                    6             22.55         20.05         6.65575
2                    1             43.1          43.1          0.0
3                    3             36.1          36.1          3.3
------------------------------------------------------------------------------
Total                10            28.67         26.65         9.6715



The StatAdvisor
---------------
This table shows sample statistics for the 3 levels of origin.
```

Figure 9-37. Summary Statistics

This information is helpful when you need to determine if other statistical analyses might be appropriate to use with the data, or when you need to determine if you should transform the data.

Use the *Summary Statistics Options* dialog box to choose the statistics you want calculated (see Figure 9-3 for an example of this dialog box).

Means Table

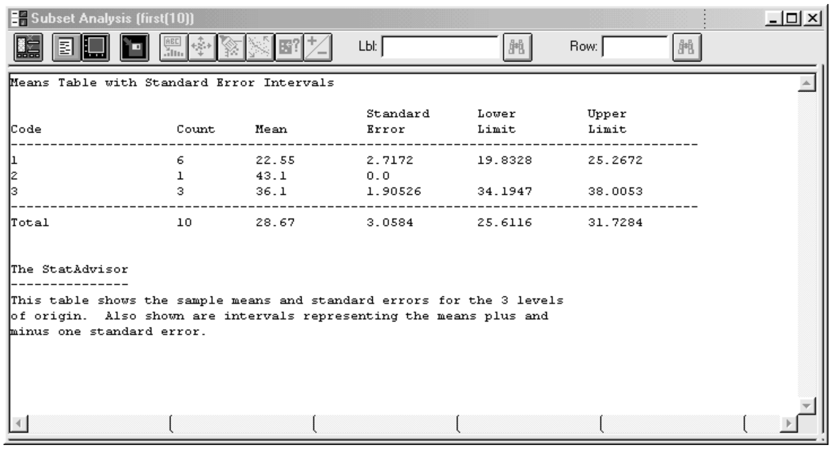
The Means Table option creates a report of the sample means and standard errors for the levels in the data (see Figure 9-38). The report also shows the intervals that represent the means, plus and minus one standard error.

Use the *Means Table Options* dialog box to choose options for the intervals and to change the confidence level (see Figure 9-39); the figure shows the defaults.

Graphical Options

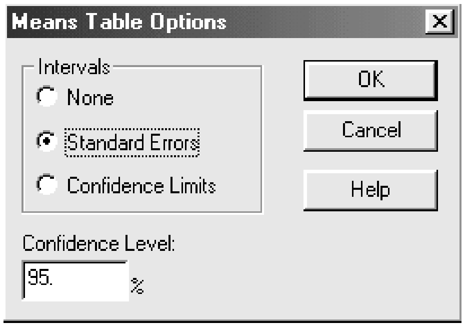
Scatterplot

The Scatterplot option creates a scatterplot that displays the pattern of points that results from plotting the values of the variable (see Figure 9-40). The levels are shown on the X-axis while the numeric values of the data are shown on the Y-axis.



```
Means Table with Standard Error Intervals

                                            Standard      Lower         Upper
Code                 Count     Mean         Error         Limit         Limit
--------------------------------------------------------------------------------------
1                    6         22.55        2.7172        19.8328       25.2672
2                    1         43.1         0.0
3                    3         36.1         1.90526       34.1947       38.0053
--------------------------------------------------------------------------------------
Total                10        28.67        3.0584        25.6116       31.7284


The StatAdvisor
---------------
This table shows the sample means and standard errors for the 3 levels
of origin.  Also shown are intervals representing the means plus and
minus one standard error.
```

Figure 9-38.   Means Table



Intervals
○ None
● Standard Errors
○ Confidence Limits

Confidence Level: 95. %

Figure 9-39. Means Table Options Dialog Box

Means Plot

The Means Plot option creates a plot that shows the sample means for the levels (see Figure 9-41). The vertical lines represent the means, plus and minus one standard error.

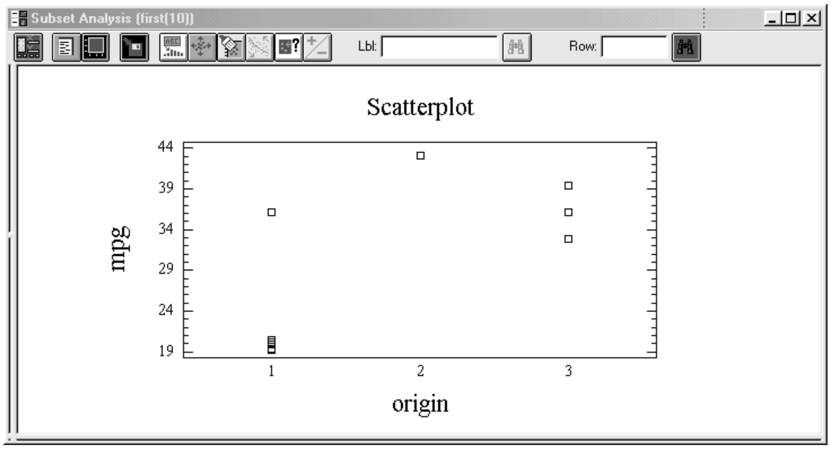


Figure 9-40. Scatterplot

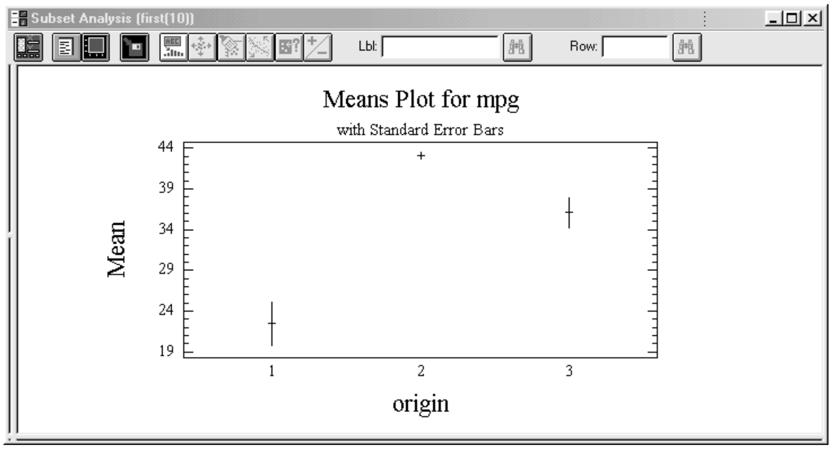
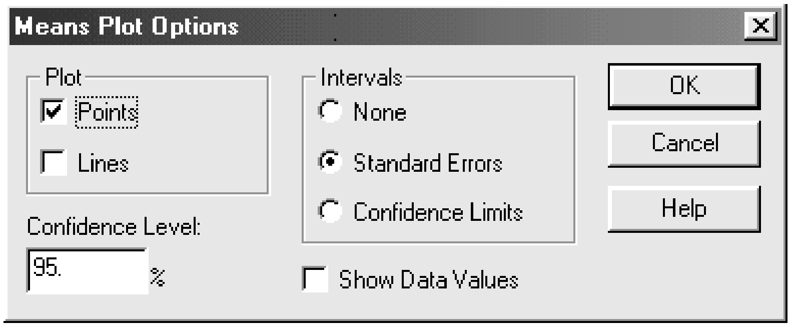


Figure 9-41. Means Plot

Use the *Means Plot Options* dialog box to indicate if Points or Lines will appear on the plot, to change the Confidence Level, to choose options for the intervals, and to indicate if you want to plot all the values on the plot (see Figure 9-42). The figure shows the defaults.



Means Plot Options

Plot

☑ Points

□ Lines

Confidence Level: 95. %

Intervals

○ None

● Standard Errors

○ Confidence Limits

□ Show Data Values

Figure 9-42. Means Plot Options Dialog Box

Median Plot

The Median Plot option creates a plot of the sample medians for the levels in the selected variables (see Figure 9-43).

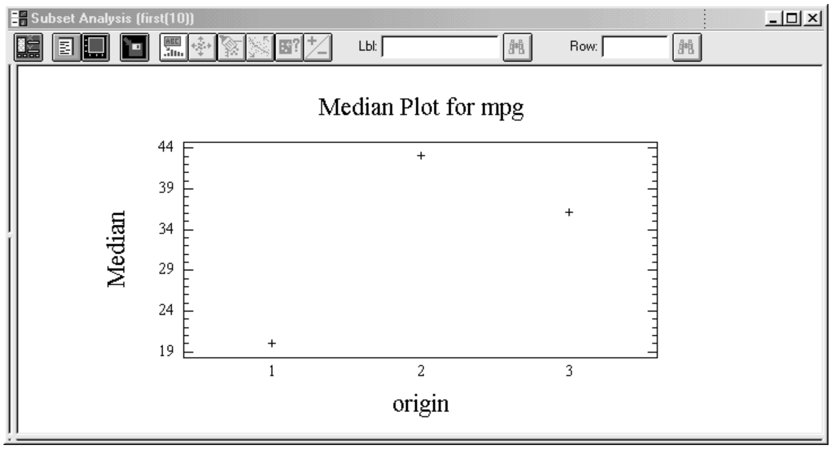


Figure 9-43. Median Plot

Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot. The default is Points.

Sigma Plot

The Sigma Plot option creates a plot that shows the sample standard deviations for the levels of the selected variables (see Figure 9-44).

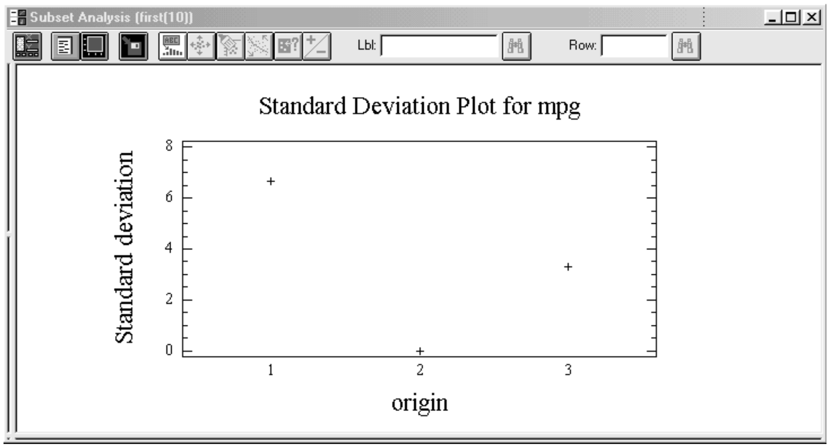


Figure 9-44. Sigma Plot

Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot. The default is points.

Range Plot

The Range Plot option creates a plot of the sample ranges for the levels of the selected variables (see Figure 9-45).

Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot. The default is points.

Box-and-Whisker Plot

The Box-and-Whisker Plot option creates a Box-and-Whisker Plot for each level of the selected variable (see Figure 9-46).

The rectangular portion of the plot extends from the lower quartile to the upper quartile, covering the center half of each sample. The center lines

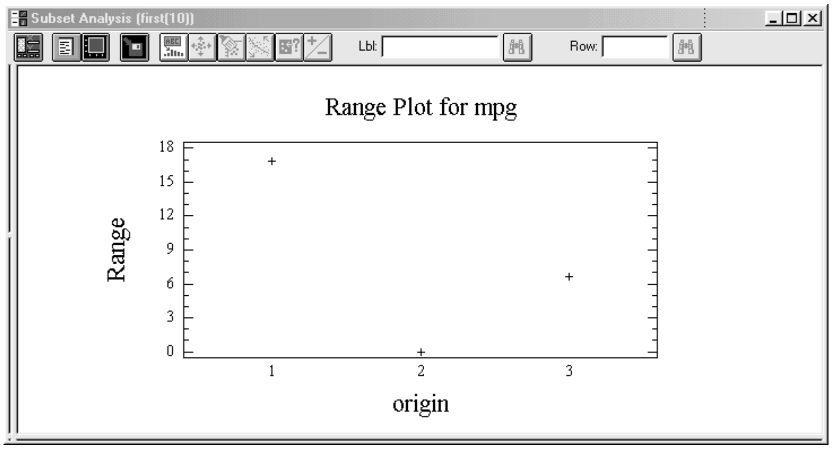


Figure 9-45.   Range Plot

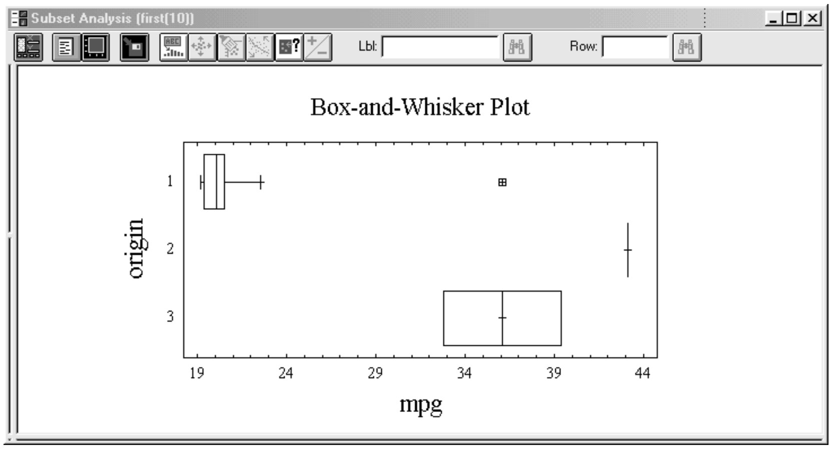


Figure 9-46. Box-and-Whisker Plot

within each box show the location of the sample medians. The plus signs represent the location of the sample means.

The whiskers extend from the box to the minimum and maximum values in each sample, except for any outside or far-outside points, which are plotted separately. Outside points lie more than 1.5 times the interquartile range above or below the box; they are shown as small squares. Far-outside points lie more than 3.0 times the interquartile range above or below the box; they are shown as small squares with plus signs through them.

Use the *Box-and-Whisker Plot Options* dialog box to indicate if the plot will appear in a vertical or horizontal direction, and to choose features for the plot such as median notch, outlier symbols, and mean marker (see Figure 9-15 for an example of this dialog box).

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are five selections: Labels, Counts, Means, Medians, and Standard Deviations.

You can also use the Target Variables text boxes on the dialog box to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

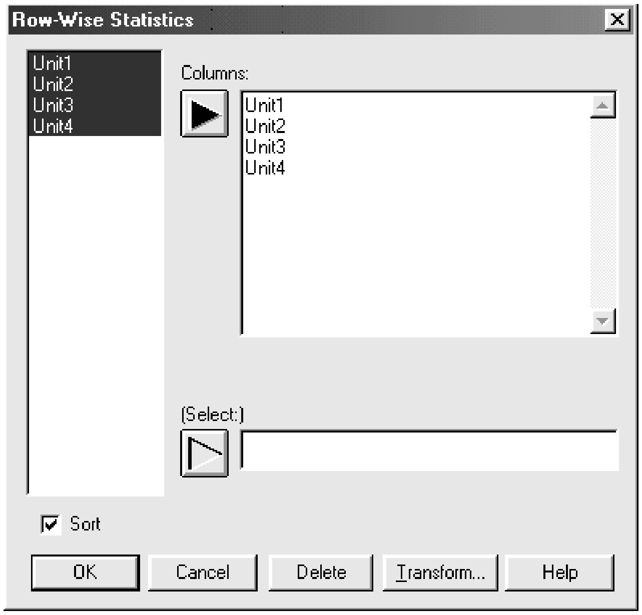
Fogiel, M. and the staff of the Research and Education Association. 1978. *The Statistical Problem Solver*. Piscataway, New Jersey: Research and Education Association.

Wonnacott, T. H. and Wonnacott, R. J. 1972. *Introductory Statistics*, second edition. New York: Wiley.

Using the Row-Wise Statistics Analysis

The Row-Wise Statistics Analysis provides the capability to calculate and save statistics on a row-wise basis for data that are in two or more columns. The analysis calculates summary statistics for each row in a set of columns. It is designed to save the statistics so you can use them in other analyses when the data are not arranged in the usual column-wise manner.

To access the analysis, from the menus, choose: DESCRIBE... NUMERIC DATA... ROW-WISE STATISTICS... to display the Analysis dialog box (see Figure 9-47).



Row-Wise Statistics

Unit1
Unit2
Unit3
Unit4

Columns:
Unit1
Unit2
Unit3
Unit4

(Select:)

☑ Sort

Figure 9-47. Row-Wise Statistics Analysis Dialog Box

Tabular Options

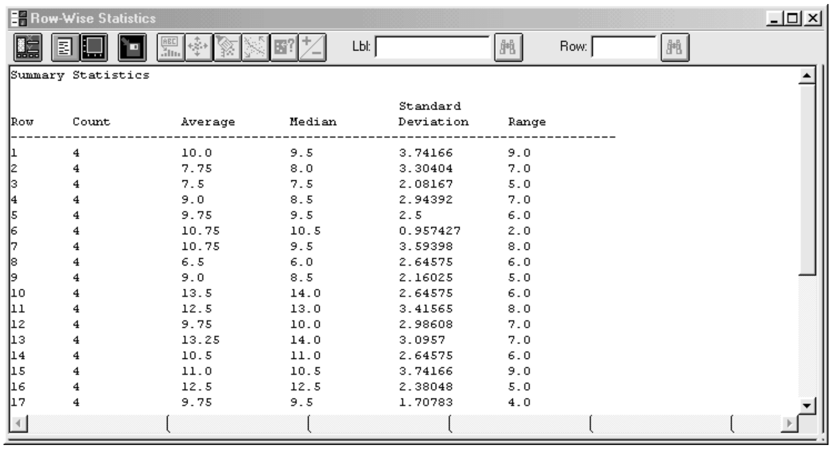
Analysis Summary

The Analysis Summary option displays the names of the selected variables (the data columns) and the number of rows.

Summary Statistics

The Summary Statistics option creates a table that shows the sample statistics for the rows in the file that have at least one nonmissing value (see Figure 9-48).

Use the *Summary Statistics Options* dialog box to choose the statistics you want calculated (see Figure 9-3 for an example of this dialog box).



```
Summary Statistics

                                             Standard
Row     Count         Average       Median        Deviation     Range
------------------------------------------------------------------------------
1       4             10.0          9.5           3.74166       9.0
2       4             7.75          8.0           3.30404       7.0
3       4             7.5           7.5           2.08167       5.0
4       4             9.0           8.5           2.94392       7.0
5       4             9.75          9.5           2.5           6.0
6       4             10.75         10.5          0.957427      2.0
7       4             10.75         9.5           3.59398       8.0
8       4             6.5           6.0           2.64575       6.0
9       4             9.0           8.5           2.16025       5.0
10      4             13.5          14.0          2.64575       6.0
11      4             12.5          13.0          3.41565       8.0
12      4             9.75          10.0          2.98608       7.0
13      4             13.25         14.0          3.0957        7.0
14      4             10.5          11.0          2.64575       6.0
15      4             11.0          10.5          3.74166       9.0
16      4             12.5          12.5          2.38048       5.0
17      4             9.75          9.5           1.70783       4.0
```

Figure 9-48. Summary Statistics

Means Table

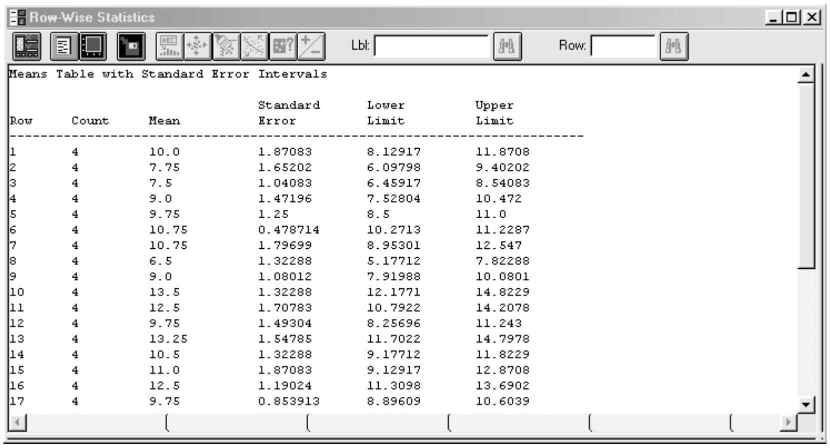
The Means Table option creates a table of the sample means and standard errors for the rows in the file that have at least one nonmissing value (see Figure 9-49). The table also shows the intervals that represent the means, plus or minus one standard error.

Use the *Means Table Options* dialog box to choose options for the intervals and to change the confidence level (see Figure 9-39 for an example of this dialog box). The figure shows the defaults.

Graphical Options

Scatterplot

The Scatterplot option creates a scatterplot that displays the pattern of points that results from plotting the rows of data (see Figure 9-50).



```
Means Table with Standard Error Intervals

                                Standard      Lower         Upper
Row     Count     Mean          Error         Limit         Limit
--------------------------------------------------------------------------
1       4         10.0          1.87083       8.12917       11.8708
2       4         7.75          1.65202       6.09798       9.40202
3       4         7.5           1.04083       6.45917       8.54083
4       4         9.0           1.47196       7.52804       10.472
5       4         9.75          1.25          8.5           11.0
6       4         10.75         0.478714      10.2713       11.2287
7       4         10.75         1.79699       8.95301       12.547
8       4         6.5           1.32288       5.17712       7.82288
9       4         9.0           1.08012       7.91988       10.0801
10      4         13.5          1.32288       12.1771       14.8229
11      4         12.5          1.70783       10.7922       14.2078
12      4         9.75          1.49304       8.25696       11.243
13      4         13.25         1.54785       11.7022       14.7978
14      4         10.5          1.32288       9.17712       11.8229
15      4         11.0          1.87083       9.12917       12.8708
16      4         12.5          1.19024       11.3098       13.6902
17      4         9.75          0.853913      8.89609       10.6039
```

Figure 9-49. Means Table

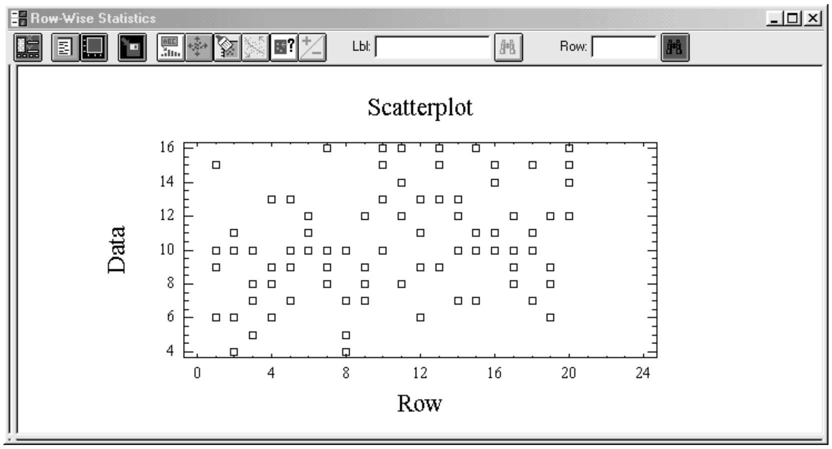


Figure 9-50. Scatterplot

Means Plot

The Means Plot option creates a plot that shows the sample means for the rows in the file that have at least one nonmissing value (see Figure 9-51). The vertical lines represent the means, plus and minus one standard error.

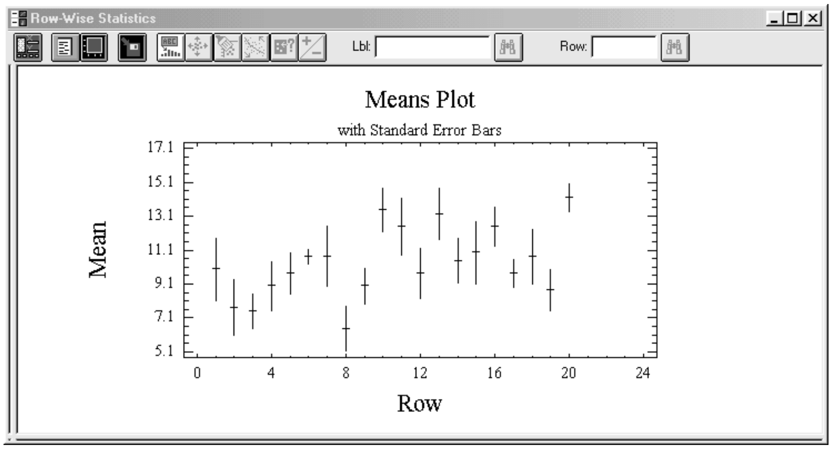


Figure 9-51. Means Plot

Use the *Means Plot Options* dialog box to indicate if points or lines will appear on the plot, to change the confidence level, to choose options for the intervals, and to indicate if you want to plot all the values on the plot (see Figure 9-42 for an example of this dialog box). The figure shows the defaults.

Median Plot

The Median Plot option creates a plot of the sample medians for the rows in the file that have at least one nonmissing value (see Figure 9-52).

Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot. The default is points.

Sigma Plot

The Sigma Plot option creates a plot that shows the sample standard deviations for the rows in the file that have at least one nonmissing value (see Figure 9-53).

Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot. The default is points.

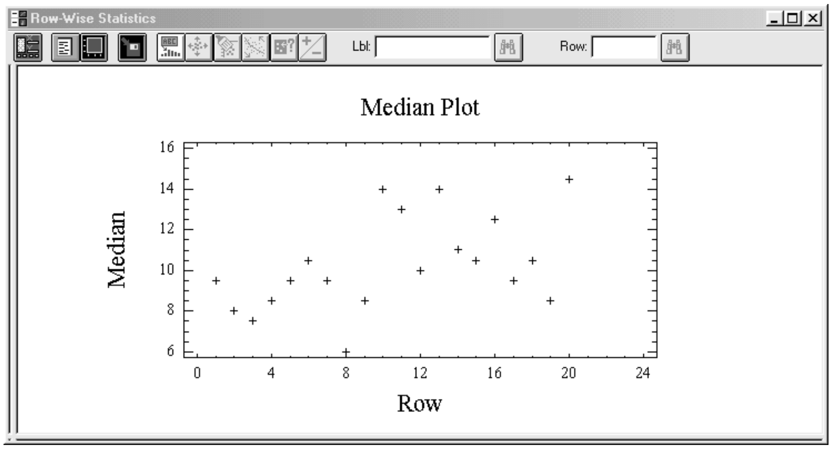


Figure 9-52. Median Plot

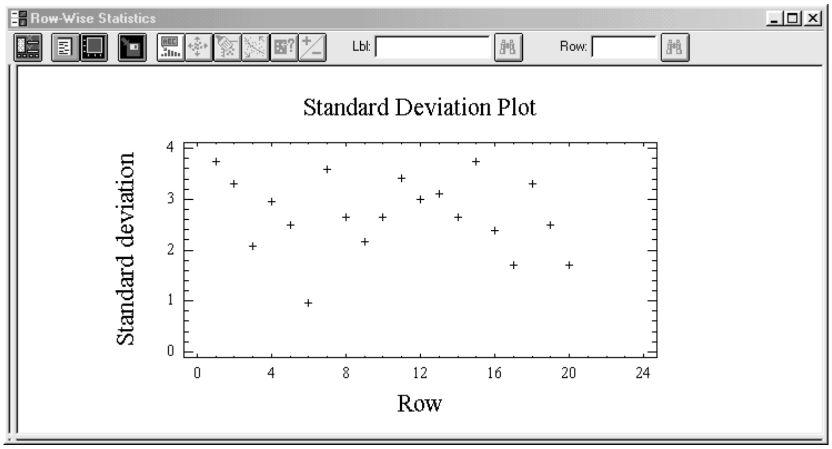


Figure 9-53. Sigma Plot

Range Plot

The Range Plot option creates a plot of the sample ranges for the rows in the file that have at least one nonmissing value (see Figure 9-54).

Use the *Plot Options* dialog box to indicate if points or lines will appear on the plot. The default is points.

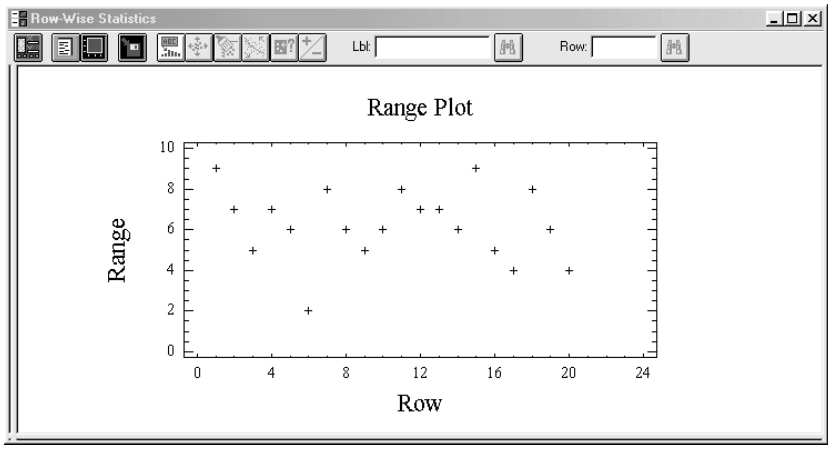


Figure 9-54. Range Plot

Box-and-Whisker Plot

The Box-and-Whisker Plot option creates a Box-and-Whisker Plot for each row of selected data (see Figure 9-55). The rectangular portion of the plot extends from the lower quartile to the upper quartile, covering the center half of each sample. The center lines within each box show the location of the sample medians. The plus signs represent the location of the sample means.

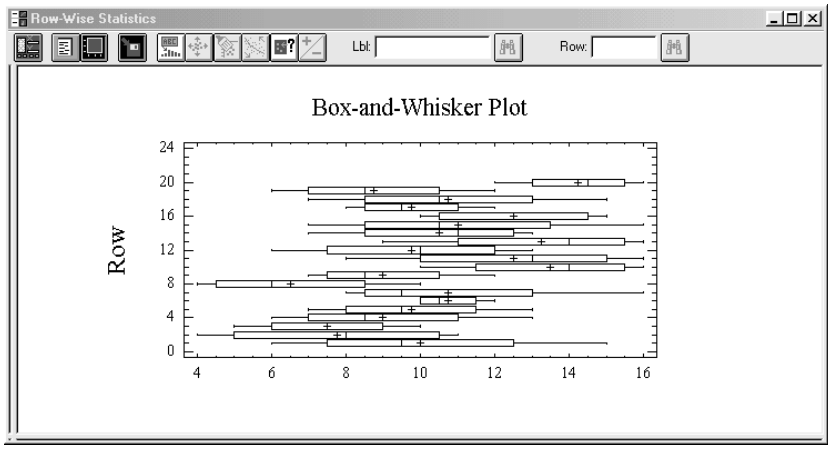


Figure 9-55. Box-and-Whisker Plot

The whiskers extend from the box to the minimum and maximum values in each sample, except for any outside or far-outside points, which are plotted separately. Outside points lie more than 1.5 times the interquartile range above or below the box; they are shown as small squares. Far-outside points lie more than 3.0 times the interquartile range above or below the box; they are shown as small squares with plus signs through them.

Use the *Box-and-Whisker Plot Options* dialog box to indicate if the plot will appear in a vertical or horizontal direction, and to choose features for the plot such as median notch, outlier symbols, and mean marker (see Figure 9-15 for an example of this dialog box). The figure shows the defaults.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are five selections: Counts, Means, Medians, Standard Deviations, and Ranges.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Montgomery, D. C. 1991. *Introduction to Statistical Quality Control*, second edition. New York: John Wiley & Sons.

Using the Power Transformations Analysis

If, after performing an exploratory analysis on a set of data, the normality of the data appears to be questionable, a transformation of the data is in order even though it is often difficult to determine what type of transformation will have the most beneficial effect. The Box-Cox power transformation automatically identifies the optimal transformations.

The program determines the optimal Box-Cox transformation by finding the value of the Lambda1 parameter, which minimizes the mean squared error. The StatAdvisor displays the transformation for the value of Lambda 1 you selected. You can compare the results of the transformation with other values of Lambda1 by choosing the MSE Comparison Table option.

To access the analysis, from the menus, choose: DESCRIBE... NUMERIC DATA... POWER TRANSFORMATIONS... to display the Analysis dialog box (see Figure 9-56).

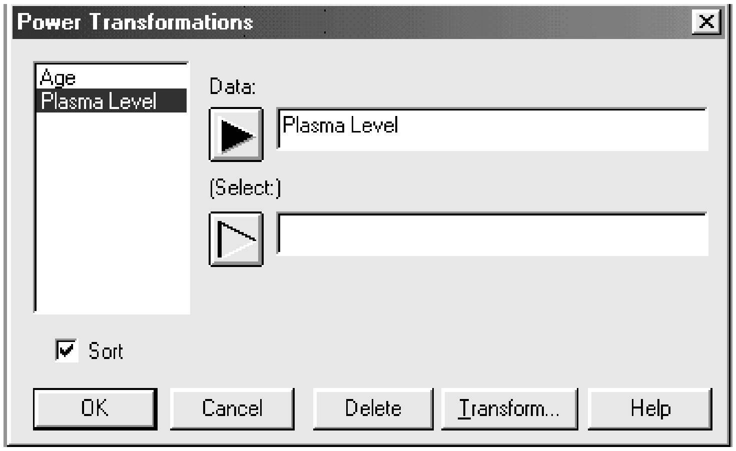


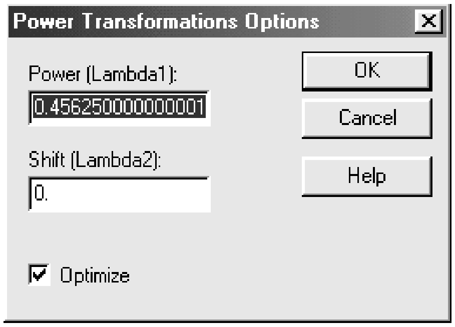
Figure 9-56. Power Transformations Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that displays the name of the variable, the number of observations, the values for the Box-Cox transformation (Lambda1 and Lambda2 optimized), and the value for the geometric mean.

Use the *Power Transformations Options* dialog box to change the values for Lambda1 and Lambda2, and to indicate if you want the value for Lambda1 optimized (see Figure 9-57).



Power (Lambda1): 0.456250000000001

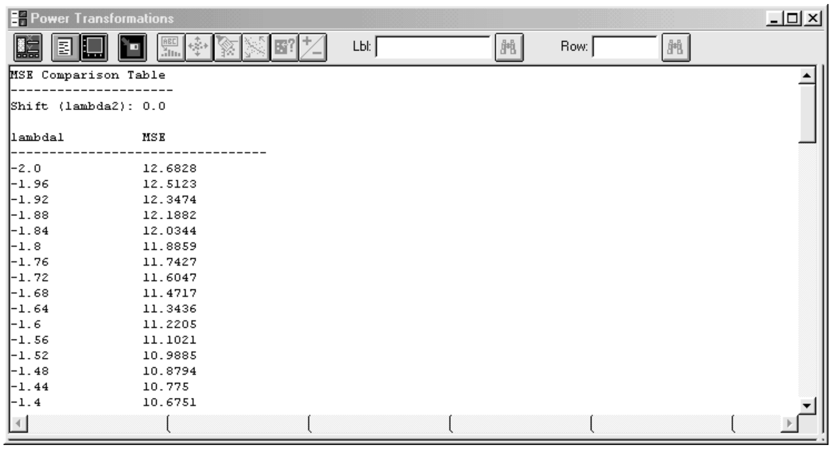
Shift (Lambda2): 0.

☑ Optimize

Figure 9-57. Power Transformations Options Dialog Box

MSE Comparison Table

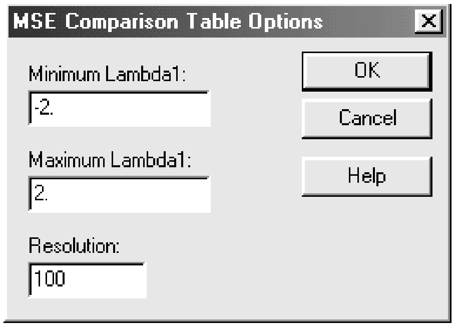
The MSE Comparison Table option creates a table that shows the Mean Squared Error (MSE) for various values of the power transformation parameter, Lambda1 (see Figure 9-58). The table shows the values for MSE for a range of different Lambdas.



```
MSE Comparison Table
---------------------
Shift (lambda2): 0.0

lambda1          MSE
---------------------------------
-2.0             12.6828
-1.96            12.5123
-1.92            12.3474
-1.88            12.1882
-1.84            12.0344
-1.8             11.8859
-1.76            11.7427
-1.72            11.6047
-1.68            11.4717
-1.64            11.3436
-1.6             11.2205
-1.56            11.1021
-1.52            10.9885
-1.48            10.8794
-1.44            10.775
-1.4             10.6751
```

Figure 9-58. MSE Comparison Table

Use the *MSE Comparison Table Options* dialog box to set the values for which the MSE will be determined, and to enter a value for the resolution (see Figure 9-59). The figure shows the defaults.



Minimum Lambda1: -2.

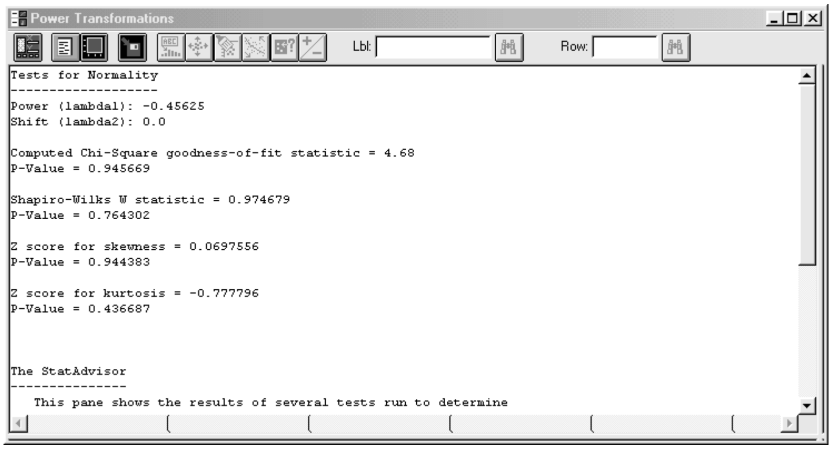
Maximum Lambda1: 2.

Resolution: 100

Figure 9-59. MSE Comparison Table Options Dialog Box

Tests for Normality

The Tests for Normality option displays the results of tests run to determine if the transformed values of the variable adequately model a normal distribution (see Figure 9-60).



```
Tests for Normality
-------------------
Power (lambda1): -0.45625
Shift (lambda2): 0.0

Computed Chi-Square goodness-of-fit statistic = 4.68
P-Value = 0.945669

Shapiro-Wilks W statistic = 0.974679
P-Value = 0.764302

Z score for skewness = 0.0697556
P-Value = 0.944383

Z score for kurtosis = -0.777796
P-Value = 0.436687



The StatAdvisor
---------------
   This pane shows the results of several tests run to determine
```

Figure 9-60. Tests for Normality

The Chi-Square Test divides the range of the transformed values of the variable into equal probability classes, and compares the number of observations in each class with the expected number, based on the fitted distribution.

The Shapiro-Wilks test is based on the comparison of the quantiles for the fitted normal distribution with the quantiles of the data.

The Standardized Skewness test looks for lack of symmetry in the data, while the Standardized Kurtosis test looks for distributional shape that is either flatter or more peaked than a normal distribution.

The Tests for Normality table shows the lowest *p*-value among the tests performed. If the *p* value is less than 0.05, you can reject the idea that the transformed values come from a normal distribution at the 95 percent confidence level.

Graphical Options

Normal Probability Plot

The Normal Probability Plot option creates a plot of the transformed values for the selected variable (see Figure 9-61). The program sorts the values from smallest to largest, then plots them versus the values (i-0.375)/(*n*+0.25), where *n*= equals the sample size.

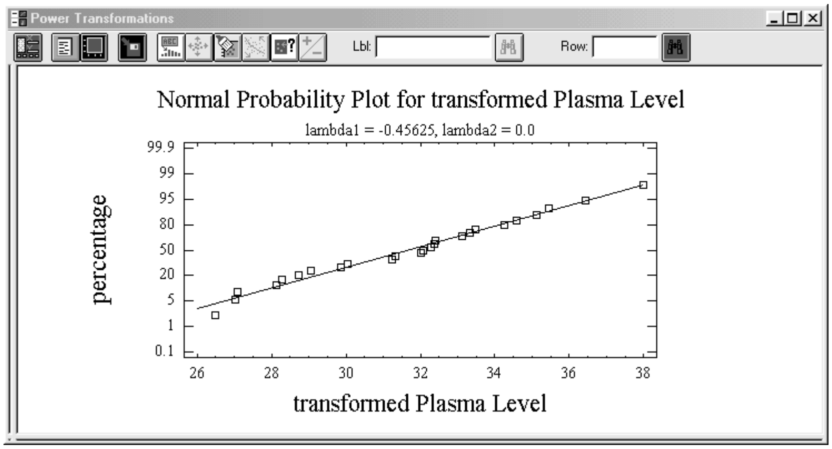
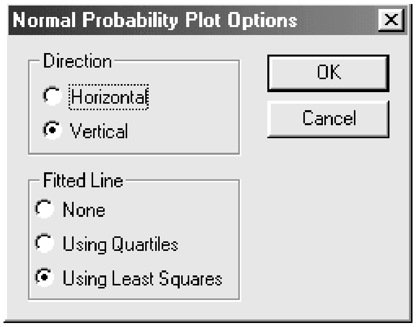


Figure 9-61. Normal Probability Plot

If the data come from a normal distribution, the points should fall approximately along a straight line. To help determine the closeness of the points to the straight line, a reference line is drawn on the plot. The reference line, by default, is fit to the data using least squares. Additional tests for normality are also available in this analysis.

Use the *Normal Probability Plot Options* dialog box to indicate if the plot will display in a horizontal or vertical direction, and to indicate if a fitted line will appear on the plot; if so, whether quartiles or least squares will be the method used to fit the line (see Figure 9-62).



Normal Probability Plot Options

Direction
○ Horizontal
● Vertical

Fitted Line
○ None
○ Using Quartiles
● Using Least Squares

Figure 9-62. Normal Probability Plot Options Dialog Box

MSE Comparison Plot

The MSE Comparison Plot option creates a plot that shows the various values of the power transformation parameter, Lambda1, between -2.0 and +2.0 (see Figure 9-63). It also shows the smallest value for MSE at Lambda1.

Use the *MSE Comparison Plot Options* dialog box to set the values for which the MSE will be determined, and to enter a value for the resolution (see Figure 9-64).

Skewness and Kurtosis Plot

The Skewness and Kurtosis Plot option creates a plot that shows the values for the standardized skewness and standardized kurtosis at various values of

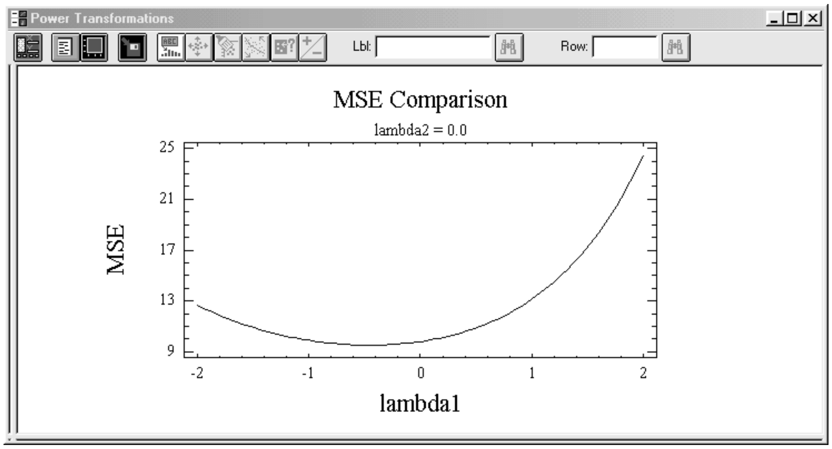
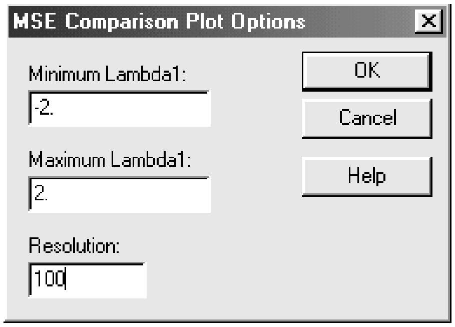


Figure 9-63. MSE Comparison Plot



Minimum Lambda1: -2.

Maximum Lambda1: 2.

Resolution: 100

Figure 9-64. MSE Comparison Plot Options Dialog Box

Lambda1 (see Figure 9-65). A vertical line is drawn at the optimal value; horizontal lines are drawn at 0 and +/-2.

Use the *Skewness and Kurtosis Plot Options* dialog box to set the values for which the MSE will be determined, and to enter a value for the resolution (see Figure 9-66).

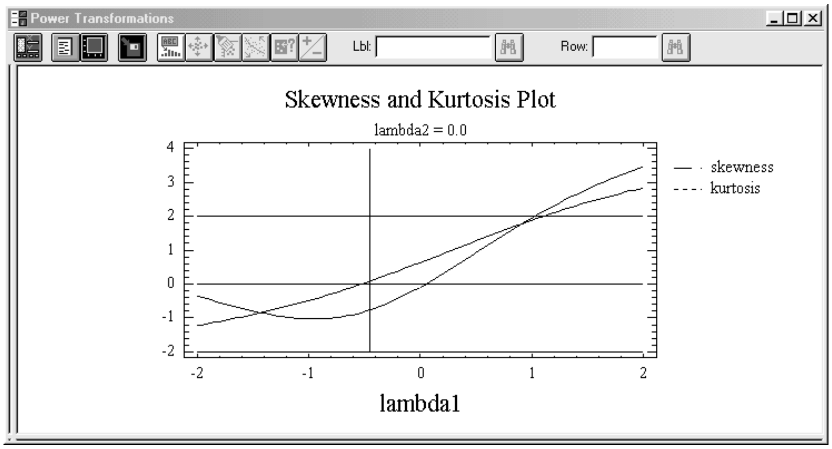
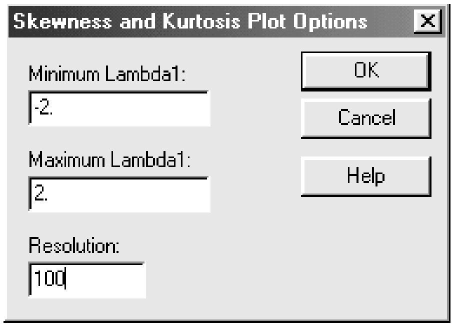


Figure 9-65. Skewness and Kurtosis Plot



Minimum Lambda1: -2.

Maximum Lambda1: 2.

Resolution: 100

Figure 9-66. Skewness and Kurtosis Plot Options Dialog Box

Saving the Results

Use the Save Results Options dialog box to choose the results you want to save. There is one selection: Transformed Data.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

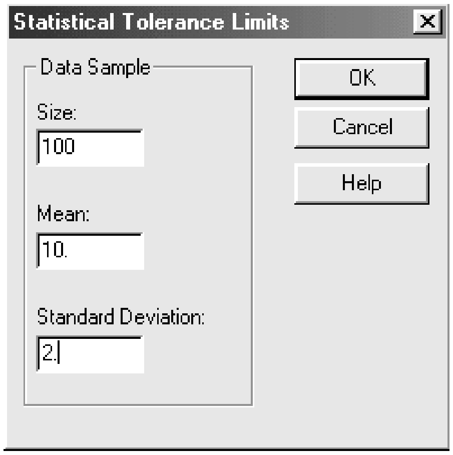
Using the Statistical Tolerance Limits Analysis

The Statistical Tolerance Limits Analysis computes statistical tolerance limits, given a sample size, mean, and standard deviation. The analysis also determines the sample size needed to obtain nonparametric limits for a given proportion of the population.

Statistical tolerance limits are limits that contain a certain proportion of the distribution of a quality characteristic. When the parameters of a distribution are known, these limits are easily calculated. However, usually the distribution and its parameters are unknown, and you must estimate the tolerance limits from a sample of population data.

This analysis provides both normal tolerance limits and nonparametric tolerance limits. Use normal tolerance limits when it can reasonably be assumed that the data collected are well described by a normal distribution. When the normality of the data cannot be assumed, use Nonparametric limits.

To access the analysis, from the menus, choose: DESCRIBE... NUMERIC DATA... STATISTICAL TOLERANCE LIMITS... to display the Analysis dialog box (see Figure 9-67).



Size: 100

Mean: 10.

Standard Deviation: 2.

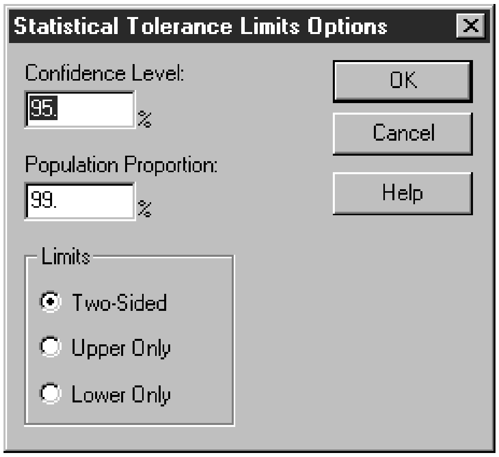
Figure 9-67. Statistical Tolerance Limits Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the size, mean, and standard deviation for the sample. It then displays the specified tolerance limits, confidence interval, and population proportions.

Use the *Statistical Tolerance Limits Options* dialog box to enter values for the confidence level and the population proportion, and to indicate if two-sided, upper, or lower limits will be used (see Figure 9-68). The figure shows the defaults.



Statistical Tolerance Limits Options

Confidence Level: 95. %

Population Proportion: 99. %

Limits

● Two-Sided

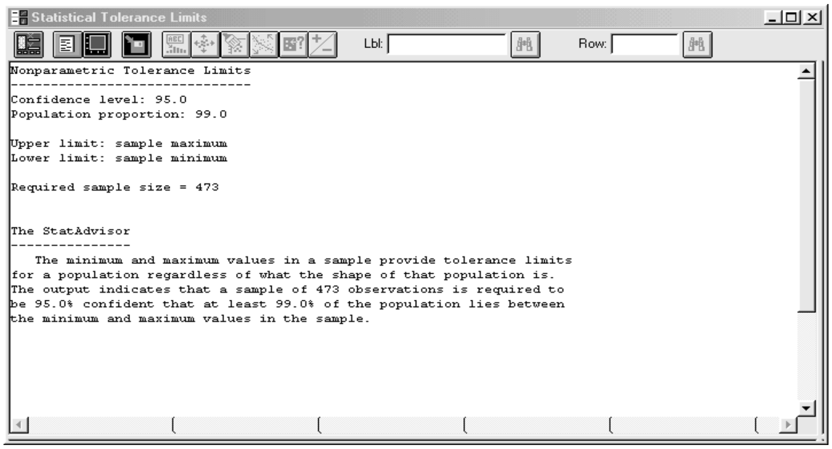
○ Upper Only

○ Lower Only

Figure 9-68. Statistical Tolerance Limits Options Dialog Box

Nonparametric Limits

The Nonparametric Limits option displays the specified tolerance limits (always the sample maximum and minimum), confidence level, and population proportion. The required sample size represents the number of observations that must be taken to ensure that, with the specified confidence level, at least the specified proportion of the population lies between the minimum and maximum values (see Figure 9-69).



```
Nonparametric Tolerance Limits
------------------------------
Confidence level: 95.0
Population proportion: 99.0

Upper limit: sample maximum
Lower limit: sample minimum

Required sample size = 473


The StatAdvisor
---------------
   The minimum and maximum values in a sample provide tolerance limits
for a population regardless of what the shape of that population is.
The output indicates that a sample of 473 observations is required to
be 95.0% confident that at least 99.0% of the population lies between
the minimum and maximum values in the sample.
```

Figure 9-69. Nonparametric Limits

Graphical Options

Tolerance Plot

The Tolerance Plot option displays a normal density function with vertical lines drawn at the location of the tolerance limits (see Figure 9-70).

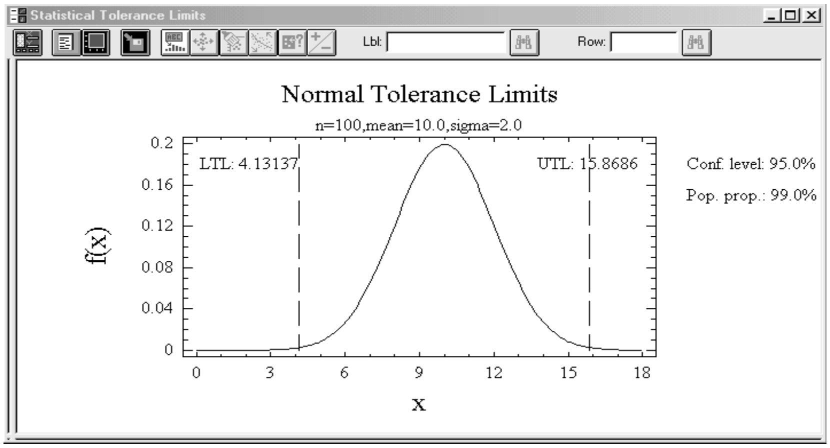


Figure 9-70. Tolerance Plot

Using the Outlier Identification Analysis

The Outlier Identification Analysis allows you to detect the presence of outliers and quantify the extent of the evidence against their validity. It also provides estimates of common parameters, such as the mean and standard deviation, which are designed to be resistant to the potential presence of such outliers. Options in this analysis allow you to calculate specific summary statistics, such as tests for normality.

To access the analysis, from the menus, choose: DESCRIBE... NUMERIC DATA... OUTLIER IDENTIFICATION... to display the Analysis dialog box (see Figure 9-71).

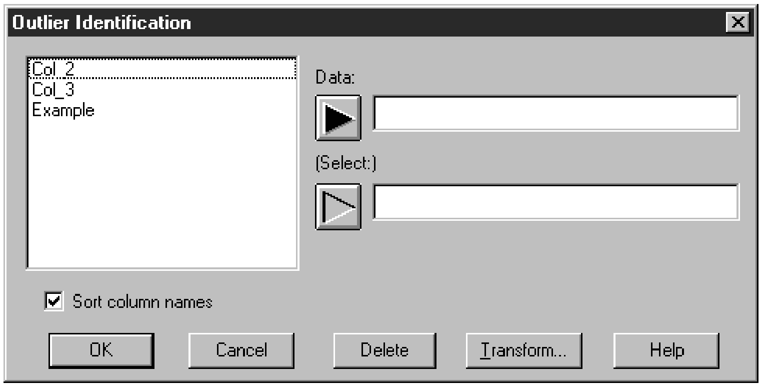


Figure 9-71. Outlier Identification Analysis Dialog Box

Tabular Options

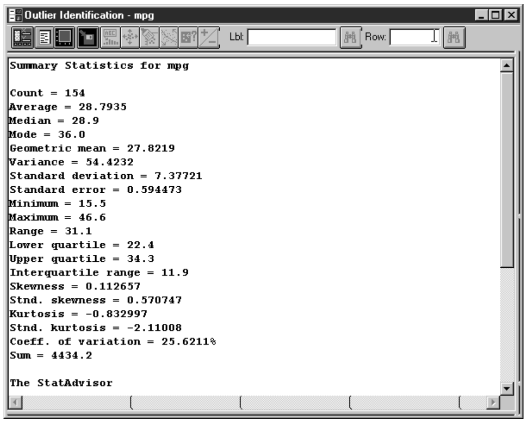
Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the name of the data variable and range; several location estimates, including the sample mean, median, trimmed mean, and Winsorized mean; several estimates of scale, including the sample standard deviation, an estimate based on the median absolute deviation (MAD), Sbi, and Winsorized sigma; confidence intervals for the mean, both based on the usual methods and using the Winsorized mean and sigma; and a table of sorted values. The table includes Studentized values (xi-xbar)/s, where xbar and s are estimated with and without each individual data value. It also includes a modified z-score based on the median absolute deviation. (The StatAdvisor notes any modified z-scores that exceed 3.5 in absolute value.) Also included on the output are the results of Grubb's test for a single outlier (if n>2) and Dixon's test for 1 or 2 outliers (if n is between 4 and 30).

Specify the confidence level for the confidence interval for the mean, the amount of trimming used in the trimmed mean and Winsorized statistics, and the number of smallest and largest values to display in the table.

Summary Statistics

The Summary Statistics option calculates summary statistics for the variable that include measures of the center, spread, and shape of the data, such as the number of values (count), average, variance, standard deviation, minimum and maximum values, range, standardized skewness and standardized kurtosis (see Figure 9-72).

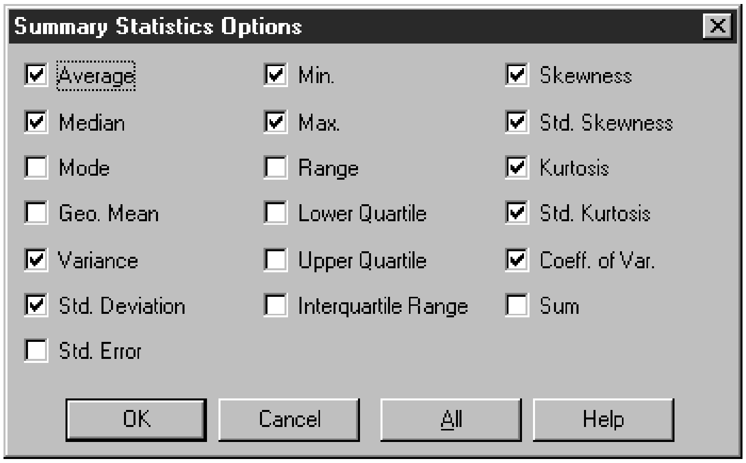


```
Summary Statistics for mpg

Count = 154
Average = 28.7935
Median = 28.9
Mode = 36.0
Geometric mean = 27.8219
Variance = 54.4232
Standard deviation = 7.37721
Standard error = 0.594473
Minimum = 15.5
Maximum = 46.6
Range = 31.1
Lower quartile = 22.4
Upper quartile = 34.3
Interquartile range = 11.9
Skewness = 0.112657
Stnd. skewness = 0.570747
Kurtosis = -0.832997
Stnd. kurtosis = -2.11008
Coeff. of variation = 25.6211%
Sum = 4434.2

The StatAdvisor
```

Figure 9-72. Summary Statistics

This information is helpful when you need to determine if other statistical analyses might be more appropriate to use with the data, or when you need to determine if you should transform the data.

Use the *Summary Statistics Options* dialog box to choose the statistics you want calculated (see Figure 9-73); the figure shows the defaults.



Summary Statistics Options

☑ Average
☑ Median
□ Mode
□ Geo. Mean
☑ Variance
☑ Std. Deviation
□ Std. Error

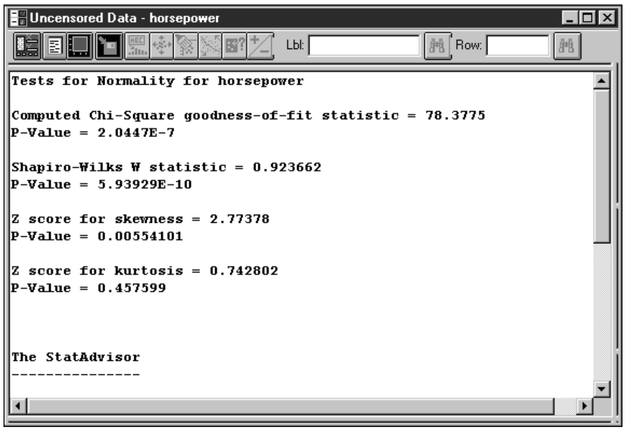
☑ Min.
☑ Max.
□ Range
□ Lower Quartile
□ Upper Quartile
□ Interquartile Range

☑ Skewness
☑ Std. Skewness
☑ Kurtosis
☑ Std. Kurtosis
☑ Coeff. of Var.
□ Sum

Figure 9-73. Summary Statistics Options Dialog Box

Tests for Normality

The Tests for Normality options displays the results of tests run to determine if the transformed values of the variable adequately model a normal distribution (see Figure 9-74).



```
Tests for Normality for horsepower

Computed Chi-Square goodness-of-fit statistic = 78.3775
P-Value = 2.0447E-7

Shapiro-Wilks W statistic = 0.923662
P-Value = 5.93929E-10

Z score for skewness = 2.77378
P-Value = 0.00554101

Z score for kurtosis = 0.742802
P-Value = 0.457599



The StatAdvisor
---------------
```

Figure 9-74. Tests for Normality

The Chi-Square Test divides the range of the transformed values of the variable into equal probability classes, and compares the number of observations in each class with the expected number, based on the fitted distribution.

The Shapiro-Wilks test is based on the comparisons of the quantiles for the fitted normal distribution with the quantiles of the data.

The Standardized Skewness test looks for lack of symmetry in the data, while the Standardized Kurtosis test looks for distributional shape that is either flatter or more peaked than a normal distribution.

The Tests for Normality table shows the lowest *p*-value among the tests performed. If the *p*-value is less than 0.05, you can reject the idea that the transformed values come from a normal distribution at the 95 percent confidence level.

Graphical Options

Outlier Plot

The Outlier Plot option creates a plot that displays the pattern of points that results from plotting the values of the variable (see Figure 9-75). The levels are shown on the X-axis while the numeric values of the data are shown on the Y-axis.

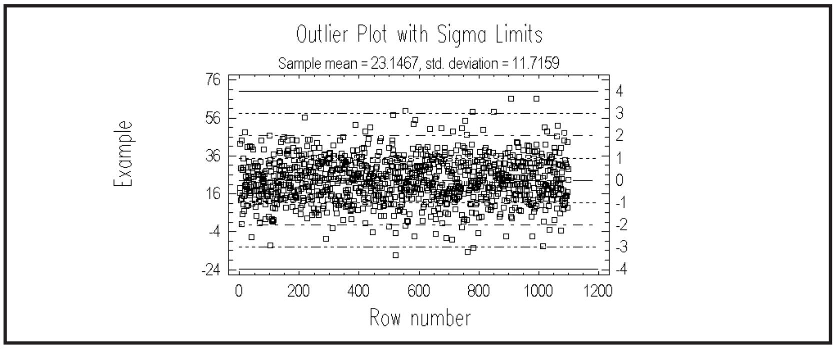
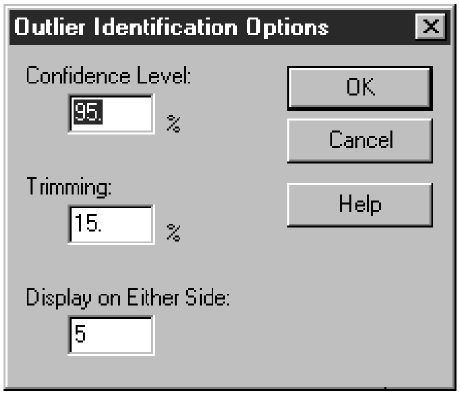


Figure 9-75. Outlier Plot

Use the *Outlier Identification Options* dialog box to indicate... (See Figure 9-76).



Outlier Identification Options

Confidence Level: 95. %

Trimming: 15. %

Display on Either Side: 5

Figure 9-76. Outlier Identification Options Dialog Box

Box-and-Whisker Plot

The Box-and-Whisker Plot option creates a plot of the data, which is divided into four equal areas of frequency (quartiles) (see Figure 9-77) . A box encloses the middle 50 percent, where the median is drawn as a vertical line inside the box.

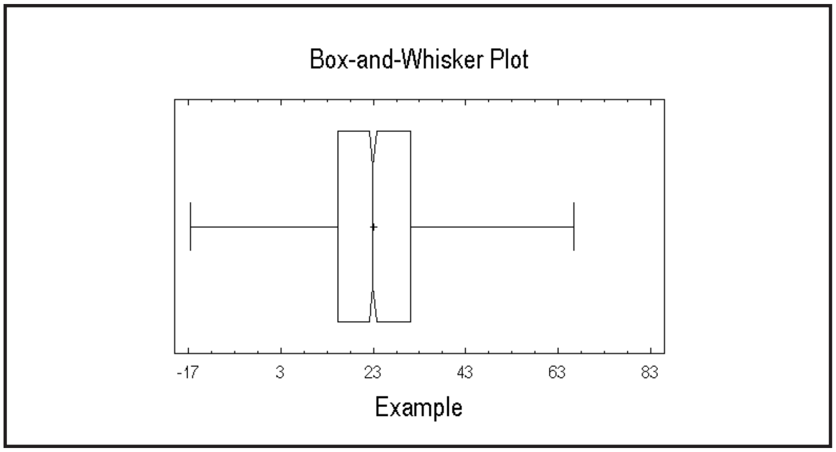
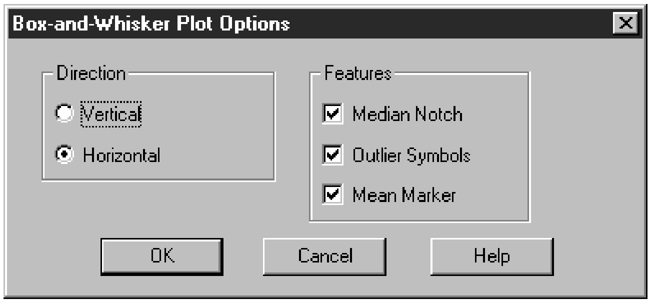


Figure 9-77. Box-and-Whisker Plot Options Dialog Box

Horizontal lines, known as whiskers, extend from each end of the box. The left (or lower) whisker is drawn from the lower quartile to the smallest point within 1.5 interquartile ranges from the lower quartile. The other whisker is drawn from the upper quartile to the largest point within 1.5 interquartile ranges from the upper quartile.

Use the *Box-and-Whisker Plot Options* dialog box to indicate if the plot will appear in a vertical or horizontal direction, and to choose features for the plot such as median notch, outlier symbols, and mean marker (see Figure 9-78); the figure shows the defaults.



Box-and-Whisker Plot Options

Direction

○ Vertical

● Horizontal

Features

☑ Median Notch

☑ Outlier Symbols

☑ Mean Marker

Figure 9-78. Box-and-Whisker Plot Options Dialog Box

Normal Probability Plot

The Normal Probability Plot option helps to determine if the data come from a normal distribution (see Figure 9-79). The plot consists of an arithmetic (interval) horizontal axis scaled for the data, and a vertical axis scaled so the cumulative distribution function of a normal distribution plots as a straight line. The closer the data are to being on a straight line, the more likely they follow a normal distribution. Significant curvature of the data indicates skewness.

Use the *Normal Probability Plot Options* dialog box to indicate the direction of the plot, to indicate if you want a fitted line on the plot and, if so, to choose the method that will be used to calculate it (see Figure 9-80); the figure shows the defaults.

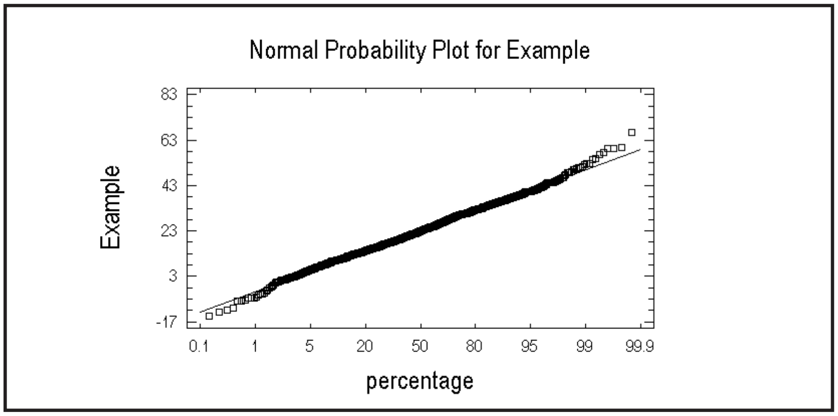
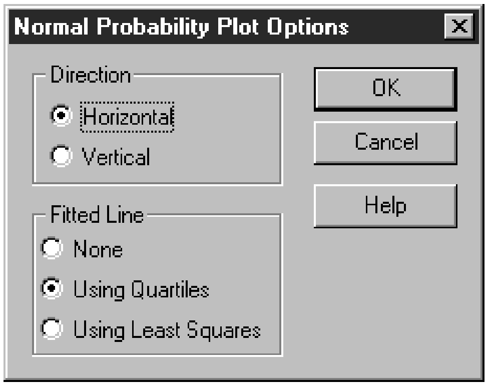


Figure 9-79. Normal Probability Plot


Normal Probability Plot Options

Direction
- ● Horizontal
- ○ Vertical

Fitted Line
- ○ None
- ● Using Quartiles
- ○ Using Least Squares

Figure 9-80. Normal Probability Plot Options Dialog Box

Saving the Results

Use the Save Results Options dialog box to choose the results you want to save. There are five selections: Trimmed Data, Select Flags, Studentized values (no deletion), Studentized values (with deletion), and Modified Z-Scores.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Iglewicz, Boris and Hoaglin, David C. 1993. *How to Detect and Handle Outliers*. Milwaukee: ASQ Quality Press.

10 Describing Categorical Data

This chapter describes the analyses you use with categorical data — data that contain numeric codes that represent discrete categories: Tabulation, Crosstabulation, and Contingency Tables.

- **Tabulation Analysis**
  This analysis produces a frequency tabulation of one variable that you can display as a table, barchart, or piechart. You use the analysis to summarize the distribution of a sample of observations.

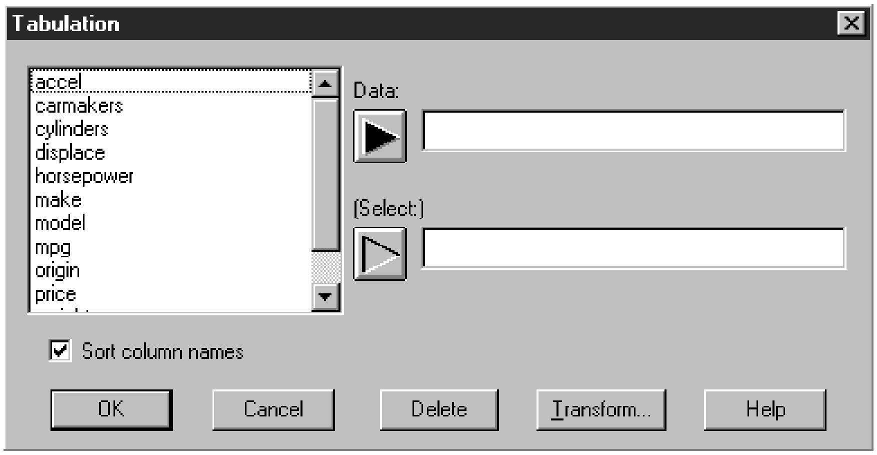
- **Crosstabulation Analysis**
  This analysis produces a frequency tabulation of two variables that you can display in various tables as well as in a barchart, mosaic plot, or a skychart. You use the analysis to summarize the joint distribution of two related variables.

- **Contingency Tables Analysis**
  This analysis determines if two classification factors are related, and if they are, how closely. The analysis allows you to enter the frequencies for each factor and displays the results in various tables as well as in a barchart, mosaic plot, or a skychart.

Using the Tabulation Analysis

The Tabulation Analysis summarizes the distribution of a single categorical variable through a frequency tabulation. You enter untabulated raw data, then the tabulation counts the number of times each value occurs. The numerical results include a summary of the analysis and a frequency table. The graphic results include a barchart and a piechart.

To access the analysis, from the menus, choose: DESCRIBE... CATEGORICAL DATA... TABULATION... (see Figure 10-1).



☑ Sort column names

Figure 10-1.   The Tabulation Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis. The summary displays the name of the selected variable, and the number of observations and unique values in the variable.

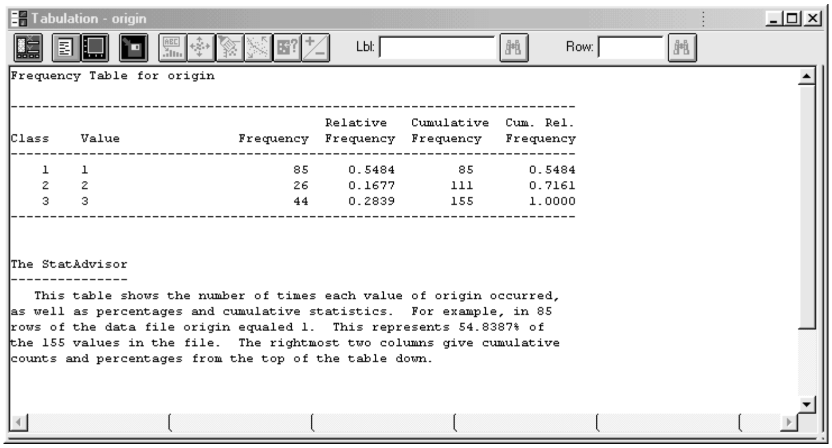
Frequency Table

The Frequency Table option creates a table that shows the number of observations that fall within each class (frequency), and the relative, cumulative, and cumulative relative frequencies (see Figure 10-2).

Graphical Options

Barchart

The Barchart option creates a plot of the frequency data. The height of each bar represents the frequencies or percentages for each category of a variable (see Figure 10-3). *For general information about barcharts, see the section"Using the Barchart Analysis," in Chapter 8, Using Basic Plots.*



```
Frequency Table for origin

------------------------------------------------------------------------
                                        Relative   Cumulative  Cum. Rel.
Class    Value               Frequency  Frequency  Frequency   Frequency
------------------------------------------------------------------------
    1    1                          85     0.5484         85      0.5484
    2    2                          26     0.1677        111      0.7161
    3    3                          44     0.2839        155      1.0000
------------------------------------------------------------------------


The StatAdvisor
---------------
   This table shows the number of times each value of origin occurred,
as well as percentages and cumulative statistics.  For example, in 85
rows of the data file origin equaled 1.  This represents 54.8387% of
the 155 values in the file.  The rightmost two columns give cumulative
counts and percentages from the top of the table down.
```

Figure 10-2. Frequency Table

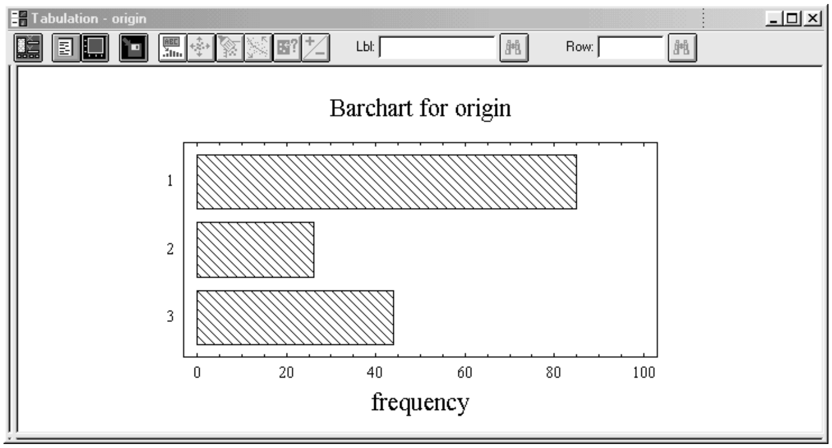
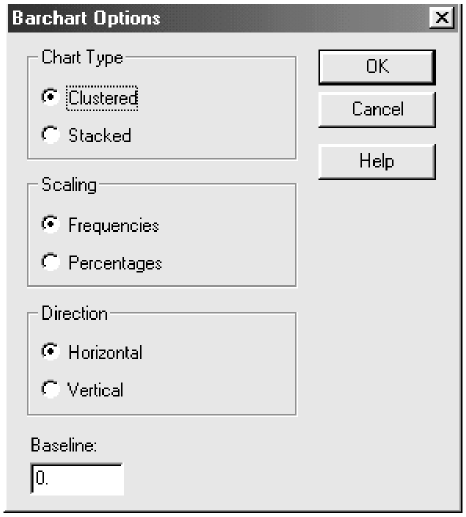


Figure 10-3. Barchart

Use the *Barchart Options* dialog box to determine a format for the bars on the chart, how the data will be plotted, and the direction of the plot. You can also enter values for the starting point of the bars (see Figure 10-4).



Barchart Options

Chart Type

● Clustered

○ Stacked

Scaling

● Frequencies

○ Percentages

Direction

● Horizontal

○ Vertical

Baseline: 0.

Figure 10-4. Barchart Options Dialog Box

Piechart

The Piechart option creates a chart that summarizes the data in a circle that is divided into segments where each segment is proportional to the number of cases in that category. You can offset a segment of the chart to display it more prominently (see Figure 10-5).

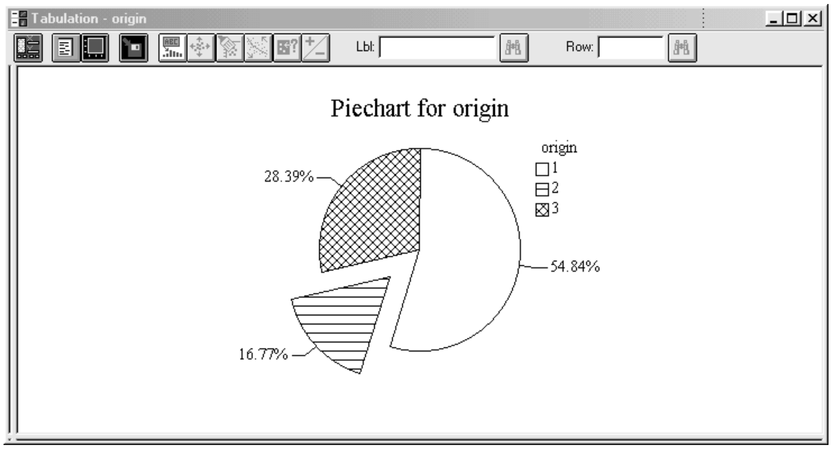
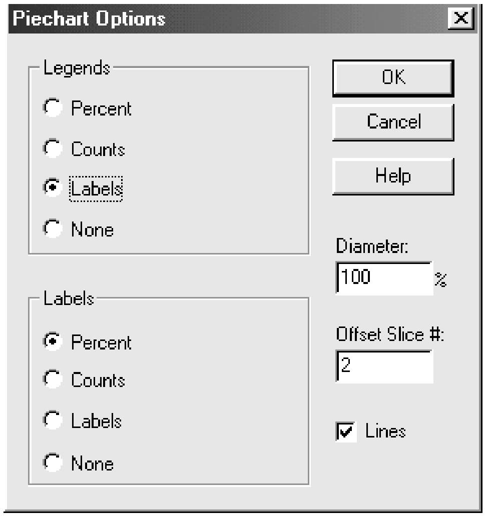


Figure 10-5. Piechart

This type of chart is useful for displaying breakdowns of percentages for up to 20 classification levels (in the factor that contains the tabulated data). *For general information on piecharts, see the section"Using the Piechart Analysis," in Chapter 8, Using Basic Plots.*

Use the *Piechart Options* dialog box to determine what the legends and labels will contain, and to enter values for the size of the circle and the segment that will be offset. You can also indicate if lines will appear from the labels to the segments (see Figure 10-6).



Piechart Options

Legends

○ Percent
○ Counts
● Labels
○ None

Labels

● Percent
○ Counts
○ Labels
○ None

Diameter: 100 %

Offset Slice #: 2

☑ Lines

Figure 10-6. Piechart Options Dialog Box

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are two options: Class Frequencies and Class Labels.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

Using the Crosstabulation Analysis

The Crosstabulation Analysis is often used simply to understand patterns of association and partial association. The analysis creates a two-way table that shows data arranged so each cell represents a unique combination of values of the two crosstabulated variables and the number of observations that fall into each combination of the values.

For example, you might conduct a study in which adults and teens were asked to choose one of two different brands of bath soap (Soap A and Soap B. The data file would look like this:

Observation	Age Group	Soap
1	Adult	A
2	Teen	A
3	Teen	B
4	Teen	B
5	Adult	A
6	Adult	B
.	.	.
.	.	.
.	.	.

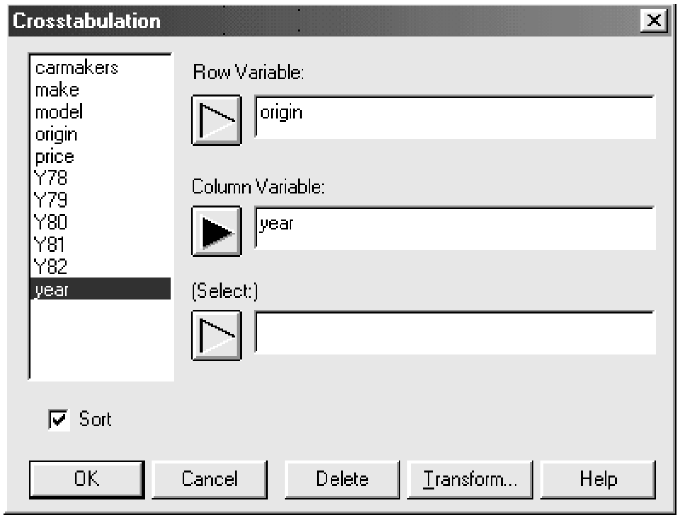
The crosstabulation would look like this:

	Soap:A	Soap:B	Total
Age Group: Adult	20 (40%)	30 (60%)	50 (50%)
Age Group: Teen	30 (60%)	20 (40%)	50 (50%)
Total	50 (50%)	50 (50%)	100 (100%)

Each cell in the table represents a combination of unique values for the two crosstabulated variables (row *Age Group,* and column *Soap*). The numbers in each cell show the number of observations that fall into each combination of values.

The crosstabulations count the number of times each unique value occurs in the first variable, then in the second. This is a useful first step in studying the relationship between two variables. To quantify or test the relationship, the analysis also allows you to create various indexes that measure the extent of association, as well as create statistical tests of the hypothesis that an association does not exist.

To access the analysis, from the menus, choose: DESCRIBE... CATEGORICAL DATA... CROSSTABULATION... (see Figure 10-7).



Crosstabulation

carmakers
make
model
origin
price
Y78
Y79
Y80
Y81
Y82
year

Row Variable: origin

Column Variable: year

(Select:)

☑ Sort

Figure 10-7. The Crosstabulation Analysis Dialog Box

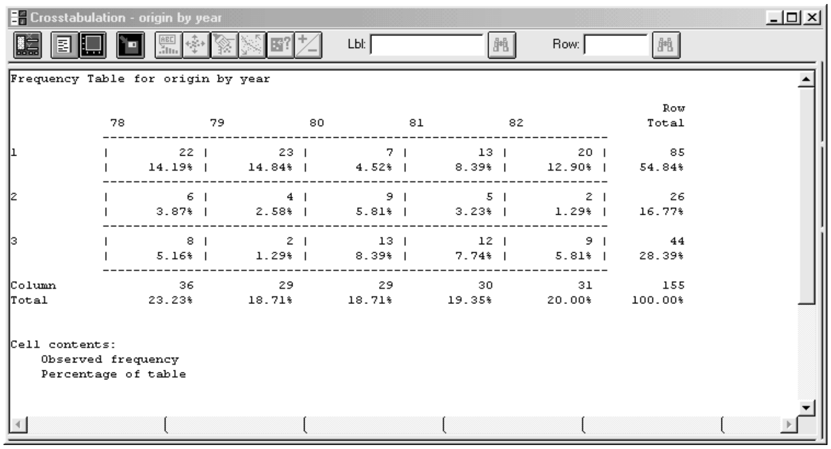
Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the names of the row and column variables, the number of observations, and the number of rows and columns.

Frequency Table

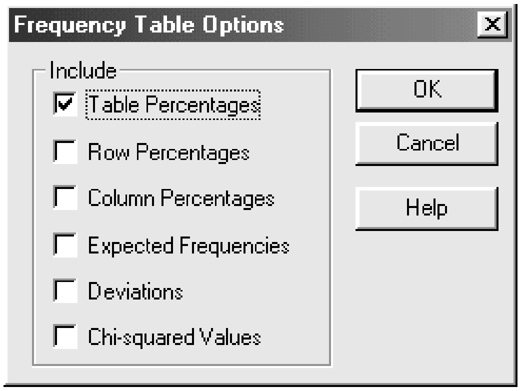
The Frequency Table option creates a 3-by-5 table that shows the number of occurrences of a value in a sample. The first number in the table is the count or frequency. The second is the percentage of the entire table represented by that cell (see Figure 10-8).



```
Frequency Table for origin by year

                                                                                     Row
             78           79           80           81           82                Total
            ------------------------------------------------------------------
1           |         22 |         23 |          7 |         13 |         20 |        85
            |     14.19% |     14.84% |      4.52% |      8.39% |     12.90% |    54.84%
            ------------------------------------------------------------------
2           |          6 |          4 |          9 |          5 |          2 |        26
            |      3.87% |      2.58% |      5.81% |      3.23% |      1.29% |    16.77%
            ------------------------------------------------------------------
3           |          8 |          2 |         13 |         12 |          9 |        44
            |      5.16% |      1.29% |      8.39% |      7.74% |      5.81% |    28.39%
            ------------------------------------------------------------------
Column                36           29           29           30           31         155
Total             23.23%       18.71%       18.71%       19.35%       20.00%     100.00%


Cell contents:
    Observed frequency
    Percentage of table
```

Figure 10-8. Frequency Table

Use the *Frequency Table Options* dialog box to determine what will appear in the Frequency Table: Table, Row, and Column Percentages; Expected Frequencies; Deviations, or Chi-Squared Values (see Figure 10-9).

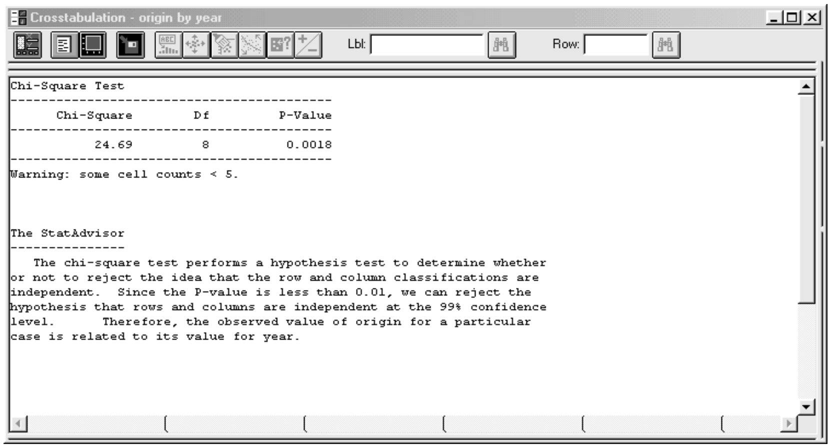


☑ Table Percentages
□ Row Percentages
□ Column Percentages
□ Expected Frequencies
□ Deviations
□ Chi-squared Values

Figure 10-9. Frequency Table Options Dialog Box

Chi-Square Test

The Chi-Square Test option performs a hypothesis test to determine if the two variables in the crosstabulation are independent of each other (see Figure 10-10). By definition, the two variables are independent if the probability that a case falls into a given cell is simply the product of the marginal probabilities of the two categories defining the cell.



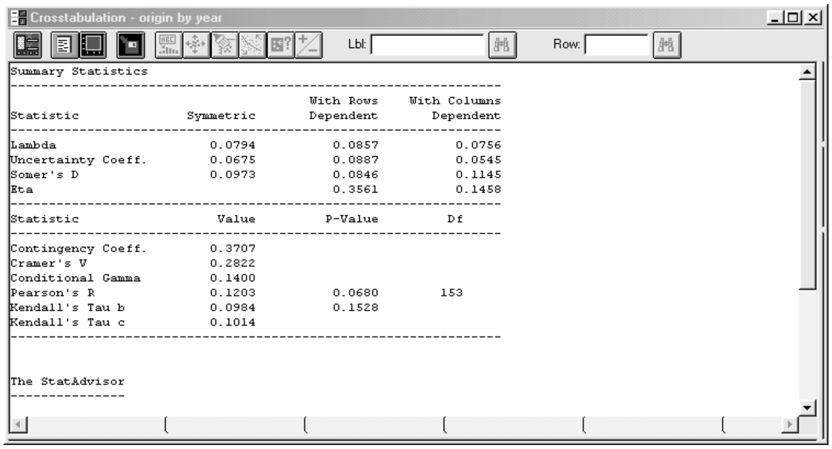
```
Chi-Square Test
------------------------------------------
      Chi-Square        Df         P-Value
------------------------------------------
           24.69         8          0.0018
------------------------------------------
Warning: some cell counts < 5.



The StatAdvisor
---------------
   The chi-square test performs a hypothesis test to determine whether
or not to reject the idea that the row and column classifications are
independent.  Since the P-value is less than 0.01, we can reject the
hypothesis that rows and columns are independent at the 99% confidence
level.      Therefore, the observed value of origin for a particular
case is related to its value for year.
```

Figure 10-10. Chi-Square Test

If the Contingency Table is a 2-by-2 table with less than 100 observations, the statistics will also include the results from Fisher's Exact test. This test is computed when a table that does not result from missing rows or columns in a larger table has a cell with an expected frequency of less than 5.

Summary Statistics

The Summary Statistics option measures the degree of association between the row and column variables. The table contains the results of various statistical tests where lambda measures the usefulness of the row (or column) factor in predicting the other factor (see Figure 10-11).



```
Summary Statistics
----------------------------------------------------------------
                                       With Rows    With Columns
Statistic              Symmetric       Dependent       Dependent
----------------------------------------------------------------
Lambda                    0.0794          0.0857          0.0756
Uncertainty Coeff.        0.0675          0.0887          0.0545
Somer's D                 0.0973          0.0846          0.1145
Eta                                       0.3561          0.1458
----------------------------------------------------------------
Statistic                  Value         P-Value         Df
----------------------------------------------------------------
Contingency Coeff.        0.3707
Cramer's V                0.2822
Conditional Gamma         0.1400
Pearson's R               0.1203          0.0680        153
Kendall's Tau b           0.0984          0.1528
Kendall's Tau c           0.1014
----------------------------------------------------------------


The StatAdvisor
---------------
```

Figure 10-11.   Summary Statistics

- **Lambda Values**
  Lambda values measure the usefulness of the row (or column) factor in predicting the other factor. Lambda always ranges between 0 and 1, where 0 is defined as being an independent variable that is of no help in predicting the dependent variable; and 1 is defined as the independent variable that correctly specifies the categories of the dependent variable. If the two variables are independent, lambda is 0, which does not always mean statistical independence. When the values of one variable are used to predict the values of the other, lambda measures association in a very precise way by imitating the reduction in error. Lambda is 0 if this type of association is absent.

- **Uncertainty Coefficients and Pearson's R**
  These statistics are calculated for numeric data only and show the degree of linear relationship between the two variables. Correlation coefficients can range from -1.00 to +1.00, where -1.00 represents a perfect negative correlation; +1.00 a perfect positive correlation; and 0.00 a lack of correlation.

- **Somers' d**
  Somers' d, d(X|Y), d(Y|X) is an asymmetric measure of association related to tau$_b$, that shows a symmetric measure of association for variables that are measured on an ordinal scale (Siegel and Castellan, 1988).

- **Eta**
  Eta is calculated when it is assumed that one variable is measured on a nominal scale and the other is measured on an interval scale. Eta is similar to the Pearson correlation coefficient, however, it is asymmetric and does not assume a linear relationship between the variables.

- **Contingency Coefficient**
  The contingency coefficient is a chi-square-based measure of the relationship between two categorical variables. It is easier to interpret than an ordinary chi-square because its range is always limited to 0 through 1, where 1 = complete independence.

- **Cramer's V and Conditional Gamma**
  These statistics are chi-square-based measures of association. Conditional gammas are displayed for three-way and higher tables. These statistics are preferred when the data contain many tied observations.

- **Kendall's Tau b and c**
  These statistics are measures of association based on a comparison of the values of both variables for all possible pairs of cases or observations. They measure the relative degree of agreement or disagreement between the two variables.

Graphical Options

Barchart

The Barchart option creates a plot of the frequency data. The height of each bar represents the frequencies or percentages for each category of a variable (see Figure 10-12). *For general information about barcharts, see the section“Using the Barchart Analysis,” in Chapter 8, Using Basic Plots.*

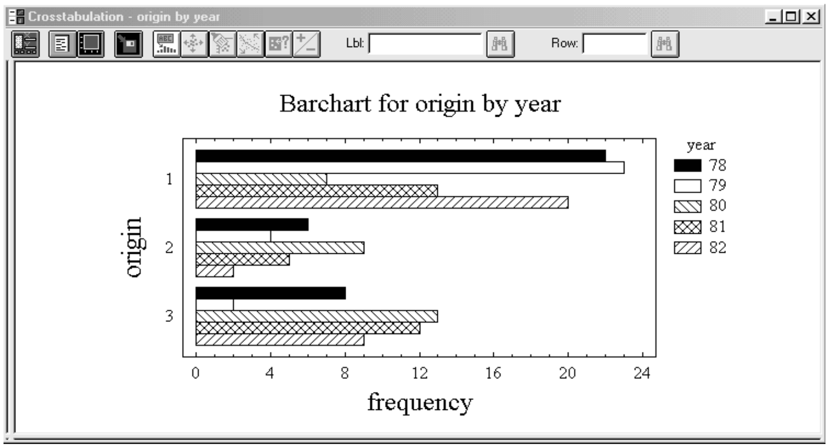


Figure 10-12. Barchart

Use the *Barchart Options* dialog box to determine a format for the bars on the chart, how the data will be plotted, and the direction of the plot. You can also enter values for the starting point of the bars (see Figure 10-4 for an example of this dialog box).

Mosaic Plot

The Mosaic Plot option creates a graphic form of the contingency table (see Figure 10-13). Rectangles, whose areas are proportional to the cell counts are constructed. The width of the bars is proportional to the percentage distribution (see Snee, 1974).

Use the *Mosaic Plot Options* dialog box to indicate the direction of the plot, either horizontal or vertical.

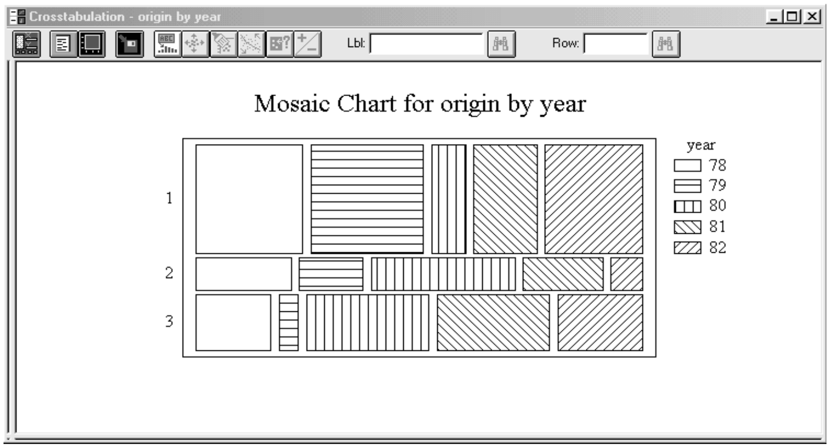

Crosstabulation - origin by year
Lbl:
Row:
Mosaic Chart for origin by year
1
2
3
year
78
79
80
81
82

Figure 10-13. Mosaic Plot

Skychart

The Skychart option creates a three-dimensional histogram, whose bar height is proportional to the number of observations in each cell (see Figure 10-14). The row variable is shown along one axis, the column variable on another axis, and the frequency on a third axis. This three-dimensional graph is an integrated view of the entire table. Its advantage is that you can evaluate the specific frequencies in each cell of the table.

Use the *Skychart Options* dialog box to indicate whether the plot will show frequency or percentage data. You can also use the Smooth/Rotate button on the Analysis toolbar to change the angle from which you view this plot. *See the section"Spinning (Rotating) Graphs," in Chapter 5, Working with Graphs and Graphics Options for a complete description of how you use the Rotate button.*

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are four selections: Cell Frequencies (single column), Row Labels, Column Labels, and Cell Frequencies (matrix).

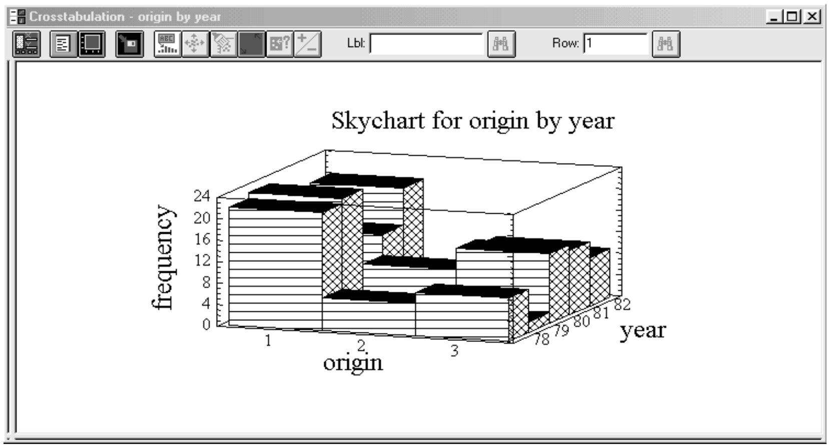


Figure 10-14. Skychart

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Everitt, B. S. 1977. *The Analysis of Contingency Tables*. New York: Routledge Chapman & Hall.

Siegel, S. and Castellan, N. J. 1988. *Nonparametric statistics for the behavioral Sciences*, second edition. New York: McGraw-Hill.

Snee, R. D. 1974. “Graphical Display of Two-way Contingency Tables,” *American Statistician*, **28**: 9-12.

Somers, R. H. 1962. “A New Symmetric Measure of Association for Ordinal Variables.” *American Sociological Review*, **27**:799-811.

Wonnacott, T. H. and Wonnacott, R. J. 1972. *Introductory Statistics*, second edition. New York: Wiley.

Using the Contingency Tables Analysis

The Contingency Tables Analysis creates separate two-way tables for each combination of the variables, which determines if two classification factors are related, and if so, how closely. You can enter frequencies for each factor.

When you analyze data as a contingency table, each variable you enter represents one column of counts for each row class. For example, if your data consists of row classes for Excellent, Average, and Poor, you might enter the following variables to represent each school grade:

grade1
grade2
grade3
grade4
. . .

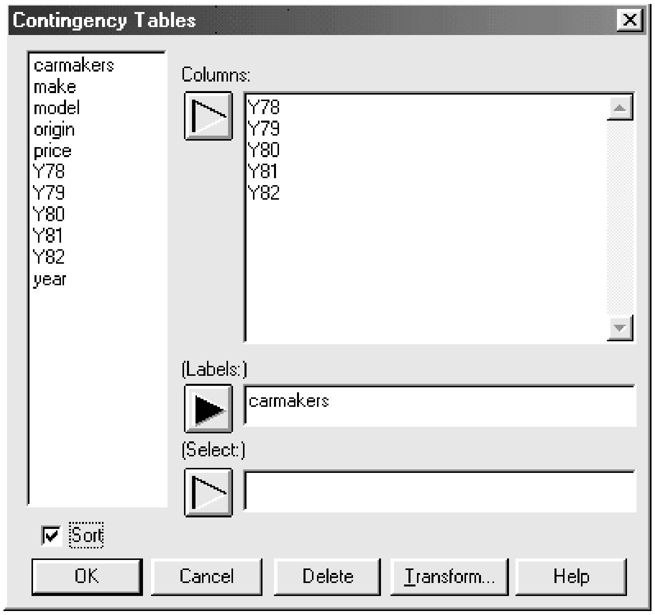
Each variable would contain three values, one value for each of the three classes: Excellent, Average, and Poor.

You can use this analysis to calculate various summary statistics, such as:

- a chi-square statistic that tests the hypothesis of independence between row and column factors (the statistic appears with the degrees of freedom and the significance level)
- the lambda values, uncertainty coefficients, and Somers' d statistics for symmetric and asymmetric cases (that is, the rows and columns are dependent)
- the contingency coefficient and other statistics.

Fisher's Exact test is performed for two-by-two matrices that have a total cell count of 100 or fewer.

To access the analysis, from the menus, choose: DESCRIBE... CATEGORICAL DATA... CONTINGENCY TABLES... (see Figure 10-15).



Columns:
Y78
Y79
Y80
Y81
Y82

(Labels:)
carmakers

(Select:)

Figure 10-15. The Contingency Tables Analysis Dialog Box

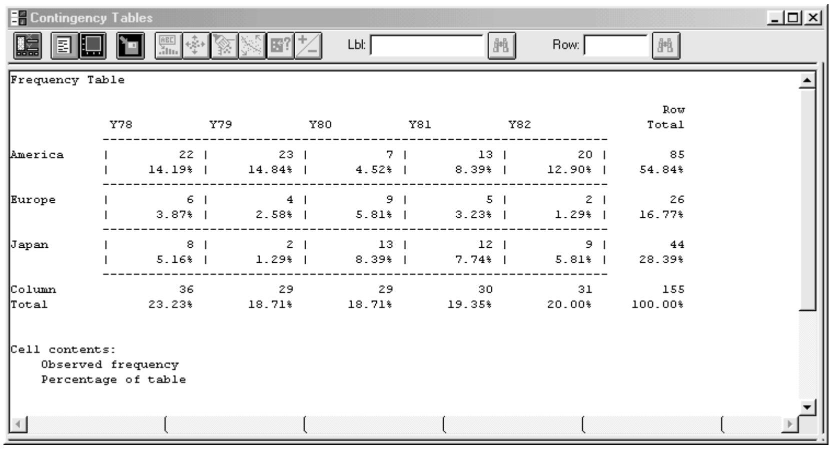
Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the names of the variables, the number of observations, and the number of rows and columns.

Frequency Table

The Frequency Table option creates the Frequency Tables in sequence (see Figure 10-16). The table displays counts, where the first number in each cell of the table is the count or frequency. The second number shows the percentage of the entire table that cell represents.



```
Frequency Table

                                                                              Row
             Y78          Y79          Y80          Y81          Y82          Total
            -----------------------------------------------------------------
America     |        22 |        23 |         7 |        13 |        20 |        85
            |    14.19% |    14.84% |     4.52% |     8.39% |    12.90% |    54.84%
            -----------------------------------------------------------------
Europe      |         6 |         4 |         9 |         5 |         2 |        26
            |     3.87% |     2.58% |     5.81% |     3.23% |     1.29% |    16.77%
            -----------------------------------------------------------------
Japan       |         8 |         2 |        13 |        12 |         9 |        44
            |     5.16% |     1.29% |     8.39% |     7.74% |     5.81% |    28.39%
            -----------------------------------------------------------------
Column               36          29          29          30          31         155
Total            23.23%      18.71%      18.71%      19.35%      20.00%     100.00%


Cell contents:
    Observed frequency
    Percentage of table
```

Figure 10-16. Frequency Table

Use the *Frequency Table Options* dialog box to determine how the percentages that appear in each cell will be calculated (see Figure 10-9 for an example of this dialog box).

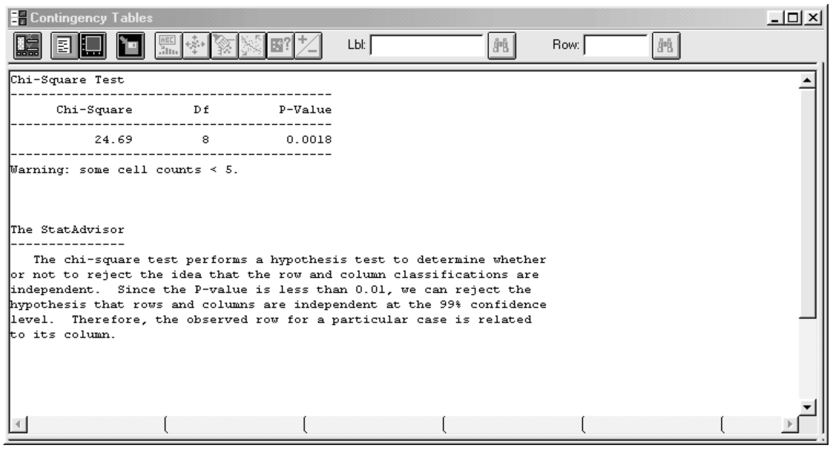
Chi-Square Test

The Chi-Square Test option performs a hypothesis test to determine if the two variables in the crosstabulation are independent of one another (see Figure 10-17). By definition, the two variables are independent if the probability that a case falls into a given cell is simply the product of the marginal probabilities of the two categories defining the cell.

If the Contingency Table is a 2 by 2 table with less than 100 observations, the statistics will also include the results from Fisher's Exact test. This test is computed when a table that does not result from missing rows or columns in a larger table has a cell with an expected frequency of less than 5.

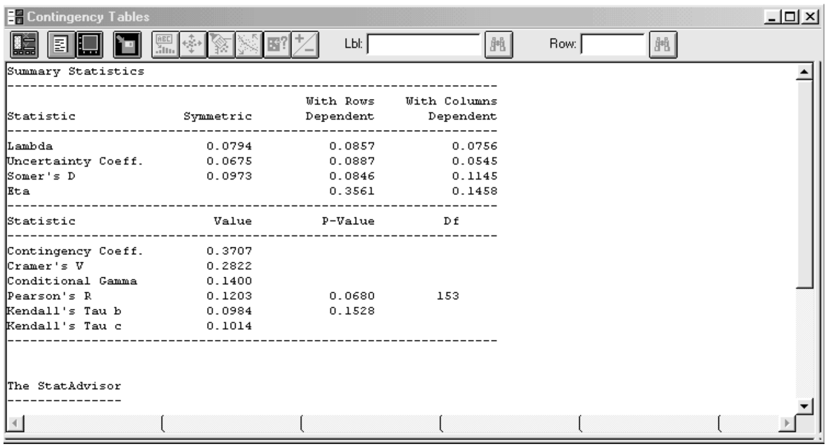
Summary Statistics

The Summary Statistics option measures the degree of association between the row and column variables. The summary contains the results of various statistical tests where lambda measures the usefulness of the row (or column) factor in predicting the other factor (see Figure 10-18).



```
Chi-Square Test
------------------------------------------
      Chi-Square        Df         P-Value
------------------------------------------
           24.69         8          0.0018
------------------------------------------
Warning: some cell counts < 5.



The StatAdvisor
---------------
   The chi-square test performs a hypothesis test to determine whether
or not to reject the idea that the row and column classifications are
independent.  Since the P-value is less than 0.01, we can reject the
hypothesis that rows and columns are independent at the 99% confidence
level.  Therefore, the observed row for a particular case is related
to its column.
```

Figure 10-17.   Chi-Square Test



```
Summary Statistics
-----------------------------------------------------------------
                                       With Rows    With Columns
Statistic               Symmetric      Dependent       Dependent
-----------------------------------------------------------------
Lambda                     0.0794         0.0857          0.0756
Uncertainty Coeff.         0.0675         0.0887          0.0545
Somer's D                  0.0973         0.0846          0.1145
Eta                                       0.3561          0.1458
-----------------------------------------------------------------
Statistic                   Value        P-Value         Df
-----------------------------------------------------------------
Contingency Coeff.         0.3707
Cramer's V                 0.2822
Conditional Gamma          0.1400
Pearson's R                0.1203         0.0680        153
Kendall's Tau b            0.0984         0.1528
Kendall's Tau c            0.1014
-----------------------------------------------------------------


The StatAdvisor
---------------
```

Figure 10-18.   Summary Statistics

- **Lambda Values**
  Lambda values measure the usefulness of the row (or column) factor in predicting the other factor. Lambda always ranges between 0 and 1, where 0 is defined as being an independent variable that is of no help in predicting the dependent variable; and 1 is defined as the independent variable that correctly specifies the categories of the dependent variable.

If the two variables are independent, lambda is 0, which does not always mean statistical independence. When the values of one variable are used to predict the values of the other, lambda measures association in a very precise way by imitating the reduction in error. Lambda is 0 if this type of association is absent.

- **Uncertainty Coefficients and Pearson's R**
  These statistics are calculated for numeric data only and show the degree of linear relationship between the two variables. Correlation coefficients can range from -1.00 to +1.00, where -1.00 represents a perfect negative correlation; +1.00 a perfect positive correlation; and 0.00 a lack of correlation.

- **Somers' d**
  Somers' d, d(X|Y), d(Y|X) is an asymmetric measure of association related to tau_b , that shows a symmetric measure of association for variables that are measured on an ordinal scale (Siegel and Castellan, 1988).

- **Eta**
  Eta is calculated when it is assumed that one variable is measured on a nominal scale and the other is measured on an interval scale. Eta is similar to the Pearson correlation coefficient, however, it is asymmetric and does not assume a linear relationship between the variables.

- **Contingency Coefficient**
  The contingency coefficient is a chi-square-based measure of the relationship between two categorical variables. It is easier to interpret than an ordinary chi-square because its range is always limited to 0 through 1, where 1 = complete independence.

- **Cramer's V and Conditional Gamma**
  These statistics are chi-square-based measures of association. Conditional gammas are displayed for three-way and higher tables. These statistics are preferred when the data contain many tied observations.

- **Kendall's Tau b and c**
  These statistics are measures of association based on a comparison of the values of both variables for all possible pairs of cases or observations. They measure the relative degree of agreement or disagreement between the two variables.

Graphical Options

Barchart

The Barchart option creates a plot of the frequency data. The height of each bar represents the frequencies or percentages for each category of a variable (see Figure 10-19). *For general information about barcharts, see the section“Using the Barchart Analysis,” in Chapter 8, Using Basic Plots.*

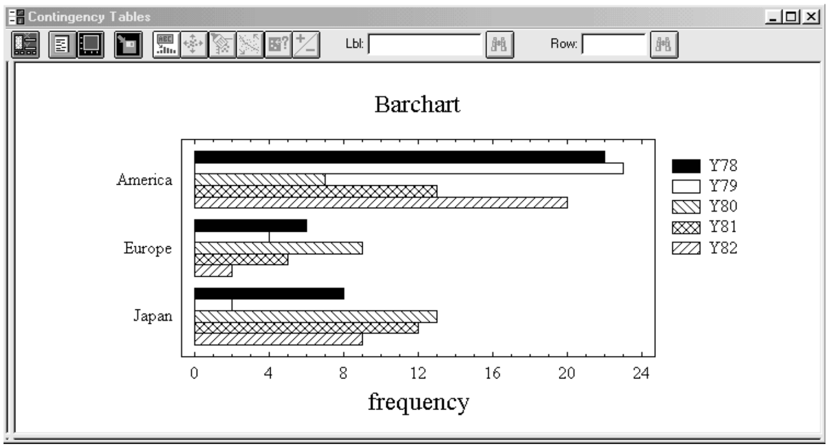


Figure 10-19. Barchart

Use the *Barchart Options* dialog box to determine a format for the bars on the chart, how the data will be plotted, and the direction of the plot. You can also enter values for the starting point of the bars (see Figure 10-4 for an example of this dialog box).

Mosaic Plot

The Mosaic Plot option creates a graphic form of the contingency table (see Figure 10-20). Rectangles, whose areas are proportional to the cell counts are constructed. The width of the bars is proportional to the percentage distribution (see Snee, 1974).

Use the *Mosaic Plot Options* dialog box to indicate the direction of the plot, either horizontal or vertical.

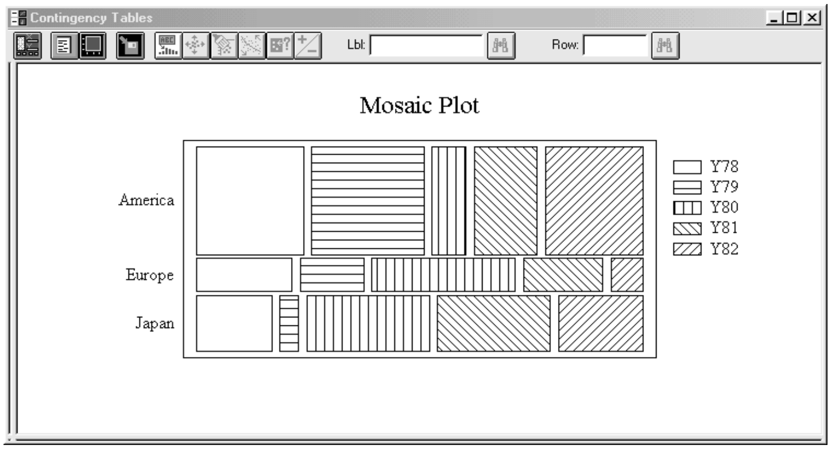


Figure 10-20. Mosaic Plot

Skychart

The Skychart option creates a three-dimensional histogram, whose bar height is proportional to the number of observations in each cell (see Figure 10-21). The row variable is shown along one axis, the column variable on another axis, and the frequency on a third axis. This three-dimensional graph is an integrated view of the entire table. Its advantage is that you can evaluate the specific frequencies in each cell of the table.

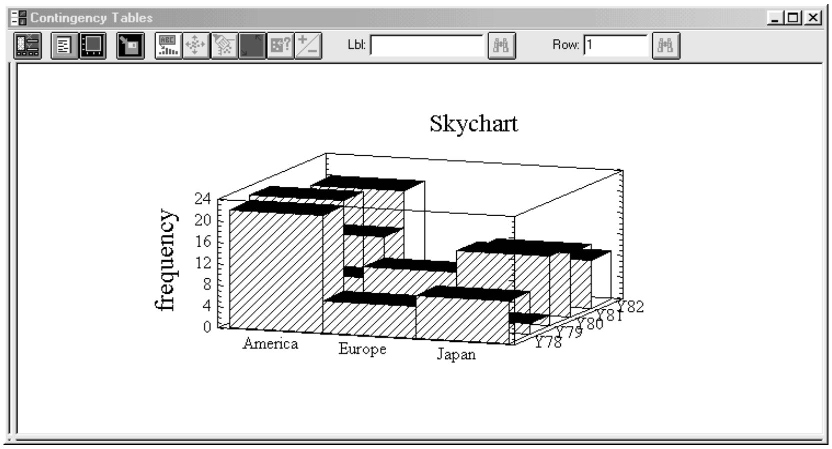


Figure 10-21. Skychart

Use the *Skychart Options* dialog box to indicate whether the plot will show frequency or percentage data. You can also use the Smooth/Rotate button on the Analysis toolbar to change the angle from which you view this plot. *See the section"Spinning (Rotating) Graphs," in Chapter 5, Working with Graphs and Graphics Options for a complete description of how you use the Rotate button.*

References

Everitt, B. S. 1977. *The Analysis of Contingency Tables*. London: Chapman and Hall.

Freund, J. E. and Williams, F. J. 1977. *Elementary Business Statistics, The Modern Approach.* New Jersey: Prentice-Hall.

Hays, W. L. 1981. *Statistics*, third edition. New York: Holt, Rinehart and Winston.

Siegel, S. and Castellan, N. J. 1988. *Nonparametric Statistics for the Behavioral Sciences*, second edition. New York: McGraw-Hill.

Somers, R. H. 1962. "A New Symmetric Measure of Association for Ordinal Variables." *American Sociological Review*, **27**:799-811.

Wonnacott, T. H. and Wonnacott, R. J. 1972. *Introductory Statistics,* second edition. New York: Wiley.

11 Working with Probability Distributions

This chapter contains analyses for using probability distributions, creating a variety of probability plots, and using uncensored and censored data to perform distribution fitting.

Using the Probability Distributions Analysis

Before you can understand or apply methods of inferential statistics to data, you need to be familiar with the science of uncertainty or *probability*. Probability is a mathematical model describing the probability of observing various values of a random variable so you can evaluate and control the likelihood that a statistical inference is correct. A random variable, also called a *statistical variable*, is a function that maps the outcome of an experiment to a real number.

In daily life, probability is normally defined as the relative frequency of the occurrence of an event that can be repeated many times. For example, you might say that the chance of winning the lottery is "one in a million." Or, if you repeatedly sample temperatures from a process and get values below 150 degrees half the time, the probability of a reading below 150 degrees is equal to 0.5 or 50 percent.

Just as a table of frequencies is called a *frequency distribution*, a table of probabilities is called a *probability distribution*. The distribution is determined by probabilities associated with the underlying sample space and by the sampling design. Many statistical applications involve estimating the mean and standard deviation of a probability distribution based on sample data, while the probability distribution of a discrete random variable is based on the probability associated with each possible value assumed by the random variable.

The probability distributions in STATGRAPHICS *Plus* contain functions that let you perform three basic operations for each of 24 different distributions:

- generate random numbers

- calculate probabilities
- create plots of probability and cumulative distributions.

The 24 distributions are

- **Bernoulli**
  A distribution whose outcome has only two possibilities: success or failure; for example, heads or tails; good or bad; defective or nondefective. The probability of a success remains the same from trial to trial. Data fit to this distribution should have values of only 0 or 1.

- **Binomial**
  A distribution for observing the number of successes in a fixed number (sequence) of independent or Bernoulli trials. There are only two possible outcomes for each trial. When you use this distribution, you must choose the number of trials (experiments). Data fit to this distribution should be integers greater than or equal to 0.

  When you use this distribution, the Number of Trials text box is activated so you can enter the number of trials that will be in the analysis.

- **Discrete Uniform**
  A distribution that allocates equal probabilities to all integer values between a Lower and an Upper limit. Data fit to this distribution should be integers.

- **Geometric**
  A distribution that characterizes the number of failures that occur before the first success in a series of Bernoulli trials; a special case of Negative Binomial distribution, where $k = 1$. Data fit to this distribution should be integers.

- **Hypergeometric**
  A distribution that arises when a random selection is made between objects of two distinct types (success, fail). The sampling occurs without replacement; that is, each time an item is drawn and studied, it is not placed back into the population. The distribution gives the probability of the number of successes. Data fit to this distribution should be integers greater than or equal to 0.

  When you use this distribution, the Number of Trials and Population Size text boxes are activated so you can enter numbers for the number of trials that will be in the analysis, as well as for the size of the population.

- **Negative Binomial**
  A distribution that characterizes the number of failures before the *k*th success in a series of Bernoulli trials. When you use this distribution, you must declare the number of successes. Data fit to this distribution should be integers greater than 0.

  When you use this distribution, the Number of Successes text box is activated so you can enter the number of successes that will be in the analysis.

- **Poisson**
  A distribution that counts the number of times a certain event occurs during a given unit of time or in a given area of volume (or weight, distance, or any other unit of measurement). The probability that an event occurs in a given unit of time, area, or volume is the same for all the units. The number of events that occur is independent of the number that occur in other units. Data fit to this distribution should be integers greater than or equal to 0.

- **Beta**
  A distribution useful for random variables that are constrained to lie between 0 and 1; characterized by two parameters: Shape 1 and Shape 2.

- **Cauchy**
  A distribution that fits data that follow a Cauchy distribution. The distribution's probability density function has no mean and an infinite variance. It is characterized by two parameters: Mode and Scale. Data fit to this distribution should be continuous with a Mode between -infinity to +infinity and a Scale greater than 0.

- **Chi-Square**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by one parameter: Degrees of Freedom. This distribution is used most often as the sampling distribution for various statistical tests.

- **Erlang**
  A distribution useful for random variables that are constrained to be greater than 0, such as the time required to complete a task; characterized by two parameters: Shape and Scale. This distribution is a special case of the Gamma distribution, which requires that the Shape parameter be an integer.

- **Exponential**
  A distribution that fits time-series data, such as arrival times, where arrivals are expected at a constant rate; useful for random variables that are constrained to be greater than 0. This distribution is a special case of both the Gamma and the Weibull distributions.

- **Extreme Value**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by two parameters: Mode and Scale. Also known as a Gumbels' distribution.

- **F (Variance Ratio)**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by two parameters: Numerator Degrees of Freedom and Denominator Degrees of Freedom. It is often used as the distribution for test statistics that are created as variance ratios.

- **Gamma**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by two parameters: Shape and Scale. This distribution is often used to model data that are positively skewed, such as the time required to complete a task.

- **Laplace**
  A distribution useful for random variables from a distribution that is more peaked than a Normal distribution; characterized by two parameters: Mean and Scale. This distribution is sometimes called the *double exponential* distribution because it looks like an exponential distribution with a mirror image.

- **Logistic**
  A distribution useful for random variables that are not constrained to be greater than or equal to 0; characterized by two parameters: Mean and Standard Deviation.

- **Lognormal**
  A distribution useful for processes in which the value is a random proportion of the previous value, such as personal incomes or particle sizes from breakage processes. The log of data that follow the lognormal distribution are normally distributed. The distribution is positively skewed and can take on various shapes. Data fit to this distribution should

be values greater than 0; characterized by two parameters: Mean and Standard Deviation.

- **Normal**
  A distribution useful in instances when you plot a Frequency Histogram of the data and the bars form a common, bell-shaped curve.

- **Pareto**
  A distribution with a decreasing density function. One parameter, Shape, is necessary to specify the distribution. Data fit to this distribution should be values greater than 0.

- **Student's *t***
  A distribution useful in forming confidence intervals when the variance is unknown, testing to determine if two sample means are significantly different, or testing to determine the significance of coefficients in a regression. The distribution is similar in shape to a Normal distribution. The mean of the *t* distribution is always equal to 0, while the standard deviation is usually slightly greater than 1. One parameter, Degrees of Freedom, is necessary to completely specify the distribution.

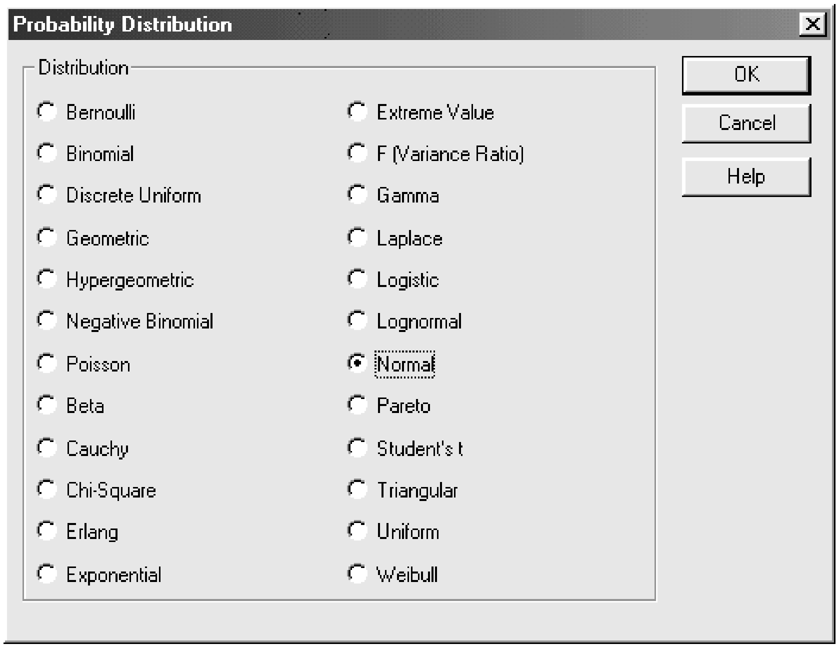
- **Triangular**
  A distribution useful for random variables that are constrained to lie between two fixed limits. Unlike the Uniform distribution, in which all the values between the limits are equally likely, the Triangular distribution peaks at some value between the two limits. This distribution is characterized by three parameters: Lower Limit, Central Value (Mode), and Upper Limit.

- **Uniform**
  A distribution useful for characterizing data that range over an interval of values, each of which is equally likely. The distribution is completely determined by the smallest possible value, *a*, and the largest possible value, *b*. The mean equals (*a* + *b*)/2, while the variance equals ((*b* - *a*)^2)/12. For discrete data, there is a related discrete uniform distribution.

- **Weibull**
  A distribution useful for random variables that are constrained to be greater than 0; characterized by two parameters: Shape and Scale. Because its failure-rate curves can take various shapes, it is an

appropriate model for product failures. The distribution is a generalization of an Exponential distribution.

To access the analysis, from the menus, choose: DESCRIBE... DISTRIBUTIONS... PROBABILITY DISTRIBUTIONS... (see Figure 11-1).



Probability Distribution

Distribution

○ Bernoulli
○ Binomial
○ Discrete Uniform
○ Geometric
○ Hypergeometric
○ Negative Binomial
○ Poisson
○ Beta
○ Cauchy
○ Chi-Square
○ Erlang
○ Exponential
○ Extreme Value
○ F (Variance Ratio)
○ Gamma
○ Laplace
○ Logistic
○ Lognormal
● Normal
○ Pareto
○ Student's t
○ Triangular
○ Uniform
○ Weibull

Figure 11-1. The Probability Distributions Analysis Dialog Box

Tabular Options

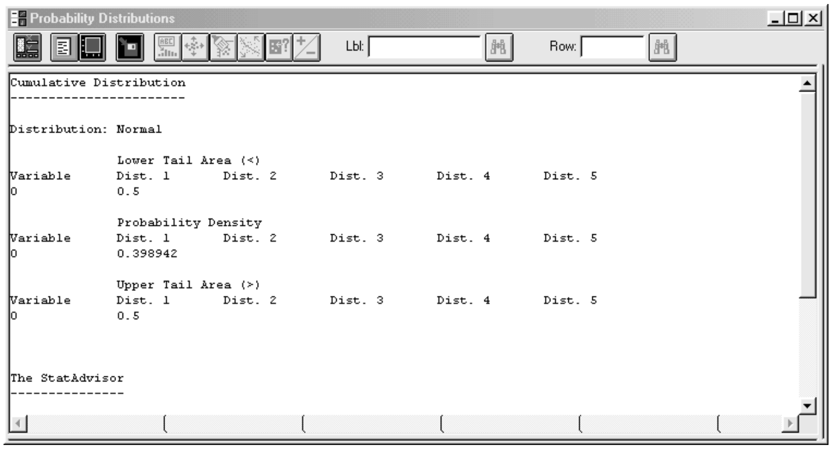
Analysis Summary

The Analysis Summary option creates and displays a summary of the distribution you selected; therefore, the contents of this pane will vary. The summary lists the parameters for the selected distribution as well as other statistics, depending on the distribution you are using.

Use the *Options* dialog boxes for the individual distributions to change the values for the parameters. *See Online Help for a description of these dialog* *boxes; for example, see Bernoulli Options Dialog Box, Binomial Options Dialog Box ... Uniform Options Dialog Box, Weibull Options Dialog Box.*

Cumulative Distribution

The Cumulative Distribution option creates a summary of the evaluation for the cumulative distribution you selected (see Figure 11-2). The summary includes the tail areas for up to five critical values of the distribution. It also displays the height of the probability density function at the value you selected.



```
Cumulative Distribution
-----------------------

Distribution: Normal

              Lower Tail Area (<)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
0             0.5

              Probability Density
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
0             0.398942

              Upper Tail Area (>)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
0             0.5



The StatAdvisor
---------------
```

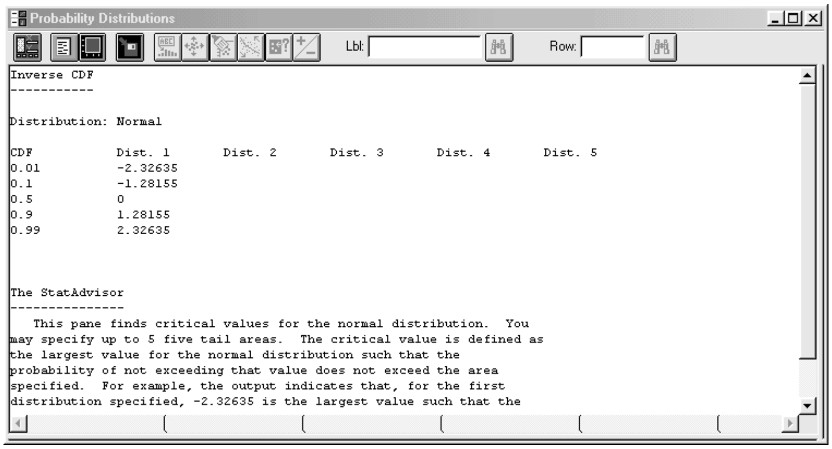
Figure 11-2. Cumulative Distribution

Use the *Cumulative Distribution Options* dialog box to enter values for up to five random variables.

Inverse CDF

The Inverse CDF option creates a summary of the critical values for the distribution you selected (see Figure 11-3). The summary includes the critical values for up to five tail areas of the distribution.

Use the *Inverse CDF Options* dialog box to enter values for the tail areas.

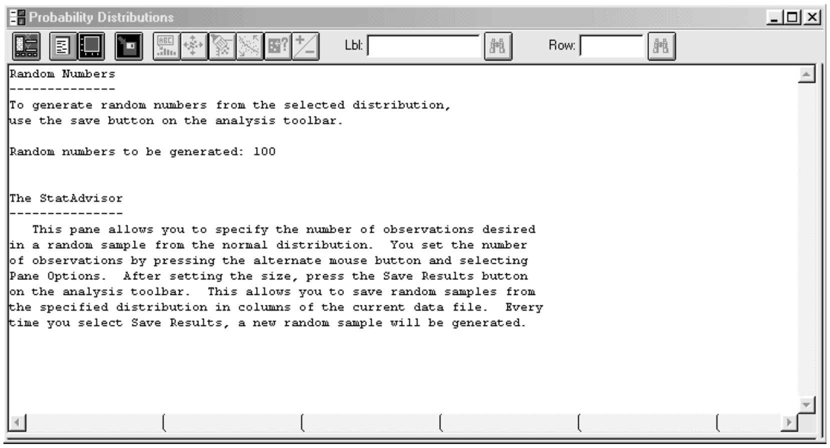


```
Inverse CDF
-----------

Distribution: Normal

CDF          Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
0.01         -2.32635
0.1          -1.28155
0.5          0
0.9          1.28155
0.99         2.32635



The StatAdvisor
---------------
   This pane finds critical values for the normal distribution.  You
may specify up to 5 five tail areas.  The critical value is defined as
the largest value for the normal distribution such that the
probability of not exceeding that value does not exceed the area
specified.  For example, the output indicates that, for the first
distribution specified, -2.32635 is the largest value such that the
```

Figure 11-3.   Inverse CDF

Random Numbers

The Random Numbers option generates random numbers from the distribution you selected (see Figure 11-4). You can save the values for the random samples for future use; for example, each time you save the results using the Save Results dialog box, the program generates a new random sample.



```
Random Numbers
--------------
To generate random numbers from the selected distribution,
use the save button on the analysis toolbar.

Random numbers to be generated: 100


The StatAdvisor
---------------
   This pane allows you to specify the number of observations desired
in a random sample from the normal distribution.  You set the number
of observations by pressing the alternate mouse button and selecting
Pane Options.  After setting the size, press the Save Results button
on the analysis toolbar.  This allows you to save random samples from
the specified distribution in columns of the current data file.  Every
time you select Save Results, a new random sample will be generated.
```

Figure 11-4.   Random Numbers

Use the *Random Numbers Options* dialog box to enter the number of observations that will be included in the random sample.

Graphical Options

Density/Mass Function Plot

The Density/Mass Function Plot option creates a plot of the probability density function for the distribution you are evaluating (see Figure 11-5). The height of the function indicates the probability of obtaining various values for the distribution you selected.

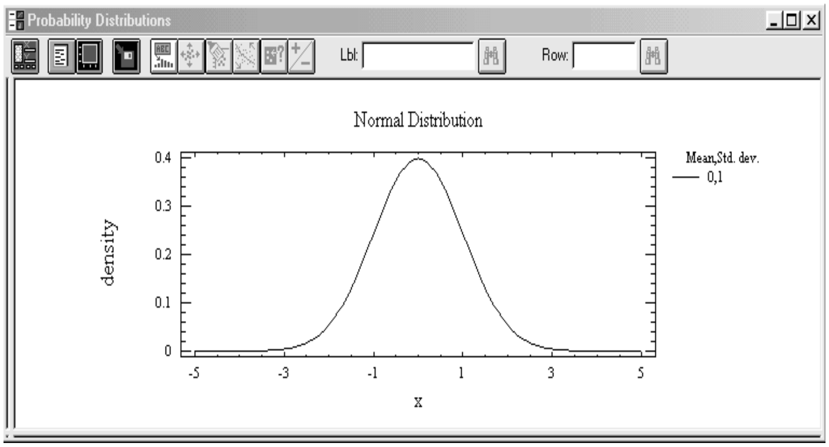


Figure 11-5. Density/Mass Function Plot

CDF Plot

The CDF Plot option creates a plot of the cumulative probability distribution for the distribution you are evaluating (see Figure 11-6).

Survivor Function Plot

The Survivor Function Plot option creates a plot of the survival probability function for the distribution you are evaluating (see Figure 11-7). The

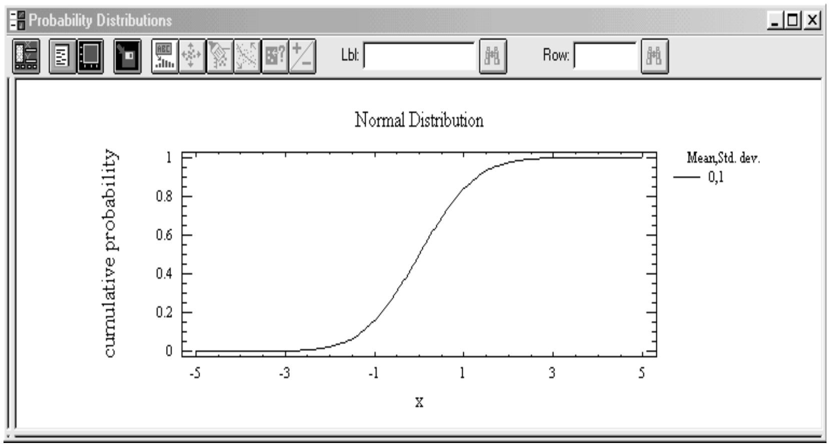


Figure 11-6. CDF Plot

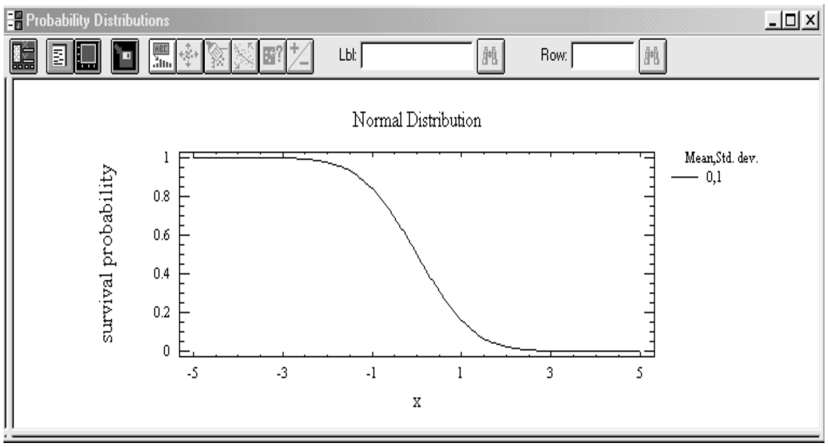


Figure 11-7. Survivor Function Plot

function indicates the probability of obtaining a value greater than or equal to the values on the X-axis.

Log Survivor Function Plot

The Log Survivor Function Plot option creates a plot of the log survival probability function for the distribution you are evaluating (see Figure 11-8). The function indicates the probability of obtaining a value greater than or equal to the values on the X-axis.

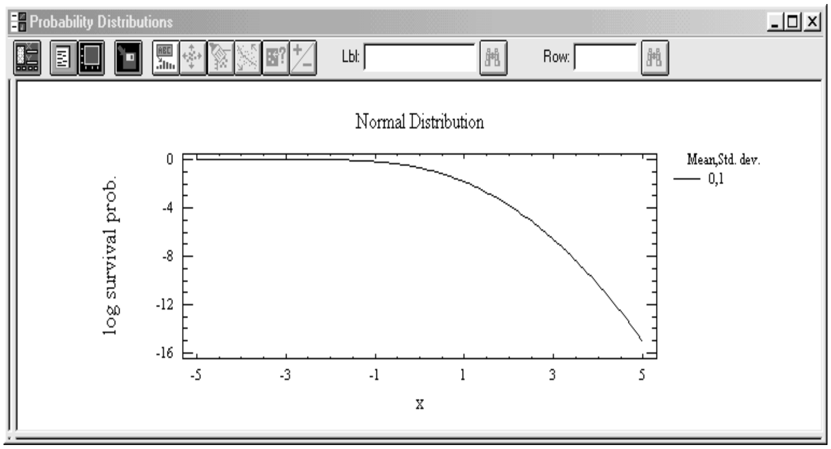


Figure 11-8. Log Survivor Function Plot

Hazard Function Plot

The Hazard Function Plot option creates a plot of the hazard function for the distribution you are evaluating (see Figure 11-9). The hazard function is equal to the probability density function divided by the survival function. When you are modeling lifetime data, the hazard function represents the instantaneous failure rate.

Using the Probability Plots Analysis

The Probability Plots Analysis allows you to create seven different types of probability plots to help determine if one variable comes from a particular type of distribution. After you create and examine the plots, you can fit a distribution using a Distribution Fitting Analysis for either uncensored or censored data.

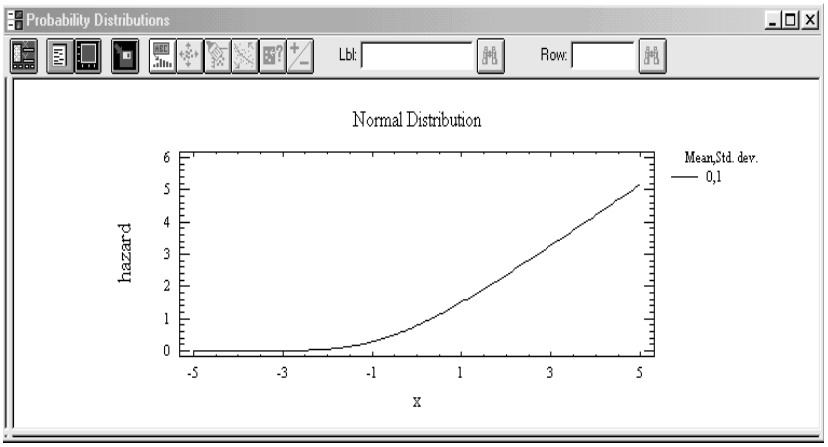
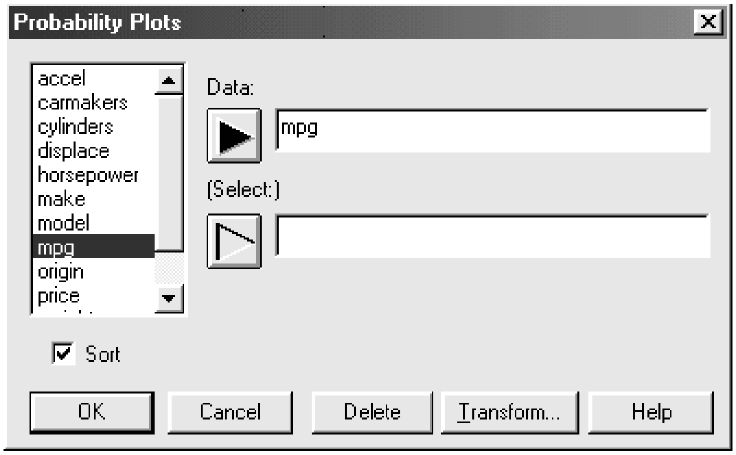


Figure 11-9. Hazard Function Plot

To access the analysis, from the menus, choose: DESCRIBE... DISTRIBUTIONS... PROBABILITY PLOTS... (see Figure 11-10).


Probability Plots

Data: mpg

(Select:)

☑ Sort

Figure 11-10. The Probability Plots Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary creates a summary that shows the name of the selected variable, the number of non-missing observations, and the number of values below the minimum and above the maximum.

Use the *Probability Plot Options* dialog box to enter a number for the values above the maximum and a number for the values below the minimum.

Graphical Options

Uniform Plot

The Uniform Plot option creates a plot for the variable you selected that helps determine if the data can be reasonably modeled using a Uniform distribution (see Figure 11-11).

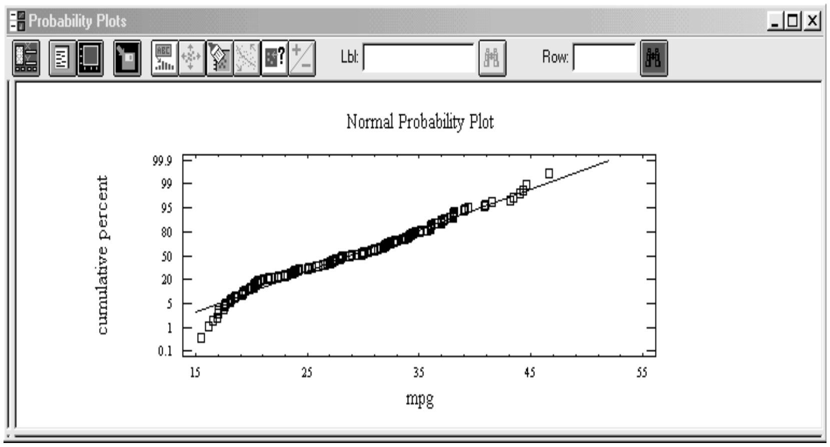


Figure 11-11. Uniform Plot

To create the plot, the program sorts the values from smallest to largest, then plots them versus the values (*i* - 0.375)/(*n* + 0.25), where *n* is the size of the sample. If the data come from a Uniform distribution, the points will lie approximately along a straight line. A reference line is superimposed on the plot to determine the closeness of the points to the line. The reference line was fit to the plot using least squares.

Normal Plot

The Normal Plot option creates a plot for the variable you selected that helps determine if the data can be reasonably modeled using a Normal distribution (see Figure 11-12).

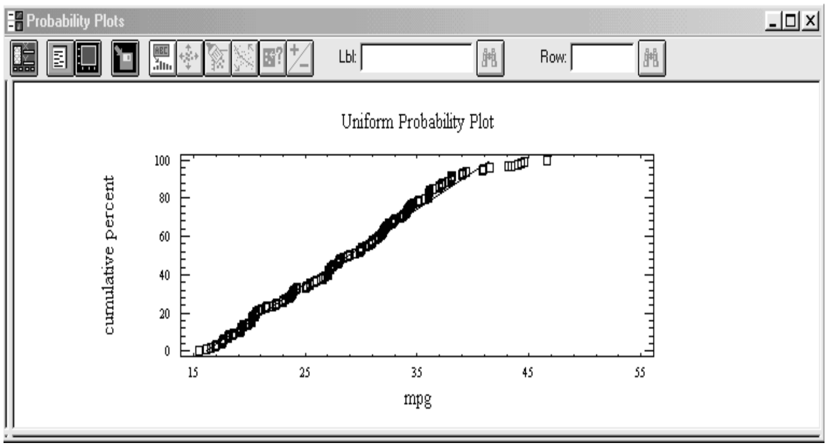


Figure 11-12. Normal Plot

To create the plot, the program sorts the values from smallest to largest, then plots them versus the values (*i*- 0.375)/(*n* + 0.25), where *n* is the size of the sample. If the data come from a Normal distribution, the points will lie approximately along a straight line. A reference line is superimposed on the plot to determine the closeness of the points to the line. The reference line was fit to the plot using least squares.

Lognormal Plot

The Lognormal Plot option creates a plot for the variable you selected that helps determine if the data can be reasonably modeled using a Lognormal distribution (see Figure 11-13).

To create the plot, the program sorts the values from smallest to largest, then plots them versus the values (*i* - 0.375)/(*n* + 0.25), where *n* is the size of the sample. If the data come from a Lognormal distribution, the points will lie approximately along a straight line. A reference line is superimposed on the plot to determine the closeness of the points to the line. The reference line was fit to the plot using least squares.

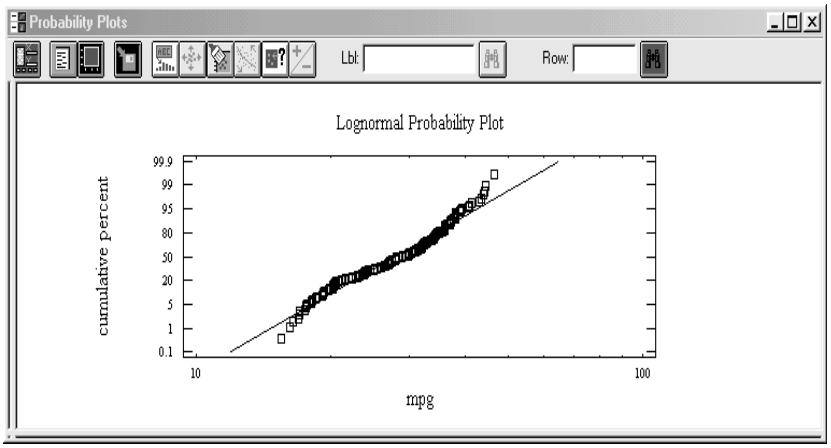


Figure 11-13. Lognormal Plot

Weibull Plot

The Weibull Plot option creates a plot for the variable you selected that helps determine if the data can be reasonably modeled using a Weibull distribution (see Figure 11-14).

To create the plot, the program sorts the values from smallest to largest, then plots them versus the values (*i* - 0.375)/(*n* + 0.25), where *n* is the size of the sample. If the data come from a Weibull distribution, the points will lie approximately along a straight line. A reference line is superimposed on the plot to determine the closeness of the points to the line. The reference line was fit to the plot using least squares.

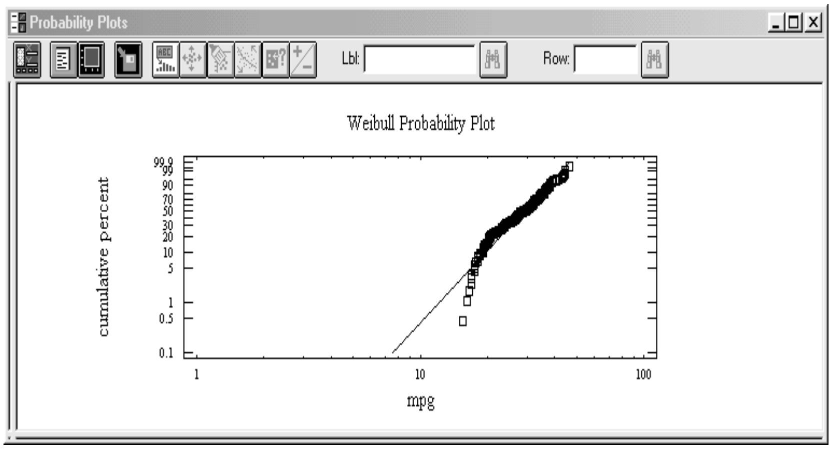


Figure 11-14. Weibull Plot

Extreme Value Plot

The Extreme Value Plot option creates a plot for the variable you selected that helps determine if the data can be reasonably modeled using an Extreme Value distribution (see Figure 11-15).

To create the plot, the program sorts the values from smallest to largest, then plots them versus the values (i - 0.375)/(n + 0.25), where n is the size of the sample. If the data come from an Extreme Value distribution, the points will lie approximately along a straight line. A reference line is superimposed on the plot to determine the closeness of the points to the line. The reference line was fit to the plot using least squares.

Logistic Plot

The Logistic Plot option creates a plot for the variable you selected that helps determine if the data can be reasonably modeled using a Logistic distribution (see Figure 11-16).

To create the plot, the program sorts the values from smallest to largest, then plots them versus the values (i - 0.375)/(n + 0.25), where n is the size of the sample. If the data come from a Logistic distribution, the points will lie approximately along a straight line. A reference line is superimposed on the plot to determine the closeness of the points to the line. The reference line was fit to the plot using least squares.

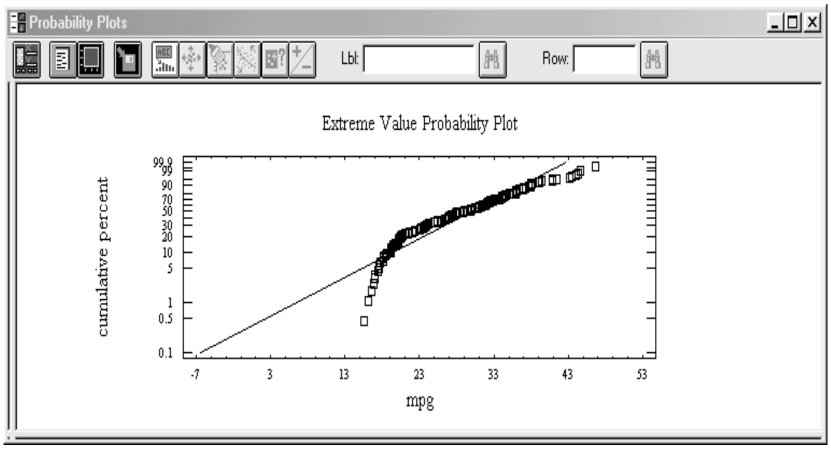


Figure 11-15. Extreme Value Plot

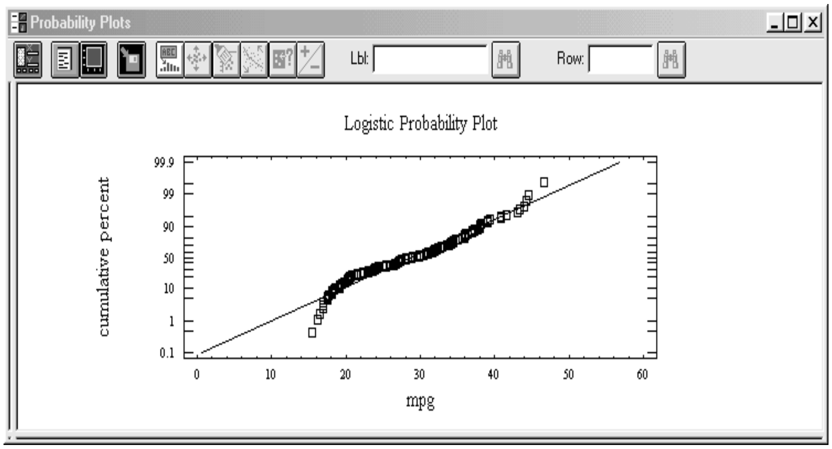


Figure 11-16. Logistic Plot

Exponential Plot

The Exponential Plot option creates a plot for the variable you selected that helps determine if the data can be reasonably modeled using an Exponential distribution (see Figure 11-17).

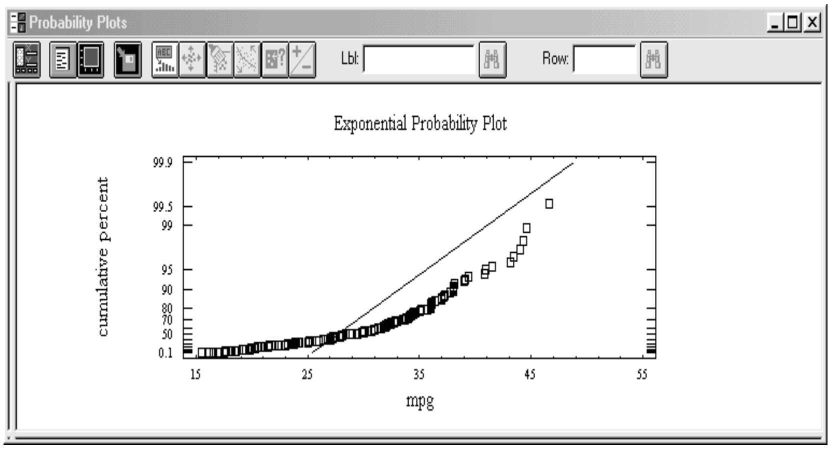


Figure 11-17. Exponential Plot

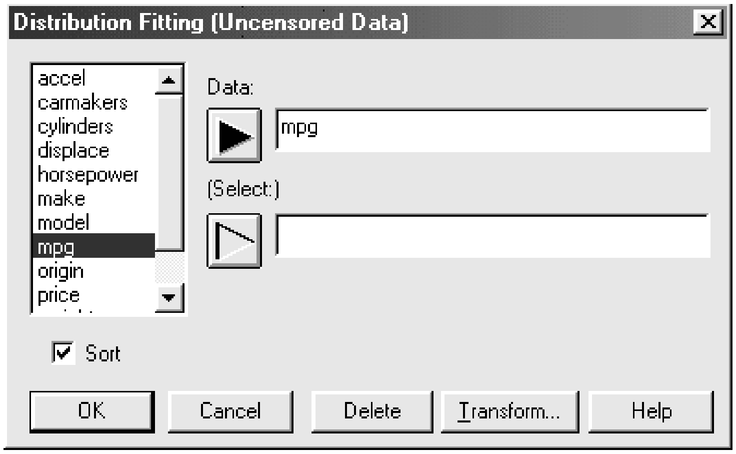
To create the plot, the program sorts the values from smallest to largest, then plots them versus the values (*i* - 0.375)/(*n* + 0.25), where *n* is the size of the sample. If the data come from an Exponential distribution, the points will lie approximately along a straight line. A reference line is superimposed on the plot to determine the closeness of the points to the line. The reference line was fitted to the plot using least squares.

Fitting Distributions Using Uncensored Data

Raw data are seldom suitable for performing an analysis, so it is necessary to convert it into a suitable form that can undergo meaningful analysis. The goal is to understand the random variability that exists in each measurement of the data. Uncensored data are data that are present throughout the entire duration of an experiment, or data that do not have to be excluded from an experiment for any reason.

The Uncensored Data Analysis allows you to create a relevant summary by fitting one of the 24 probability distribution functions in STATGRAPHICS *Plus* to a set of data. The analysis provides a way to determine if uncensored data follow a Normal (the default), or another type of distribution. The analysis calculates and displays estimated parameters for each distribution you choose.

To access the analysis, from the menus, choose: DESCRIBE... DISTRIBUTIONS... DISTRIBUTION FITTING (UNCENSORED DATA)... (see Figure 11-18).



Data: mpg

☑ Sort

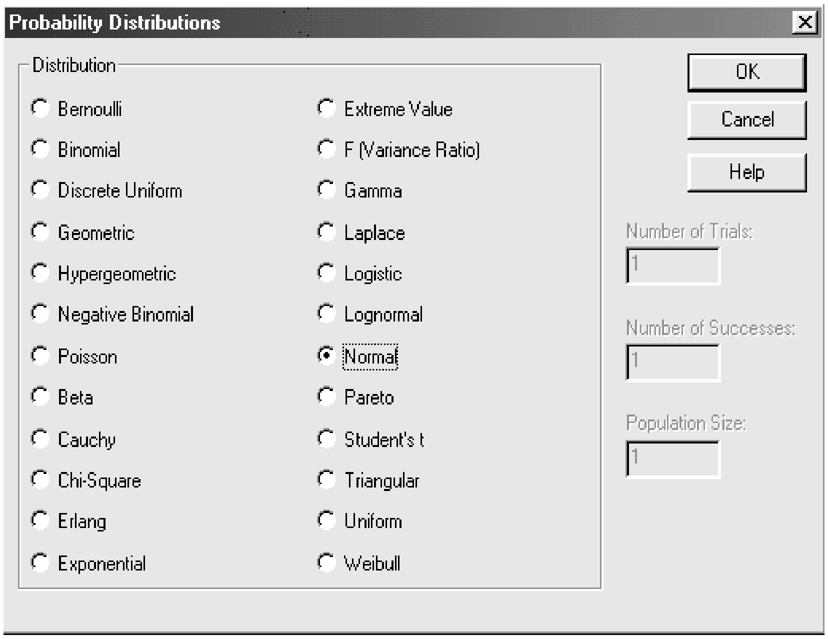
Figure 11-18. The Distribution Fitting (Uncensored Data) Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that includes the name of the variable you selected, the number of values and their range, the name of the distribution you are currently using, and the estimated parameters for that distribution.

Use the *Probability Distributions Options* dialog box to choose a different distribution, to enter the number of trials, and to enter the number of successes (see Figure 11-19). Each of the 24 distributions has its own options dialog box. *See Online Help for a description of these dialog boxes; for example, see Bernoulli Options Dialog Box, Binomial Options Dialog Box ... Uniform Options Dialog Box, Weibull Options Dialog Box.*



Distribution

○ Bernoulli
○ Binomial
○ Discrete Uniform
○ Geometric
○ Hypergeometric
○ Negative Binomial
○ Poisson
○ Beta
○ Cauchy
○ Chi-Square
○ Erlang
○ Exponential
○ Extreme Value
○ F (Variance Ratio)
○ Gamma
○ Laplace
○ Logistic
○ Lognormal
● Normal
○ Pareto
○ Student's t
○ Triangular
○ Uniform
○ Weibull

Number of Trials: 1

Number of Successes: 1

Population Size: 1

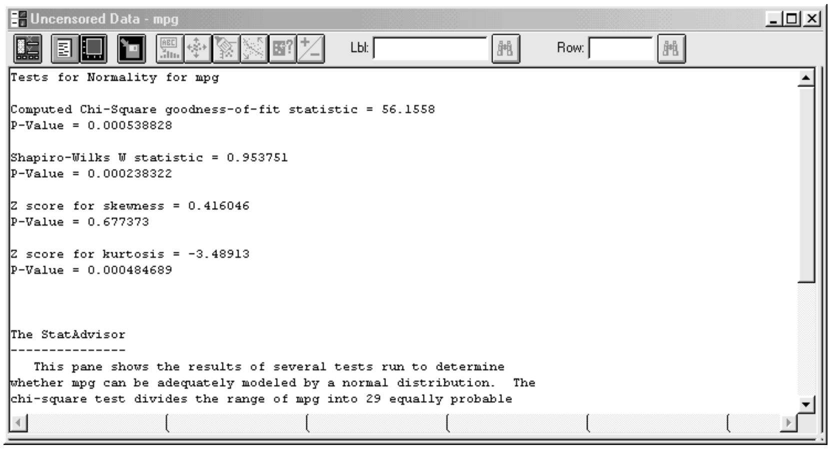
Figure 11-19. Probability Distributions Options Dialog Box

Tests for Normality

The Tests for Normality option calculates the results for several tests that determine if the data can be adequately modeled by the Normal distribution (see Figure 11-20). The results include the values and the *p*-values for these tests: Chi-Square Goodness-of-Fit, Shapiro-Wilk's W, *z*-Score for Skewness, and *z*-Score for Kurtosis.

- **Chi-Square Goodness-of-Fit**
  The Chi-Square Goodness-of-Fit statistic compares the frequency of the data (observed frequency) with a distribution you select (expected frequency). This is an updated version of the standard chi-square goodness-of-fit statistic (see Madansky, 1988).

- **Shapiro-Wilk's W-Statistic**
  The Shapiro-Wilk's W-Statistic tests the assumption that the data follow a Normal rather than another distribution (see Madansky, 1988).



```
Tests for Normality for mpg

Computed Chi-Square goodness-of-fit statistic = 56.1558
P-Value = 0.000538828

Shapiro-Wilks W statistic = 0.953751
P-Value = 0.000238322

Z score for skewness = 0.416046
P-Value = 0.677373

Z score for kurtosis = -3.48913
P-Value = 0.000484689



The StatAdvisor
---------------
   This pane shows the results of several tests run to determine
whether mpg can be adequately modeled by a normal distribution.  The
chi-square test divides the range of mpg into 29 equally probable
```

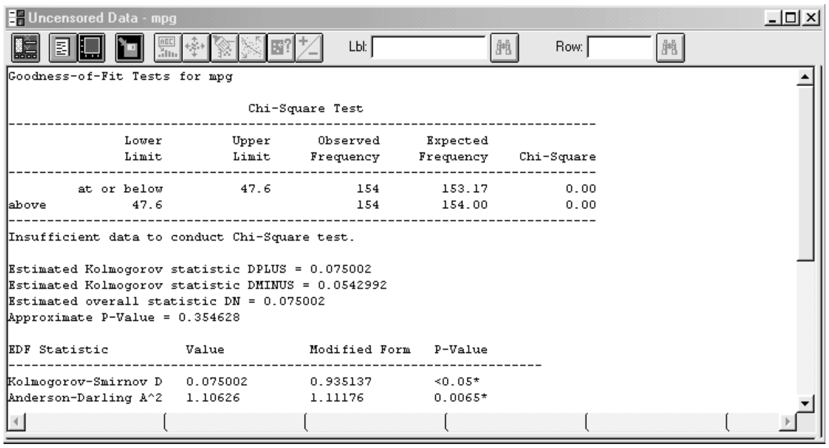
Figure 11-20. Tests for Normality

- ***z*-Score for Skewness**
  The *z*-Score for Skewness statistic determines if the data are symmetrically distributed; and, if so, how likely they are to be normally distributed. Unlike standardized skewness, the *z*-score for skewness is valid for small sample sizes (*n* is less than or equal to 8) (see D'Agostino and Stephens, 1986).

- ***z*-Score for Kurtosis**
  The *z*-Score for Kurtosis statistic determines the degree of peakedness in the distribution. Unlike standardized kurtosis, the *z*-score for kurtosis is valid for small sample sizes (*n* is less than or equal to 20) (see D'Agostino and Stephens, 1986).

Goodness-of-Fit Tests

The Goodness-of-Fit Tests option calculates results for tests of fit based on the Empirical Distribution Function (EDF). EDF is a step function that is calculated from the sample, which estimates the population distribution function. EDF statistics are used for testing the fit of the sample to the distribution (D'Agostino and Stephens, 1986). The tests are Chi-Square Goodness of Fit, Kolmogorov-Smirnov D, Kuiper V, Cramer-von Mises W^2, Watson U^2, and Anderson-Darling A^2 (see Figure 11-21).



```
Goodness-of-Fit Tests for mpg

                              Chi-Square Test
----------------------------------------------------------------------------
               Lower         Upper      Observed      Expected
               Limit         Limit     Frequency     Frequency    Chi-Square
----------------------------------------------------------------------------
         at or below          47.6           154        153.17          0.00
above           47.6                         154        154.00          0.00
----------------------------------------------------------------------------
Insufficient data to conduct Chi-Square test.

Estimated Kolmogorov statistic DPLUS = 0.075002
Estimated Kolmogorov statistic DMINUS = 0.0542992
Estimated overall statistic DN = 0.075002
Approximate P-Value = 0.354628

EDF Statistic          Value           Modified Form   P-Value
---------------------------------------------------------------------
Kolmogorov-Smirnov D   0.075002        0.935137        <0.05*
Anderson-Darling A^2   1.10626         1.11176         0.0065*
```

Figure 11-21. Goodness-of-Fit Tests

D'Agostino and Stephens (1986) provide the following explanation of these EDF statistics:

> The first two EDF statistics, D^+ and D^-, are respectively, the largest vertical difference when $F_n(x)$ is smaller than F(x), and the largest vertical difference when $F_n(x)$ is smaller than F(x). . . . The most well-known EDF statistic is D, introduced by Kolmogorov (1933).... A closely related statistic V, given by Kuiper (1960), is useful for observations on a circle. . . . A second and wide class of measures of discrepancy is given by the Cramer-von Mises family . . . when
> y(x) = 1, the statistic is the Cramer-von Mises statistic, now usually called W^2, and when
>
> $y(x) =[\{F(x)\}\{1 - F(x)\}]^{-1}$
>
> the statistic is the Anderson-Darling statistic, called A^2. A modification of W^2, also devised originally for the circle, is the Watson statistic U^2.

- **Chi-Square**
  The Chi-Square statistic divides the range of the data into nonoverlapping intervals and compares the number of observations in each class with the number of expected frequencies, based on the fitted distribution.

- **Kolmogorov-Smirnov D**
  The Kolmogorov-Smirnov D statistic calculates the maximum distance between the cumulative distribution of the data and the cumulative

distribution function of the fitted distribution. This calculation is a nonparametric method that tests the overall goodness of fit between the distribution of the data and the distribution you choose.

- **Kuiper V**
  The Kuiper V statistic improves the Kolmogorov-Smirnov test in the tails of a distribution. If there are only two class levels, it is a scaled value for an asymptotic distribution, and determines the probability of observing a larger test statistic.

- **Cramer-von Mises W^2**
  The Cramer-von Mises W^2 statistic determines if a one-dimensional data sample is compatible with being a random sample from a given distribution. It determines if two data samples are compatible with being random samples from the same unknown distribution. It is similar to the Kolmogorov-Smirnov test except it is more complex.

- **Watson U^2**
  The Watson U^2 statistic is a modification of the Cramer-von Mises statistic; it was originally devised for a circle.

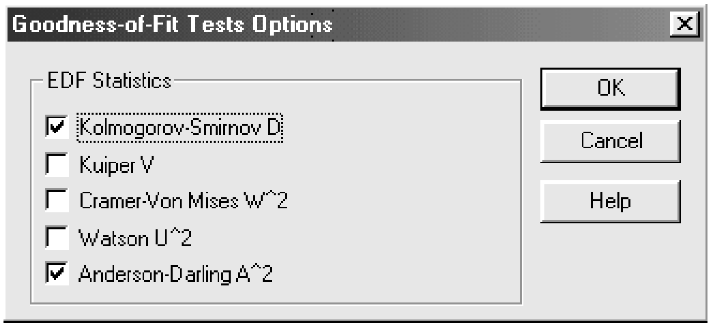
- **Anderson-Darling A^2**
  The Anderson-Darling statistic is a general test for complete datasets (without censored observations) that compares the fit of an observed cumulative distribution function with an expected cumulative distribution function.

Use the *Goodness-of-Fit Tests Options* dialog box to choose an EDF statistic (see Figure 11-22).

Tail Areas

The Tail Areas option calculates the area under the density curve at specified points (see Figure 11-23). The test uses the cumulative distribution function to calculate the probability that a random variable will fall below a given value. The test calculates tail areas for up to five critical values.

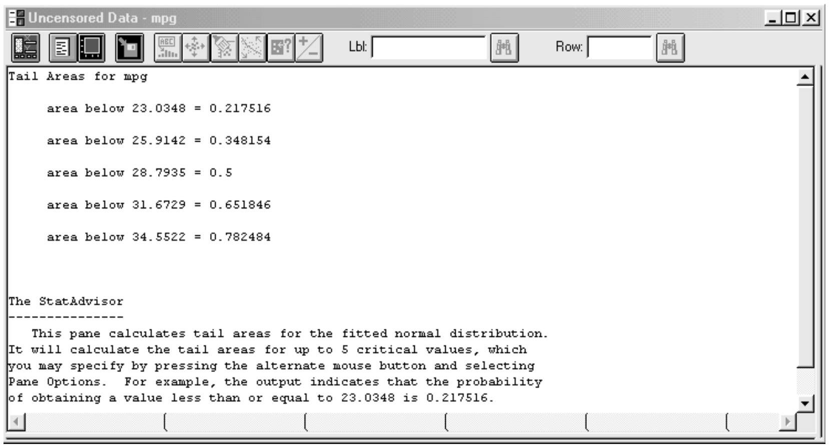
Use the *Tail Areas Options* dialog box to enter values that will be used to calculate the area under the curve.



EDF Statistics

☑ Kolmogorov-Smirnov D
□ Kuiper V
□ Cramer-Von Mises W^2
□ Watson U^2
☑ Anderson-Darling A^2

Figure 11-22. Goodness-of-Fit Tests Options Dialog Box



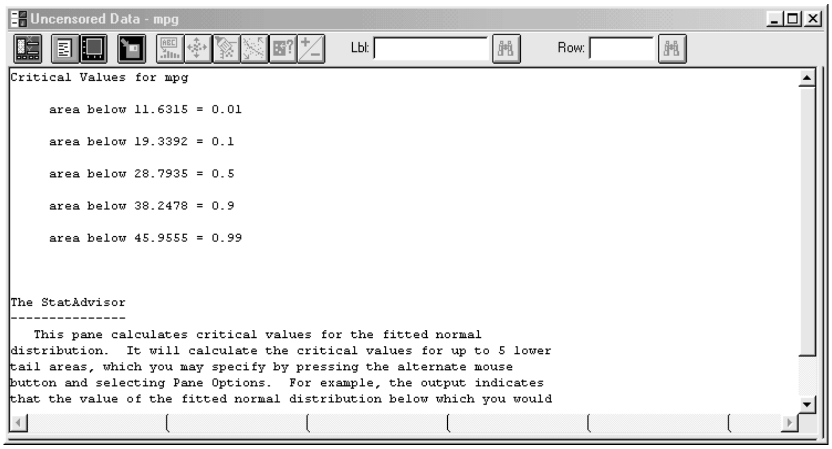
```
Tail Areas for mpg

     area below 23.0348 = 0.217516

     area below 25.9142 = 0.348154

     area below 28.7935 = 0.5

     area below 31.6729 = 0.651846

     area below 34.5522 = 0.782484



The StatAdvisor
---------------
   This pane calculates tail areas for the fitted normal distribution.
It will calculate the tail areas for up to 5 critical values, which
you may specify by pressing the alternate mouse button and selecting
Pane Options.  For example, the output indicates that the probability
of obtaining a value less than or equal to 23.0348 is 0.217516.
```

Figure 11-23. Tail Areas

Critical Values

The Critical Values option calculates critical values for the distribution (see Figure 11-24). Critical values are the smallest values for the area that fall under the distribution curve. The values for the distribution curve are those that are no less than the values you selected for the probability. The test calculates critical values for up to five lower tail areas.

Use the *Critical Values Options* dialog box to enter values for the probabilities that will be used to calculate the corresponding critical values.

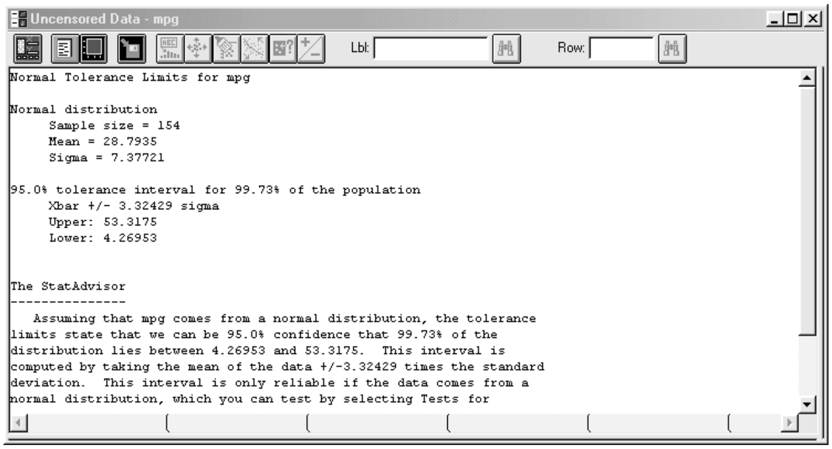


```
Critical Values for mpg

     area below 11.6315 = 0.01

     area below 19.3392 = 0.1

     area below 28.7935 = 0.5

     area below 38.2478 = 0.9

     area below 45.9555 = 0.99



The StatAdvisor
---------------
   This pane calculates critical values for the fitted normal
distribution.  It will calculate the critical values for up to 5 lower
tail areas, which you may specify by pressing the alternate mouse
button and selecting Pane Options.  For example, the output indicates
that the value of the fitted normal distribution below which you would
```

Figure 11-24. Critical Values

Normal Tolerance Limits

The Normal Tolerance Limits option creates normal tolerance limits for normally distributed data (see Figure 11-25). Tolerance limits are the values between which you can expect to find a specified proportion of the population. The results show the sample size you selected, the estimated parameters for the Normal distribution, and the Upper and Lower tolerance limits for the specified population proportion and confidence level.



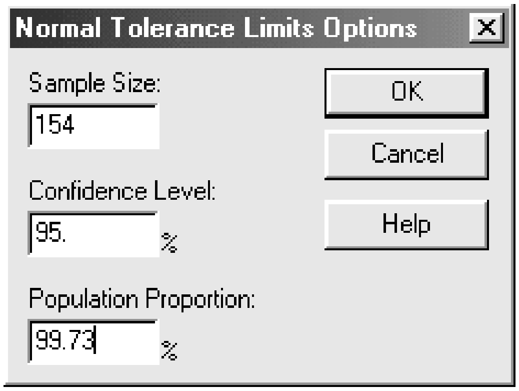
```
Normal Tolerance Limits for mpg

Normal distribution
     Sample size = 154
     Mean = 28.7935
     Sigma = 7.37721

95.0% tolerance interval for 99.73% of the population
     Xbar +/- 3.32429 sigma
     Upper: 53.3175
     Lower: 4.26953


The StatAdvisor
---------------
   Assuming that mpg comes from a normal distribution, the tolerance
limits state that we can be 95.0% confidence that 99.73% of the
distribution lies between 4.26953 and 53.3175.  This interval is
computed by taking the mean of the data +/-3.32429 times the standard
deviation.  This interval is only reliable if the data comes from a
normal distribution, which you can test by selecting Tests for
```

Figure 11-25. Normal Tolerance Limits

The program calculates the Upper and Lower tolerance limits by taking the mean of the data, plus or minus, respectively, *K* times the standard deviation, where *K* is a constant whose value is based on the sample size, the confidence level, and the population proportion specified for the test. The value of *K* is also shown in the results. The analysis uses the estimated parameters for a column of data to determine the tolerance limits for future samples taken from the same population.

If you believe your data do not follow a Normal distribution and you want nonparametric tolerance limits, use the Distribution-Free Tolerance Limits option.

Use the *Normal Tolerance Limits Options* dialog box to enter values for the size of the sample, the confidence level, and the population proportion (see Figure 11-26).



Normal Tolerance Limits Options

Sample Size: 154

Confidence Level: 95. %

Population Proportion: 99.73 %

Figure 11-26. Normal Tolerance Limits Options Dialog Box

Distribution-Free Tolerance Limits

The Distribution-Free Tolerance Limits option creates nonparametric limits, as well as the number of counts; the minimum, maximum, and median of the data; and the confidence level for the limits and population proportion (see Figure 11-27). The program calculates the interval using the smallest and largest value(s) from the dataset. Use nonparametric tolerance limits when you believe the data may not be from a Normal distribution.

Use the *Distribution-Free Tolerance Limits Options* dialog box to enter values for the confidence level, population proportion, and the interval depth (see Figure 11-28).

Graphical Options

Density Trace

The Density Trace option creates a plot, which is a smoothed histogram of the shape of the distribution, especially the variations in density over the range of the data (see Figure 11-29). The plot uses overlapping intervals and a weight function to smooth the densities, which result in a continuous line rather than a group of rectangles.

Use the *Density Trace Options* dialog box to indicate the method that will be used to estimate the density trace; to enter the degree of overlap that will be used to compute the density trace; and to enter the number of density values that will be calculated (see Figure 11-30).

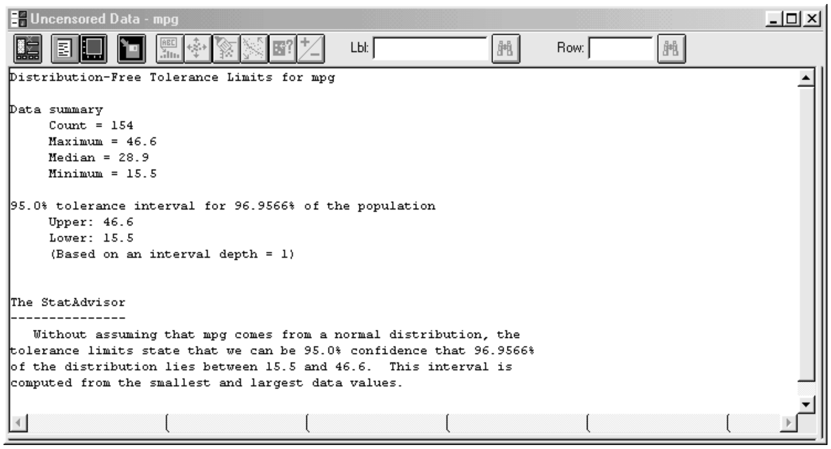
Symmetry Plot

The Symmetry Plot option creates a plot that helps assess the symmetry of the data. The program constructs the plot by first sorting the values from smallest to largest (see Figure 11-31). It then selects the values to the left and right of the median and plots them as points a respective distance from the median. It repeats this process for the pair of points that are second closest to the median, then third closest, and so on.

If the distribution is symmetric, the points will lie close to the diagonal line. If the distribution is positively skewed, the points will deviate above the line. If the distribution is negatively skewed, the points will deviate below the line

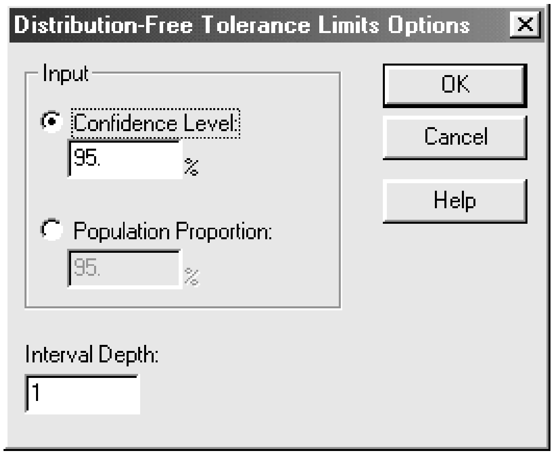
Frequency Histogram

The Frequency Histogram option creates a histogram that shows the distribution curve as an overlay (see Figure 11-32). The program divides the data into sets of nonoverlapping intervals and plots bars for each interval; the height of each bar is proportional to the number of observations that fall within that interval. Use the plot to determine if the data follow the distribution you selected.



```
Distribution-Free Tolerance Limits for mpg

Data summary
     Count = 154
     Maximum = 46.6
     Median = 28.9
     Minimum = 15.5

95.0% tolerance interval for 96.9566% of the population
     Upper: 46.6
     Lower: 15.5
     (Based on an interval depth = 1)


The StatAdvisor
---------------
   Without assuming that mpg comes from a normal distribution, the
tolerance limits state that we can be 95.0% confidence that 96.9566%
of the distribution lies between 15.5 and 46.6.  This interval is
computed from the smallest and largest data values.
```

Figure 11-27. Distribution-Free Tolerance Limits



Input

● Confidence Level: 95. %

○ Population Proportion: 95. %

Interval Depth: 1

Figure 11-28. Distribution-Free Tolerance Limits Options Dialog Box

Use the *Frequency Tabulation Options* dialog box to enter the number of classes into which the data will be grouped, and to enter values for the Lower and Upper limits. You can also indicate if the current scaling should be retained if you change the data in the Analysis dialog box (see Figure 11-33).

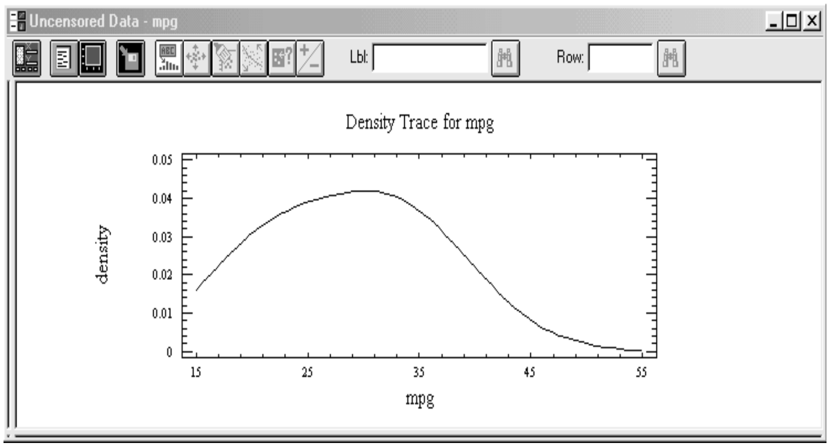


Figure 11-29. Density Trace

Note: Changes you make to this plot can affect the Goodness-of-Fit tabular option.

Quantile Plot

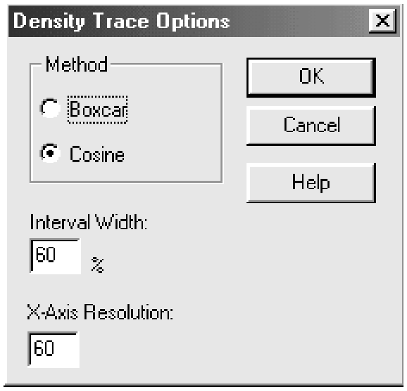
The Quantile Plot option creates a plot of the quantiles for the data (see Figure 11-34). The Y-axis represents the proportion of values that are below a particular value.

Quantile/Quantile Plot

The Quantile/Quantile Plot option creates a plot of the sorted values against the corresponding quartiles of the fitted distribution so you can compare the cumulative distributions for the two samples (see Figure 11-35). If the distributions are a good fit, the points fall along the line.

Distribution Functions 1 and 2

The Distribution Functions 1 and 2 options each create one of several different types of distribution function plots (see Figures 11-36 and 11-37).



Density Trace Options

Method

○ Boxcar

● Cosine

Interval Width: 60 %

X-Axis Resolution: 60

Figure 11-30. Density Trace Options Dialog Box

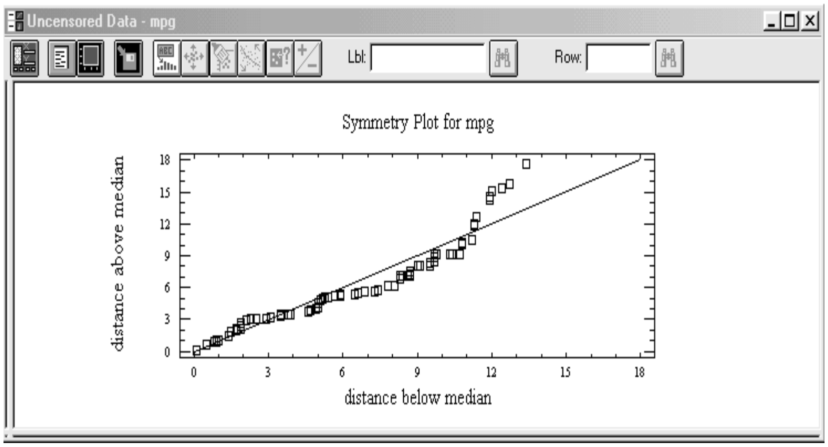


Figure 11-31. Symmetry Plot

The type of functions you can display are: Density Function, Cumulative Distribution Function, Survivor Function, Log Survivor Function, and Hazard Function. The default for Distribution Function 1 is the Density Function plot. The default for Distribution Function 2 is the Cumulative Distribution Function plot.

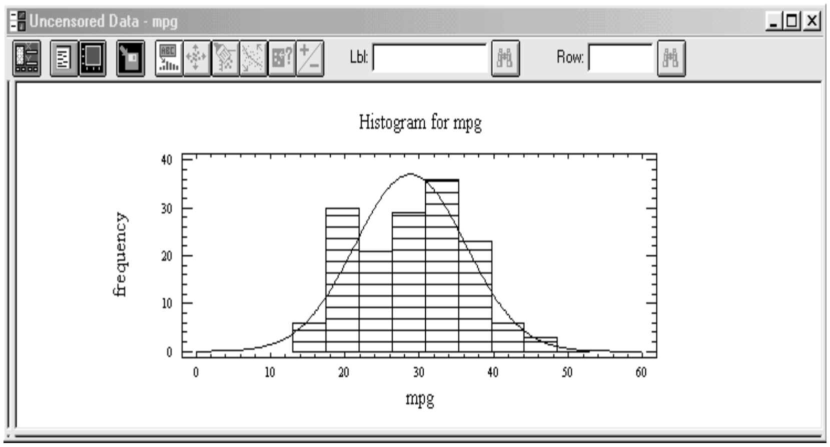
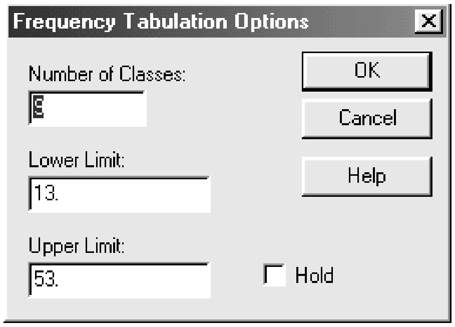


Figure 11-32. Frequency Histogram


Frequency Tabulation Options

Number of Classes: 8

Lower Limit: 13.

Upper Limit: 53.

□ Hold

Figure 11-33. Frequency Tabulation Options Dialog Box

Use the *Distribution Functions Options* dialog box to indicate the type of function distribution you want to plot, and to enter the number of density values that will be calculated (see Figure 11-38).

References

D'Agostino, R. B. and Stephens, M. A. 1986. *Goodness-of-Fit Techniques*. New York: Marcel Dekker, Inc.

Hastings, N. A. J. and Peacock, J. B. 1975. *Statistical Distributions*. London: Butterworth and Co.

Law, A. M. and Kelton, W. D. 1982. *Simulation Modeling and Analysis*. New York: McGraw-Hill.

Madansky, A. 1988. *Prescriptions for Working Statisticians*. New York: Springer-Verlag.

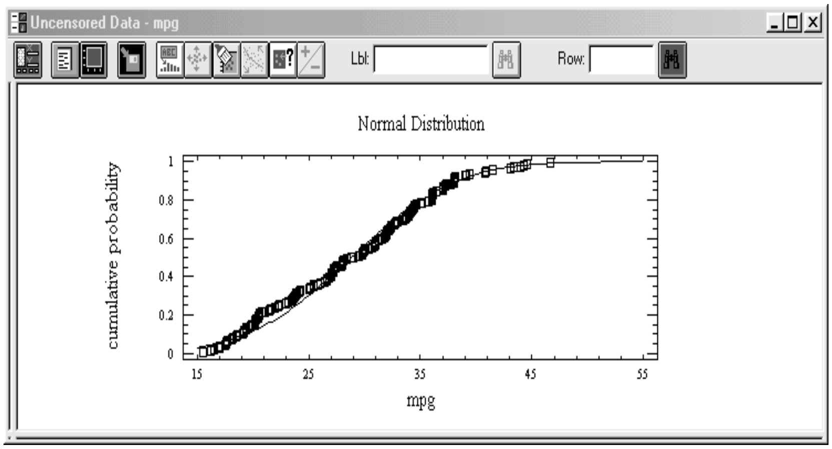


Figure 11-34. Quantile Plot

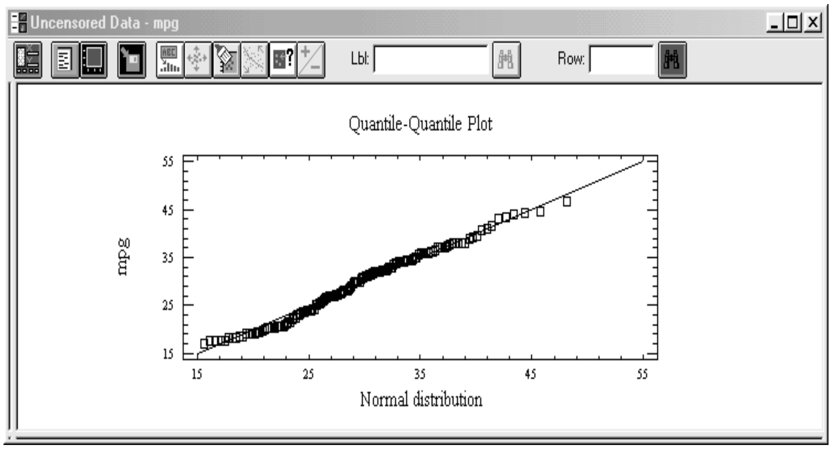


Figure 11-35. Quantile/Quantile Plot

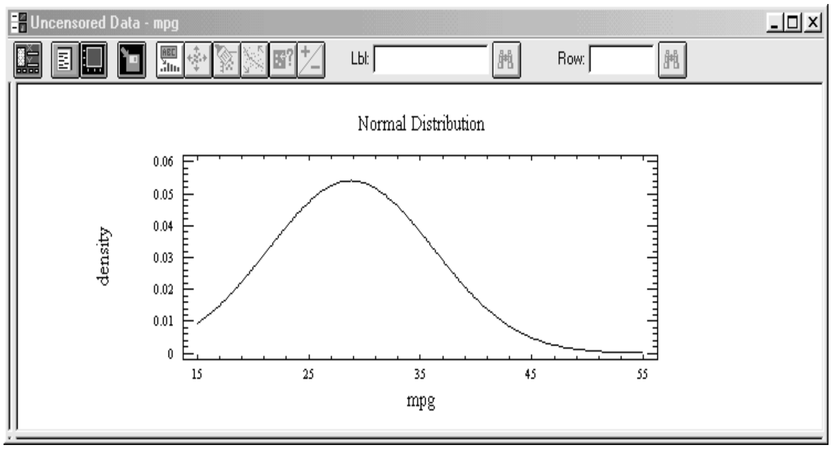


Figure 11-36.   Distribution Function 1 Plot

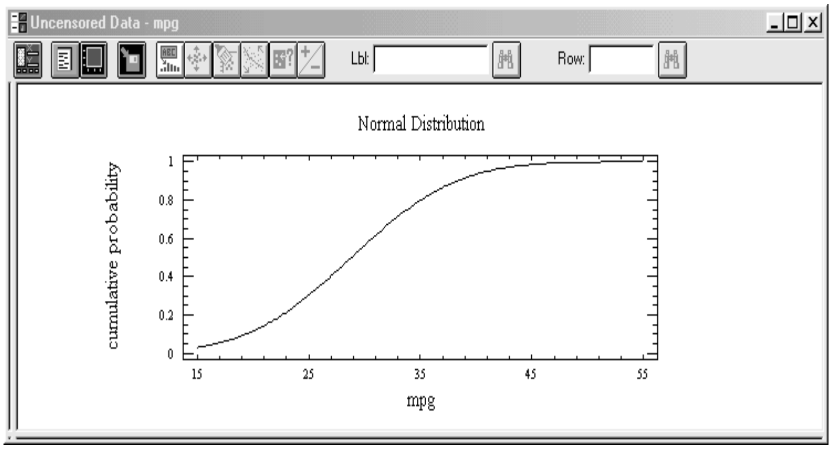
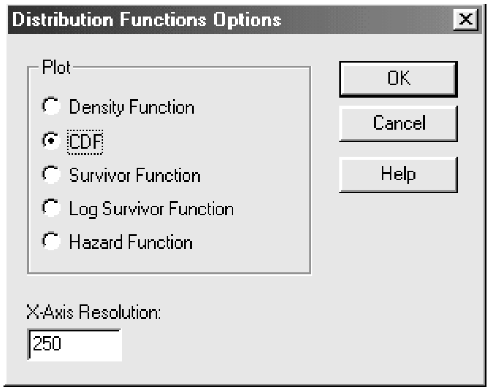


Figure 11-37.   Distribution Function 2 Plot



Distribution Functions Options

Plot

○ Density Function
● CDF
○ Survivor Function
○ Log Survivor Function
○ Hazard Function

X-Axis Resolution: 250

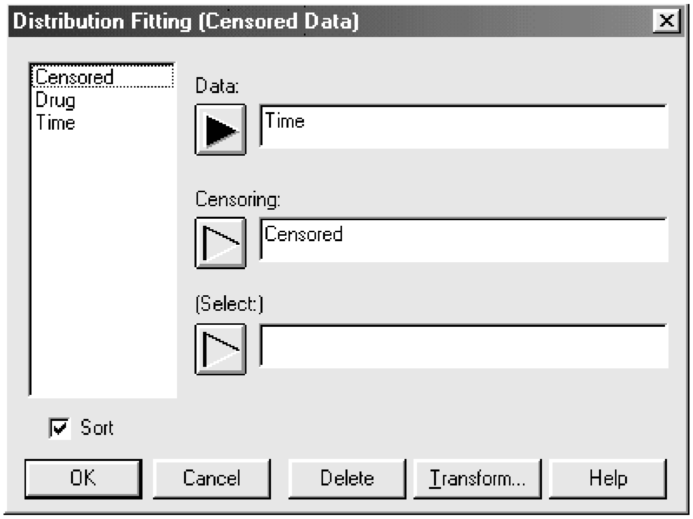
Figure 11-38. Distribution Functions Options Dialog Box

Fitting Distributions Using Censored Data

Censored data are data or samples that are incomplete in some way, such as when certain values are unknown or ignored. For example, in a study of the 1993 college graduation rate of individuals born in 1970, some of those born then may not have completed college by 1993; however, they could easily have done so after the end of the study. The data in the study would be censored because the number of individuals from the 1970 group who finished their degrees after 1993 would be unknown.

The Censored Data Analysis fits censored data to one of the 24 probability distribution functions available in STATGRAPHICS *Plus*. The analysis fits a given distribution to a set of data, then calculates the estimated parameters for the distribution.

To access this analysis, from the menus, choose: DESCRIBE... DISTRIBUTIONS... DISTRIBUTION FITTING (CENSORED DATA)... (see Figure 11-39).



Distribution Fitting (Censored Data)

Data: Time

Censoring: Censored

(Select:)

☑ Sort

Figure 11-39. Distribution Fitting (Censored Data) Analysis Dialog Box

Tabular Options

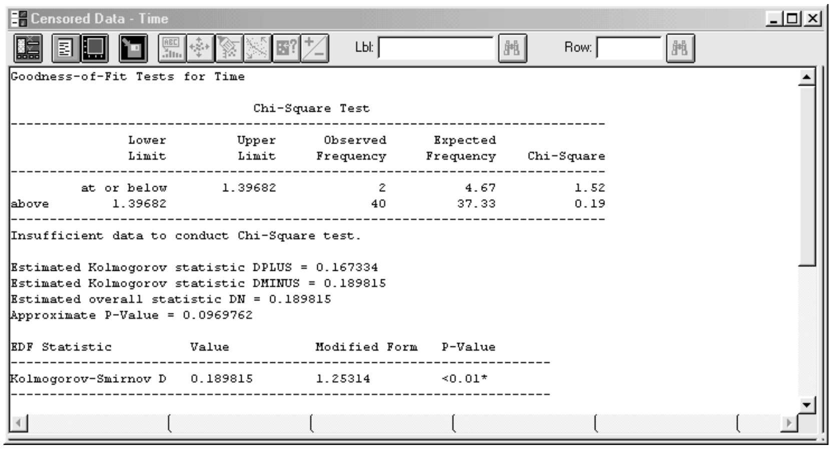
Analysis Summary

The Analysis Summary option creates a summary of the results, which includes the number of values and their ranges, as well as the number of left- and right-censored (positive/negative) observations. It then displays the results of the fitted distribution; for example, it shows the values for the mean and standard deviation for a Lognormal distribution if that is the distribution you are using.

Use the *Probability Distributions Options* dialog box to choose a different distribution, to enter the number of trials, and to enter the number of successes (see Figure 11-19 for an example of this dialog box). Each of the 24 distributions has its own Options dialog box. *See Online Help for a description of these dialog boxes and their use; for example, see Bernoulli Options Dialog Box, Binomial Options Dialog Box ... Uniform Options Dialog Box, Weibull Options Dialog Box.*

Goodness-of-Fit Tests

The Goodness-of-Fit Tests option calculates results for tests of fit based on the Empirical Distribution Function (EDF). EDF is a step function that is calculated from the sample, which estimates the population distribution function. EDF statistics are used for testing the fit of the sample to the distribution (D'Agostino and Stephens, 1986). The two tests are Kolmogorov-Smirnov D and Kuiper V (see Figure 11-40).



```
Goodness-of-Fit Tests for Time

                               Chi-Square Test
----------------------------------------------------------------------------
               Lower         Upper      Observed      Expected
               Limit         Limit     Frequency     Frequency    Chi-Square
----------------------------------------------------------------------------
         at or below       1.39682             2          4.67          1.52
above        1.39682                          40         37.33          0.19
----------------------------------------------------------------------------
Insufficient data to conduct Chi-Square test.

Estimated Kolmogorov statistic DPLUS = 0.167334
Estimated Kolmogorov statistic DMINUS = 0.189815
Estimated overall statistic DN = 0.189815
Approximate P-Value = 0.0969762

EDF Statistic          Value           Modified Form   P-Value
---------------------------------------------------------------------
Kolmogorov-Smirnov D   0.189815        1.25314         <0.01*
---------------------------------------------------------------------
```

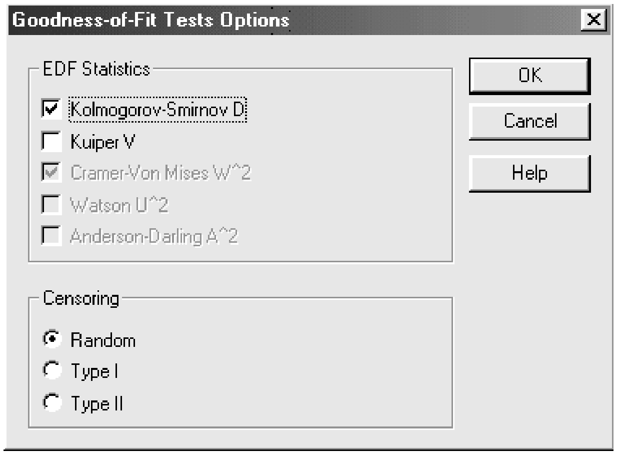
Figure 11-40.   Goodness-of-Fit Tests

- **Kolmogorov-Smirnov D**
  The Kolmogorov-Smirnov D test calculates the maximum distance between the cumulative distribution of the data and the cumulative distribution function of the fitted distribution. This calculation is a nonparametric method used to test the overall goodness of fit between the distribution of the data and the distribution you choose. If the *p* value is less than .05 (for a 95 percent confidence level), the data do not fit the distribution. The Kolmogorov-Smirnov test is available when you are fitting data to continuous distributions.

- **Kuiper V**
  The Kuiper V statistic improves the Kolmogorov-Smirnov test in the tails of a distribution. If there are only two class levels, it is a scaled value for

an asymptotic distribution, and determines the probability of observing a larger test statistic.

Use the *Goodness-of-Fit Tests Options* dialog box to choose the EDF statistic and the censoring method (see Figure 11-41).



Goodness-of-Fit Tests Options

EDF Statistics

- ☑ Kolmogorov-Smirnov D
- □ Kuiper V
- ☑ Cramer-Von Mises W^2
- □ Watson U^2
- □ Anderson-Darling A^2

Censoring

- ● Random
- ○ Type I
- ○ Type II

Figure 11-41. Goodness-of-Fit Tests Options Dialog Box

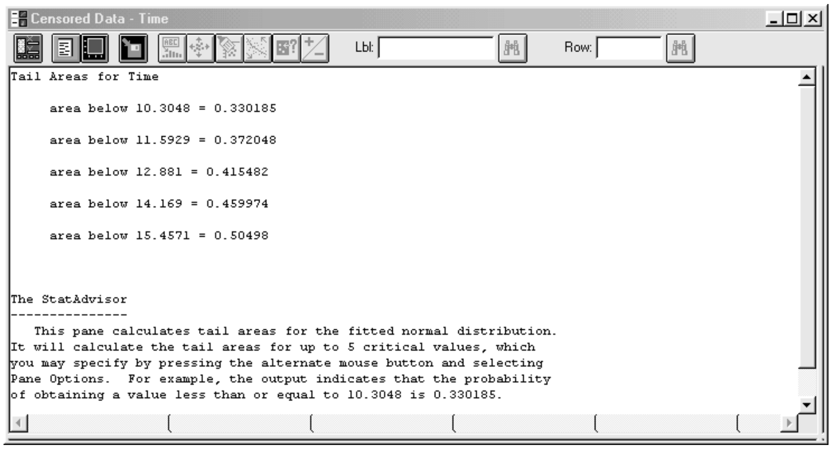
Tail Areas

The Tail Areas option calculates the area under the density curve at specified points. The test uses the cumulative distribution function to calculate the probability that a random variable will fall below a given value. The test calculates tail areas for up to five critical values (see Figure 11-42).

Use the *Tail Areas Options* dialog box to enter values that will be used to calculate the areas under the curve.

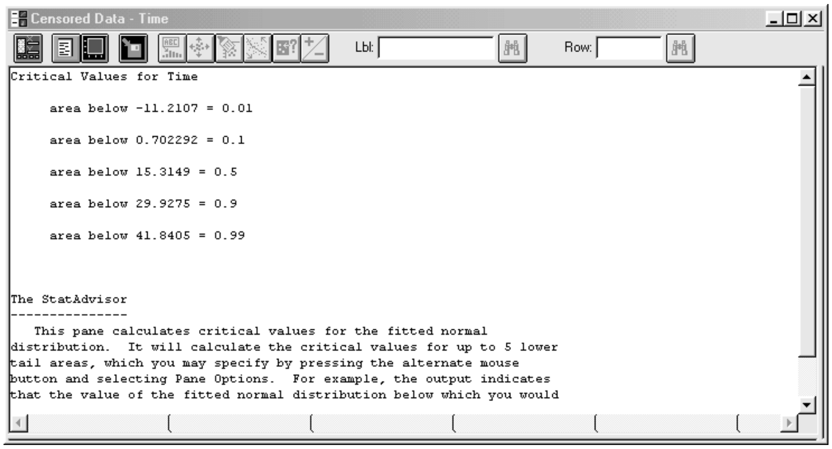
Critical Values

The Critical Values option calculates critical values for the distribution. Critical values are the smallest values for the area that fall under the distribution curve (see Figure 11-43). The values for the distribution curve are those that are no less than the values you selected for the probability. The test calculates critical values for up to five lower tail areas.



```
Tail Areas for Time

     area below 10.3048 = 0.330185

     area below 11.5929 = 0.372048

     area below 12.881 = 0.415482

     area below 14.169 = 0.459974

     area below 15.4571 = 0.50498



The StatAdvisor
---------------
   This pane calculates tail areas for the fitted normal distribution.
It will calculate the tail areas for up to 5 critical values, which
you may specify by pressing the alternate mouse button and selecting
Pane Options.  For example, the output indicates that the probability
of obtaining a value less than or equal to 10.3048 is 0.330185.
```

Figure 11-42.   Tail Areas



```
Critical Values for Time

     area below -11.2107 = 0.01

     area below 0.702292 = 0.1

     area below 15.3149 = 0.5

     area below 29.9275 = 0.9

     area below 41.8405 = 0.99



The StatAdvisor
---------------
   This pane calculates critical values for the fitted normal
distribution.  It will calculate the critical values for up to 5 lower
tail areas, which you may specify by pressing the alternate mouse
button and selecting Pane Options.  For example, the output indicates
that the value of the fitted normal distribution below which you would
```

Figure 11-43. Critical Values

Use the *Critical Values Options* dialog box to enter values for the probabilities that will be used to calculate the corresponding critical values.

Graphical Options

Frequency Histogram

The Frequency Histogram option creates a histogram with the distribution curve shown as an overlay (see Figure 11-44). The program divides the data into sets of nonoverlapping intervals and plots bars for each interval; the height of each bar is proportional to the number of observations that fall within that interval. Use the plot to determine if the data follow the distribution you selected.

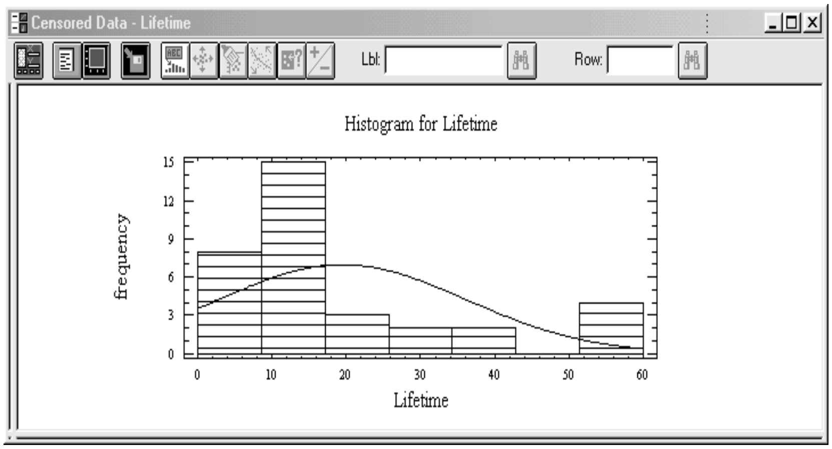


Figure 11-44. Frequency Histogram

Use the *Frequency Tabulation Options* dialog box to enter the number of classes into which the data will be grouped, and to enter values for the Lower and Upper limits. You can also indicate if the current scaling should be retained if you change the data in the Analysis dialog box (see Figure 11-33 for an example of this dialog box).

Quantile Plot

The Quantile Plot option creates a plot of the quantiles for the data (see Figure 11-45). The Y-axis represents the proportion of values that are below a particular value.

Quantile/Quantile Plot

The Quantile/Quantile Plot option creates a plot of the sorted values against the corresponding quartiles of the fitted distribution so you can compare the cumulative distributions for the two samples (see Figure 11-46). If the distributions are a good fit, the points fall along the line.

Distribution Functions 1 and 2

The Distribution Functions 1 and 2 options each create one of several different types of distribution function plots (see Figures 11-47 and 11-48). The type of functions you can display are: Density Function, Cumulative Distribution Function, Survivor Function, Log Survivor Function, and Hazard Function. The default for Distribution Function 1 is the Density Function plot. The default for Distribution Function 2 is the Cumulative Distribution Function plot.

Use the *Distribution Functions Options* dialog box to indicate the type of function distribution you want to plot and to enter the number of density values that will be calculated (see Figure 11-38 for an example of this dialog box).

References

Ansell, J. I. and Phillips, M. J. 1994. *Practical Methods for Reliability Data Analysis*. London: Oxford Science Publications, Clarendon Press.

Collett, D. 1996. *Modelling Survival Data in Medical Research*. London: Chapman & Hall.

Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons.

D'Agostino, R. B. and Stephens, M. A. 1986. *Goodness-of-Fit Techniques*. New York: Marcel Dekker, Inc.

Nelson, W. B. 1982. *Applied Life Data*. New York: Wiley.

Tobias, P. A. and Trindade, D. C. 1995. *Applied Reliability, second edition. London: Chapman & Hall.*

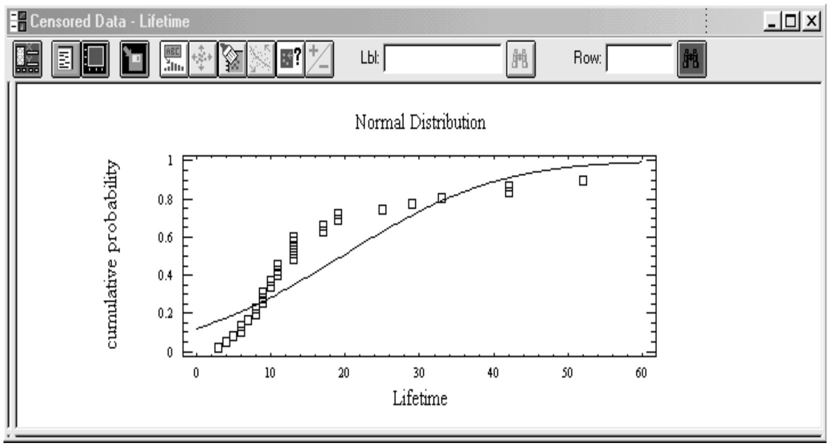


Figure 11-45. Quantile Plot

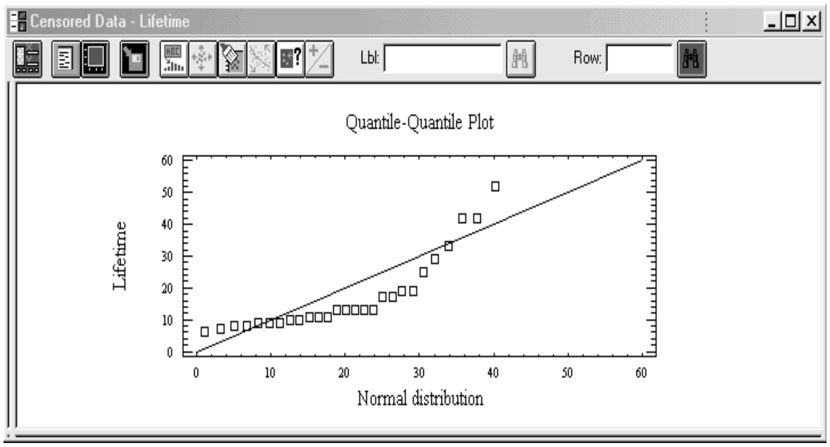


Figure 11-46. Quantile/Quantile Plot

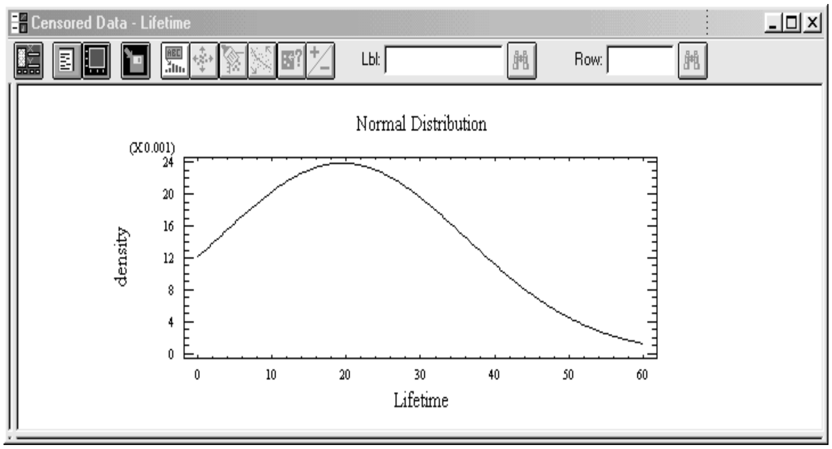


Figure 11-47. Distribution Function 1 Plot

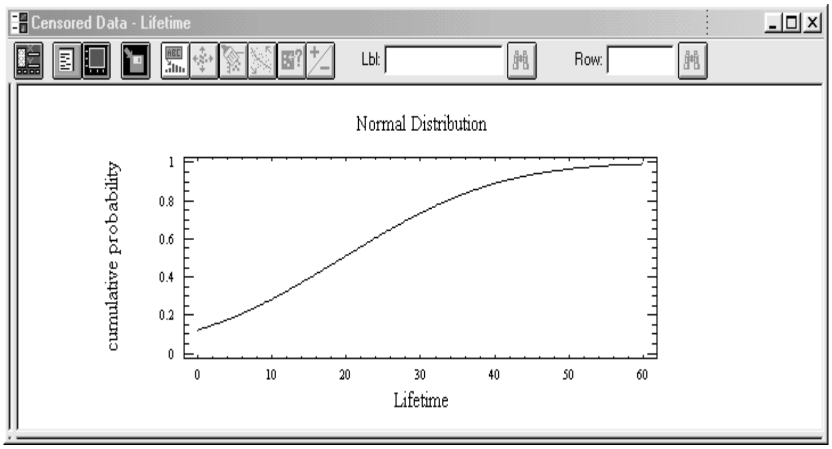


Figure 11-48. Distribution Function 2 Plot

12 Performing Analyses Using Life Data

One of the oldest and most widely used methods for portraying lifetime data — the survival experience of a group of individuals — are life tables, which are usually assumed to be a random sample from some population. Life tables provide estimates of survival probabilities for a population.

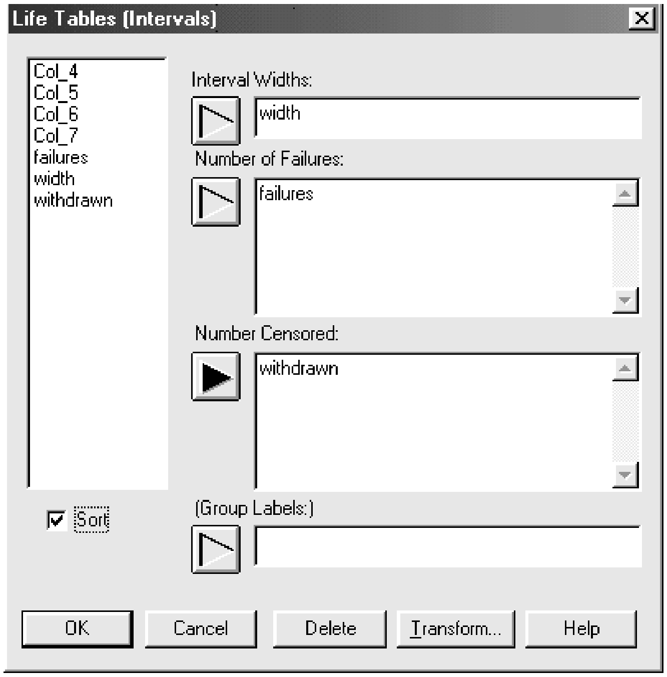
The life tables portion of STATGRAPHICS *Plus* consists of four analyses: life tables for both intervals and times, as well as the Weibull and Arrhenius Plots analyses. The Life Tables (Intervals) Analysis uses uncensored data to create life tables based on counts of failures in intervals, while the Life Tables (Times) Analysis uses censored data to create life tables from a series of failure times. The Weibull Analysis is a flexible distribution-modeling method that can be successfully applied to many product-failure mechanisms to create a wide variety of possible failure-rate curves (Tobias and Trindade, 1995). Under normal conditions, there are often situations where it is not possible to obtain sufficient amounts of data for fitting life distributions. The Arrhenius Plots Analysis is used in these instances to increase stress levels beyond normal operating conditions. The analysis then fails an adequate number of items, letting you collect the data you need, then attempts to fit the distributions.

Using the Life Tables (Intervals) Analysis

Sample items lost from a study are called *censored data*, which means you only know that the sample items survive until they leave the study. For example, imagine studying a sample of 100 automobile tires to determine their average expected tread life. During the study, some tires are lost because they are pierced and flattened; therefore, they can no longer be used for the duration of the study. Or, suppose you are studying 100 individuals to determine the effects of a pain-relieving medication. During the study some portion of the sample group moves to other cities and you lose contact; they can no longer be used for the duration of the study.

The Life Tables (Intervals) Analysis creates a life table from the counts of failures in a set of intervals. Using a random sample of censored data, the program creates the life table, which estimates the survival probabilities for a population over time; that is, it calculates the probability that items in the sample group will survive at least as long as the beginning of an interval.

To access the analysis, from the menus, choose: DESCRIBE... LIFE DATA... LIFE TABLES (INTERVALS)... (see Figure 12-1).



Col_4
Col_5
Col_6
Col_7
failures
width
withdrawn

☑ Sort

Interval Widths: width

Number of Failures: failures

Number Censored: withdrawn

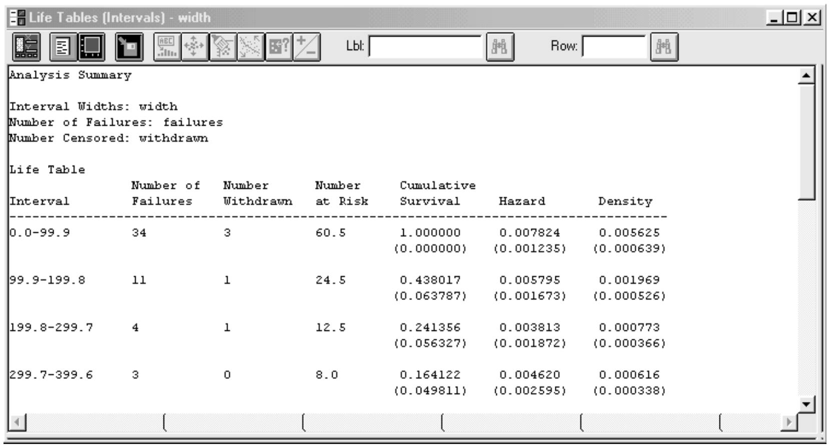
(Group Labels:)

Figure 12-1. The Life Tables (Intervals) Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option compares the groups in the study and creates a table of results (see Figure 12-2). The table shows the estimated survival probabilities within the number of intervals for the total number of items in the study. The results show the number of items at risk at the start of each interval, the number that failed before the end of the interval, and the number that were withdrawn (censored) during the interval.



```
Analysis Summary

Interval Widths: width
Number of Failures: failures
Number Censored: withdrawn

Life Table
               Number of   Number      Number     Cumulative
Interval       Failures    Withdrawn   at Risk    Survival      Hazard       Density
------------------------------------------------------------------------------------
0.0-99.9       34          3           60.5        1.000000     0.007824     0.005625
                                                  (0.000000)   (0.001235)   (0.000639)

99.9-199.8     11          1           24.5        0.438017     0.005795     0.001969
                                                  (0.063787)   (0.001673)   (0.000526)

199.8-299.7    4           1           12.5        0.241356     0.003813     0.000773
                                                  (0.056327)   (0.001872)   (0.000366)

299.7-399.6    3           0           8.0         0.164122     0.004620     0.000616
                                                  (0.049811)   (0.002595)   (0.000338)
```

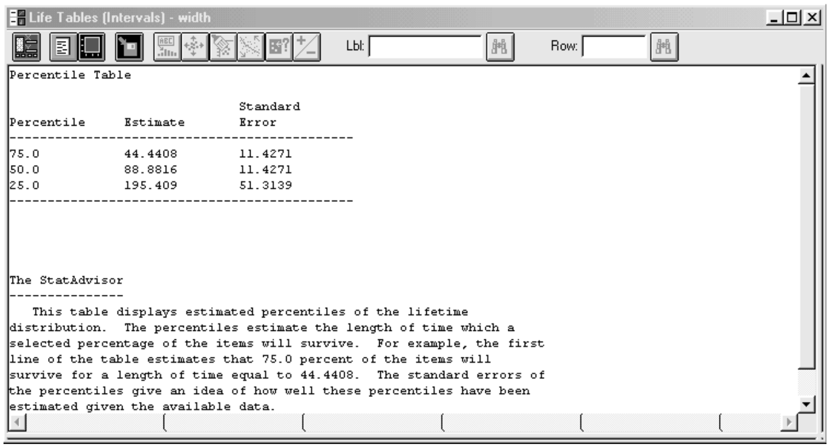
Figure 12-2. Analysis Summary

The Cumulative Survival column shows the estimated probability that an item will survive as least as long as the beginning of the interval. The Hazard column shows the estimated hazard function (instantaneous failure rate) over each interval. The Density column provides an estimate of the density function for the corresponding lifetime distribution. Standard errors are shown in parentheses for each of the three functions.

Percentiles

The Percentiles option creates a table of percentages for the lifetime distribution. The percentiles provide an estimation of the length of time that a selected percentage of the items will survive (see Figure 12-3). The standard errors of the percentiles show the accuracy of the estimates for the percentiles, given the available data.

Use the *Percentiles Options* dialog box to enter values for other percentiles. If you do not want to calculate percentiles, enter 0.

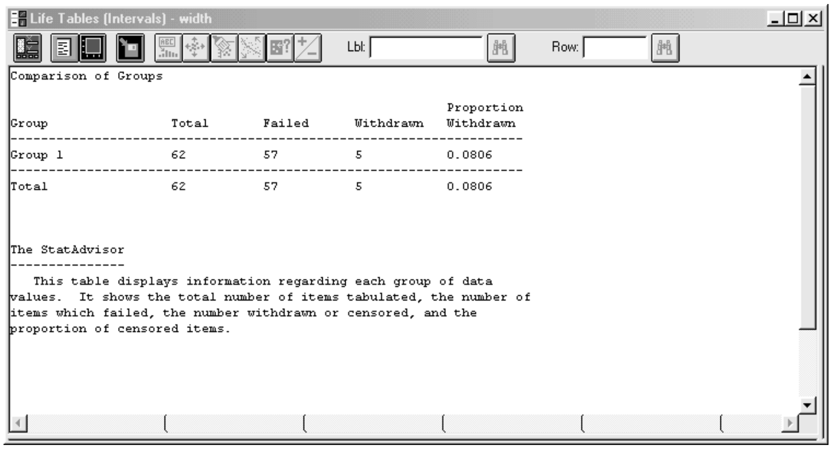


```
Percentile Table

                              Standard
Percentile     Estimate       Error
---------------------------------------------
75.0           44.4408        11.4271
50.0           88.8816        11.4271
25.0           195.409        51.3139
---------------------------------------------



The StatAdvisor
---------------
   This table displays estimated percentiles of the lifetime
distribution.  The percentiles estimate the length of time which a
selected percentage of the items will survive.  For example, the first
line of the table estimates that 75.0 percent of the items will
survive for a length of time equal to 44.4408.  The standard errors of
the percentiles give an idea of how well these percentiles have been
estimated given the available data.
```

Figure 12-3. Percentiles

Group Comparisons

The Group Comparisons option creates a table that contains information about each group of values in the data (see Figure 12-4). The table shows the total number of items that failed, the number that were withdrawn or censored, and the proportion of censored items.



```
Comparison of Groups

                                                        Proportion
Group                 Total       Failed      Withdrawn   Withdrawn
-------------------------------------------------------------------
Group 1               62          57          5           0.0806
-------------------------------------------------------------------
Total                 62          57          5           0.0806



The StatAdvisor
---------------
   This table displays information regarding each group of data
values.  It shows the total number of items tabulated, the number of
items which failed, the number withdrawn or censored, and the
proportion of censored items.
```

Figure 12-4. Group Comparisons

Graphical Options

Survival Function

The Survival Function option creates a plot of the estimated survival function, which is the probability that an item will not fail before a length of time you specify (see Figure 12-5).

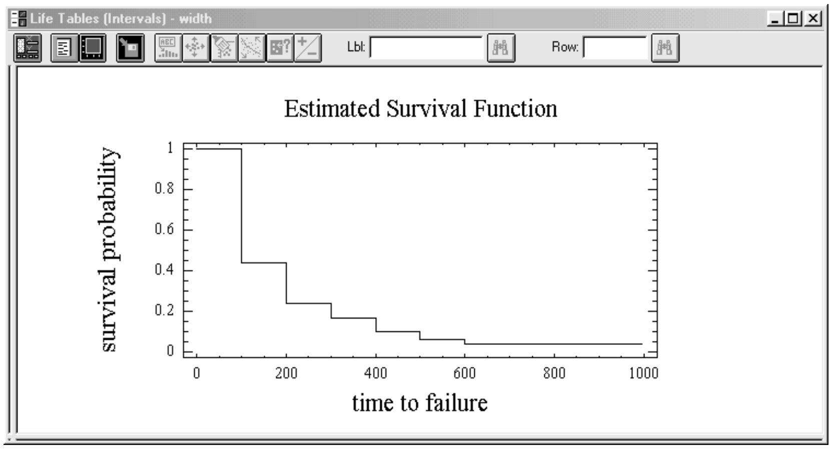


Figure 12-5. Survival Function

Log Survival Function

The Log Survival Function option creates a plot of the estimated log survival function, which is the probability that an item will not fail before a length of time you specify (see Figure 12-6).

Cumulative Hazard Function

The Cumulative Hazard Function option creates a plot of the estimated cumulative hazard function, which is helpful for suggesting possible parametric models for failure data (see Figure 12-7).

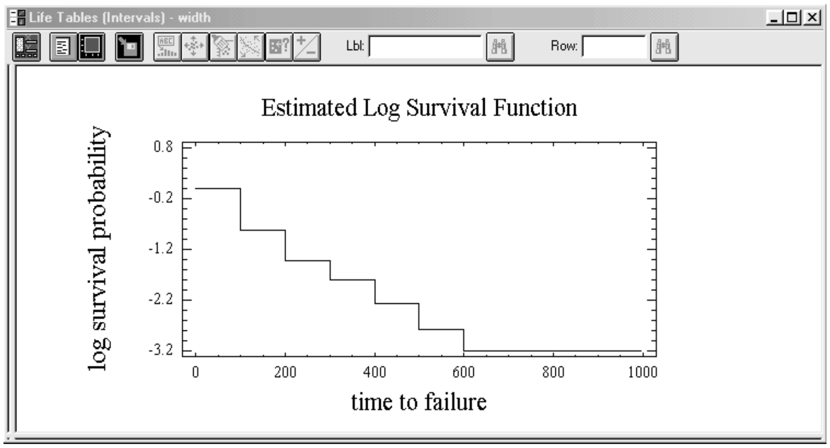


Figure 12-6. Log Survival Function

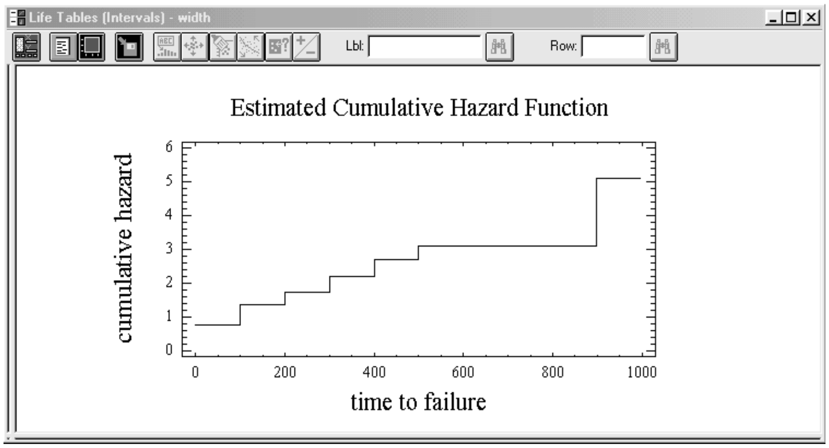

Life Tables (Intervals) - width
Lbl:
Row:
Estimated Cumulative Hazard Function
cumulative hazard
6
5
4
3
2
1
0
0
200
400
600
800
1000
time to failure

Figure 12-7. Cumulative Hazard Function

Death Density Function

The Death Density Function option creates a plot of the estimated death density function, which corresponds to the probability distribution of the item's lifetime (see Figure 12-8).

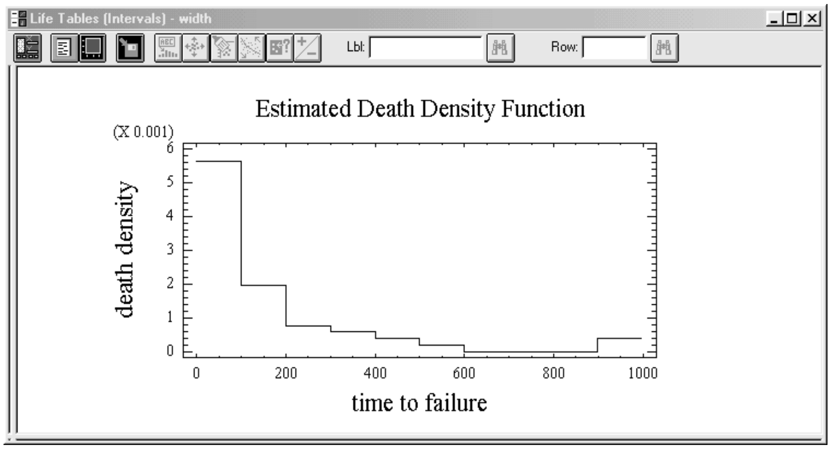


Figure 12-8. Death Density Function

Use the *Life Table Plots Options* dialog box to indicate if you want to include confidence limits on the plot and, if so, to enter a number for the confidence level that will be used to calculate the confidence intervals.

Hazard Function

The Hazard Function option creates a plot of the estimated hazard function, which shows the failure rate conditional to an item surviving to a point in time you specify (see Figure 12-9).

Use the *Life Table Plots Options* dialog box to indicate if you want to include confidence limits on the plot and, if so, to enter a number for the confidence level that will be used to calculate the confidence intervals.

References

Berkson, B. J. and Gage, R. P. 1950. "Calculations of Survival Rates for Cancer," *Proc. Staff Meet.* Mayo Clinic, **25**:270-286.

Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data.* New York: John Wiley & Sons.

Nelson, W. B. 1982. *Applied Life Data*. New York: Wiley.

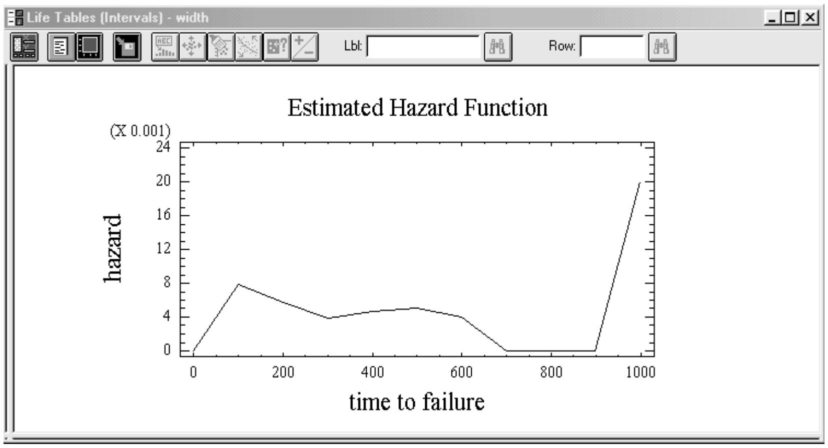


Figure 12-9.   Hazard Function

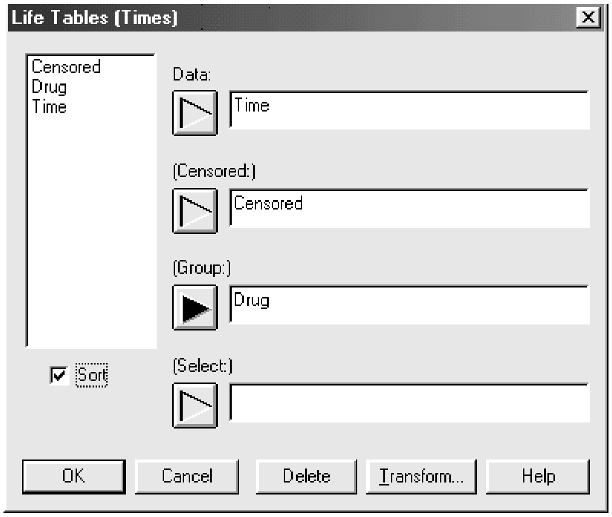
Tobias, P. and Trindade, D. 1986. *Applied Reliability*, second edition. London: Chapman & Hall.

Using the Life Tables (Times) Analysis

The Life Tables (Times) Analysis creates a life table from the failure times of items, which provides a nonparametric estimate of an empirical survival function.

The analysis is helpful when you must obtain survival probabilities for multicensored data. It is also particularly useful in reliability studies when there is need to estimate the lifetime of products and when, for various reasons, items must be withdrawn from a test. Another use is for studying and comparing clinical trials of survival rates for patients who are undergoing different treatments for the same medical condition.

To access the analysis, from the menus, choose: DESCRIBE... LIFE DATA... LIFE TABLES (TIMES)... (see Figure 12-10).



Life Tables (Times)

Data: Time

(Censored:) Censored

(Group:) Drug

(Select:)

☑ Sort

Figure 12-10. The Life Tables (Times) Analysis Dialog Box

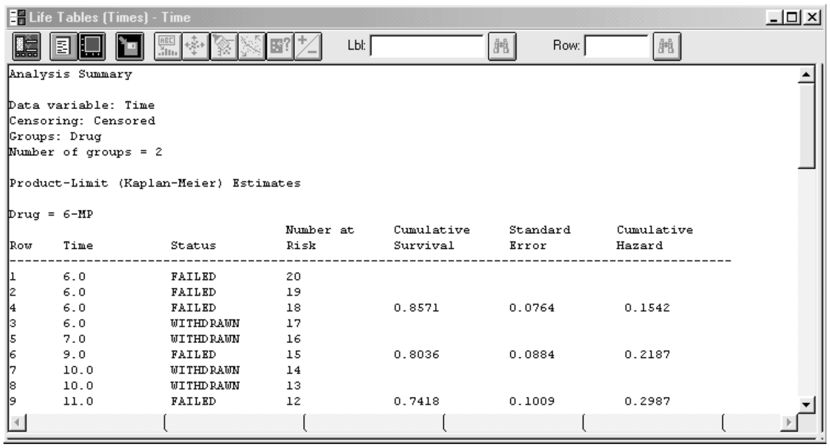
Tabular Options

Analysis Summary

The Analysis Summary option creates a table of the estimated survival probabilities (see Figure 12-11). Each row of the table represents a single value shown in increasing order. If the value represents a failure or death, the Status column reads FAILED. If the value represents a censored observation, the Status column reads WITHDRAWN.

The Time column displays the time an item failed or was withdrawn from the study, while the Number at Risk column displays the number of items that did not fail, were not withdrawn, or were not censored before the observed time.

The Cumulative Survival values represent the probability that an item will survive to a particular point in time. The values in the Standard Error and Cumulative Hazard columns represent, respectively, the standard deviation of the product-limit estimates, and one minus the cumulative survival probability.

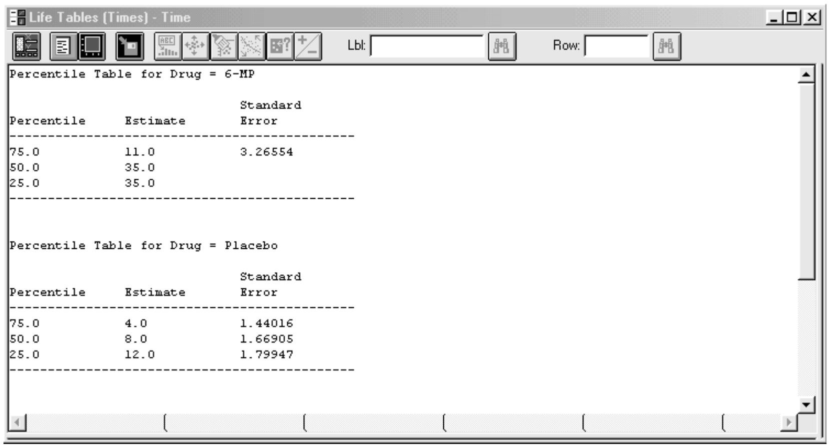


```
Analysis Summary

Data variable: Time
Censoring: Censored
Groups: Drug
Number of groups = 2

Product-Limit (Kaplan-Meier) Estimates

Drug = 6-MP
                                    Number at     Cumulative     Standard      Cumulative
Row    Time          Status         Risk          Survival       Error         Hazard
-------------------------------------------------------------------------------------------
1      6.0           FAILED         20
2      6.0           FAILED         19
4      6.0           FAILED         18            0.8571         0.0764         0.1542
3      6.0           WITHDRAWN      17
5      7.0           WITHDRAWN      16
6      9.0           FAILED         15            0.8036         0.0884         0.2187
7      10.0          WITHDRAWN      14
8      10.0          WITHDRAWN      13
9      11.0          FAILED         12            0.7418         0.1009         0.2987
```

Figure 12-11. Analysis Summary

Percentiles

The Percentiles option creates a table of the estimated percentiles for the lifetime distribution. The percentiles estimate the length of time that a selected percentage of the items will survive (see Figure 12-12). The standard errors of the percentiles provide an estimate of the accuracy for the percentiles, given the available data.



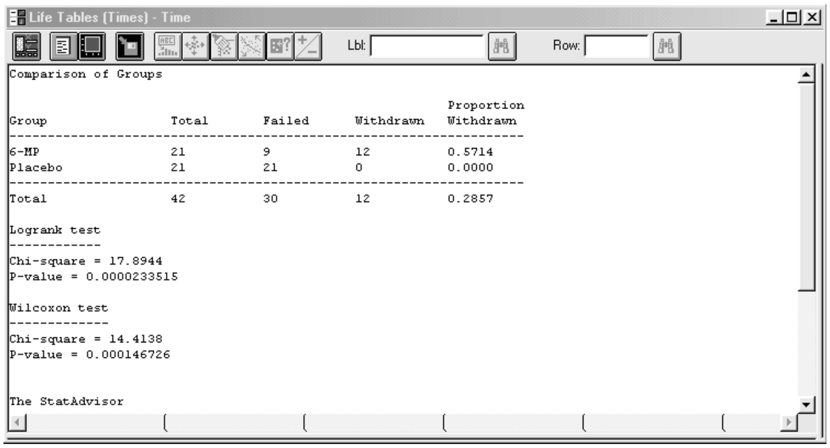
```
Percentile Table for Drug = 6-MP

                              Standard
Percentile     Estimate       Error
---------------------------------------------
75.0           11.0           3.26554
50.0           35.0
25.0           35.0
---------------------------------------------


Percentile Table for Drug = Placebo

                              Standard
Percentile     Estimate       Error
---------------------------------------------
75.0           4.0            1.44016
50.0           8.0            1.66905
25.0           12.0           1.79947
---------------------------------------------
```

Figure 12-12. Percentiles

Use the *Percentiles Options* dialog box to enter values for other percentiles. If you do not want to calculate percentiles, enter 0.

Group Comparisons

The Group Comparisons option creates a table that shows the results of a comparison of the groups in the study using the Logrank and Wilcoxon tests (see Figure 12-13). The table displays the number of items at risk and the number of terminations for each group. The Chi-Square statistic shows the results of testing for the equivalence of death rates between the two groups. Other values show the associated significance level so you can determine if the survival rates of the two groups are significantly different.



```
Comparison of Groups

                                                        Proportion
Group                Total       Failed      Withdrawn   Withdrawn
--------------------------------------------------------------------
6-MP                 21          9           12          0.5714
Placebo              21          21          0           0.0000
--------------------------------------------------------------------
Total                42          30          12          0.2857

Logrank test
------------
Chi-square = 17.8944
P-value = 0.0000233515

Wilcoxon test
-------------
Chi-square = 14.4138
P-value = 0.000146726


The StatAdvisor
```

Figure 12-13. Group Comparisons

Graphical Options

Survival Function

The Survival Function option creates a plot of the estimated survival function, which is the probability that an item will not fail before a specified length of time (see Figure 12-14).

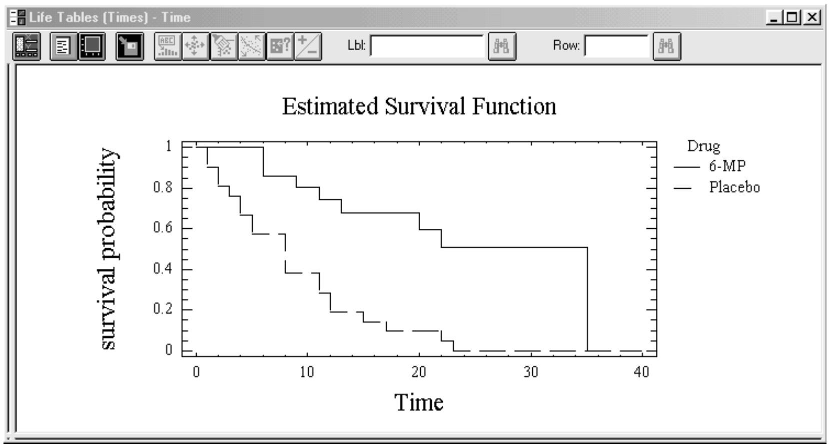


Figure 12-14. Survival Function

Log Survival Function

The Log Survival Function option creates a plot of the estimated log survival function, which is the probability that an item will not fail before a length of time you specify (see Figure 12-15).

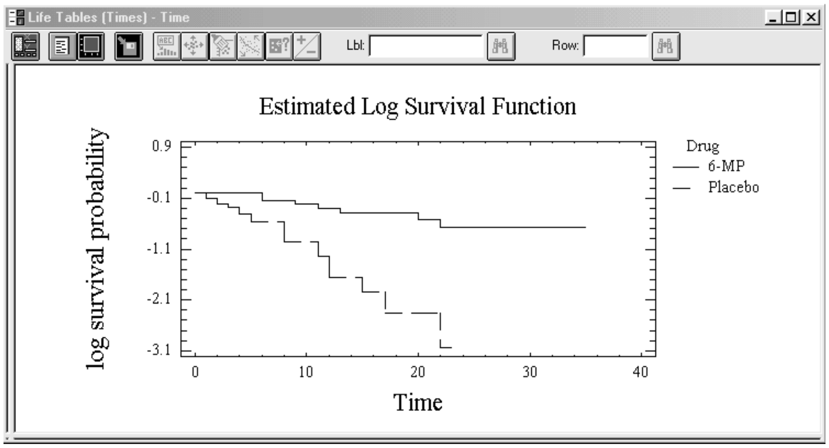


Figure 12-15. Log Survival Function

Cumulative Hazard Function

The Cumulative Hazard Function option creates a plot of the estimated cumulative hazard function, which is helpful when you are using failure data and need suggestions for possible parametric models (see Figure 12-16).

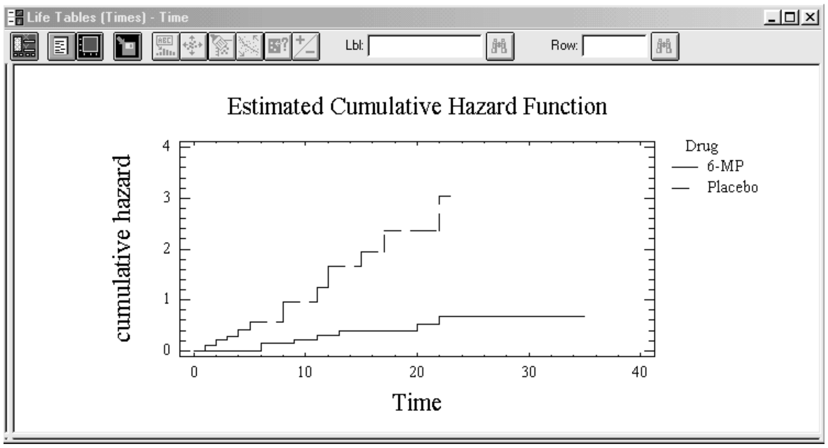

Life Tables (Times) - Time
Lbl:
Row:
Estimated Cumulative Hazard Function
cumulative hazard
0
1
2
3
4
0
10
20
30
40
Time
Drug
6-MP
Placebo

Figure 12-16. Cumulative Hazard Function

References

Armitage, P. and Berry, G. 1987. *Statistical Methods in Medical Research,* second edition. Oxford: Blackwell Scientific Publications.

Lawless, J. F. 1982. *Statistical Methods for Lifetime Data.* New York: Wiley.

Using the Weibull Analysis

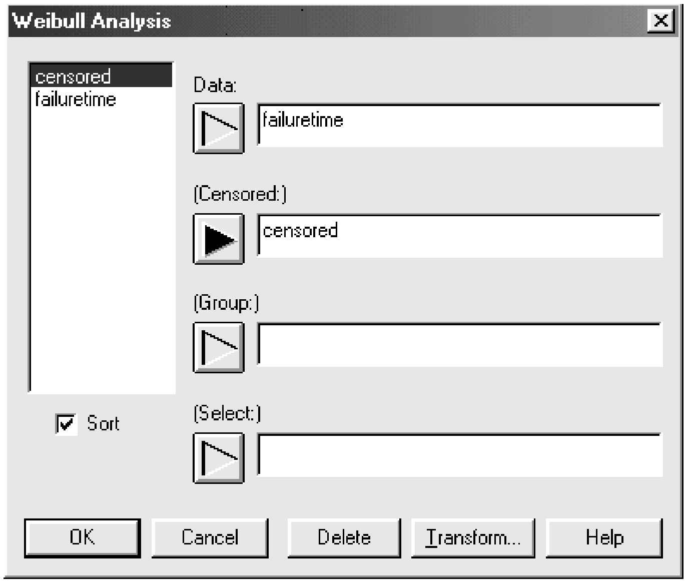
Weibull analysis originated in 1951 when Weibull claimed that his distribution or his “family” of distributions was applicable to a wide variety of problems that ranged from the yield strength of steel to the size of adult males born in the British Isles. Weibull never claimed that the distributions always worked or that they were always the best choice; however, time has shown that his statements were correct (Abernethy, Breneman, Medlin, and Reinman, 1983).

Today his analyses are used in numerous applications, particularly in fitting failure-time distributions to lifetime data. The flexible shape of the distribution makes it an appropriate fit for modeling the failure times of many different and diverse processes; therefore, the distribution can accurately describe most observed lifetimes.

The Weibull Analysis in STATGRAPHICS *Plus* handles various types of censored lifetime data. The data can include time-to-failure for manufactured items or survival times for individuals undergoing medical treatment in clinical trials. The basis for most applications of the analysis is a single failure class or mode from a single part or component. Usually the analysis begins with a few failures embedded in a large number of successful, unfailed, or "censored" units.

The analysis creates estimates for distribution parameters that are based on three estimation methods: Rank Regression, Maximum Likelihood, or Weibayes. It then calculates a variety of goodness-of-fit tests, as well as tail areas and critical values, and graphic displays that help determine how the survival rate will change over time.

To access the analysis, from the menus, choose: DESCRIBE... LIFE DATA... WEIBULL ANALYSIS... (see Figure 12-17).


Weibull Analysis

Data: failuretime

(Censored:) censored

(Group:)

(Select:)

☑ Sort

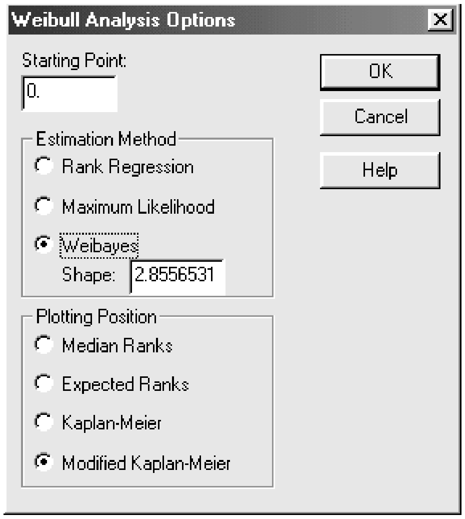
Figure 12-17. The Weibull Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the results of fitting a Weibull distribution to a set of data. The summary includes the name of the data and the censoring variables, as well as the name of the estimation method used. The summary then displays the sample size, the number of failures, the estimated parameters for the distribution you selected, and the starting point.

Use the *Weibull Analysis Options* dialog box to choose an origin for the distribution starting point, an estimation method, and a plotting position (see Figure 12-18). The estimation methods include rank regression, maximum likelihood, and Weibayes. The methods used to calculate the plotting position are median ranks, expected ranks, Kaplan-Meier, and modified Kaplan-Meier. *See Online Help for detailed descriptions for each of these methods.* This dialog box is available from all of the tabular and graphical options.



Starting Point: 0.

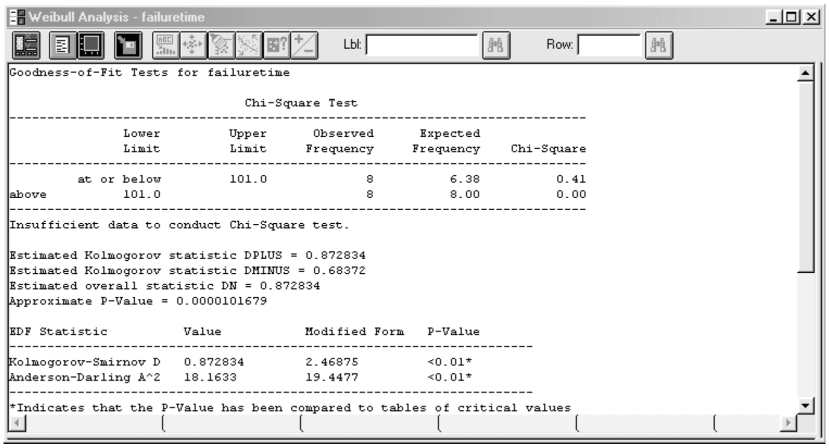
Estimation Method
- ○ Rank Regression
- ○ Maximum Likelihood
- ● Weibayes
  Shape: 2.8556531

Plotting Position
- ○ Median Ranks
- ○ Expected Ranks
- ○ Kaplan-Meier
- ● Modified Kaplan-Meier

Figure 12-18. Weibull Analysis Options Dialog Box

Goodness-of-Fit Tests

The Goodness-of-Fit Tests option calculates results for tests based on the Empirical Distribution Function (EDF). EDF is a step function that is calculated from the sample, which estimates the population distribution function. The tests are Chi-Square, Kolmogorov-Smirnov D, Kuiper V, Cramer-von Mises W^2, Watson U^2, and Anderson-Darling A^2 (see Figure 12-19).



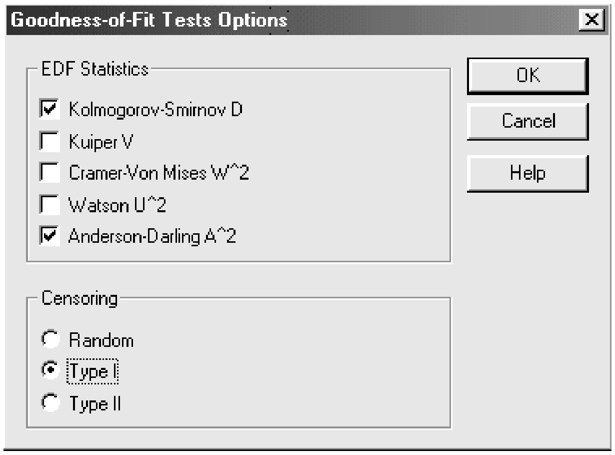
```
Goodness-of-Fit Tests for failuretime

                               Chi-Square Test
----------------------------------------------------------------------------
               Lower         Upper      Observed      Expected
               Limit         Limit     Frequency     Frequency    Chi-Square
----------------------------------------------------------------------------
         at or below         101.0             8          6.38          0.41
above          101.0                           8          8.00          0.00
----------------------------------------------------------------------------
Insufficient data to conduct Chi-Square test.

Estimated Kolmogorov statistic DPLUS = 0.872834
Estimated Kolmogorov statistic DMINUS = 0.68372
Estimated overall statistic DN = 0.872834
Approximate P-Value = 0.0000101679

EDF Statistic          Value           Modified Form   P-Value
---------------------------------------------------------------------
Kolmogorov-Smirnov D   0.872834        2.46875         <0.01*
Anderson-Darling A^2   18.1633         19.4477         <0.01*
---------------------------------------------------------------------
*Indicates that the P-Value has been compared to tables of critical values
```

Figure 12-19. Goodness-of-Fit Tests

Use the *Goodness-of-Fit Tests Options* dialog box to select the EDF statistics you want to calculate, as well as a censoring method (see Figure 12-20). The EDF statistics are Chi-Square, Kolmogorov-Smirnov D, Kuiper V, Cramer-von Mises W^2, Watson U^2, and Anderson-Darling $A^{2.}$ The censoring types are Random, Type I and Type II. *See Online Help for descriptions of EDF Statistics, in general, as well as for each of the separate tests.*

Tail Areas

The Tail Areas option calculates the area under the density curve at specified points (see Figure 12-21). The test uses the cumulative distribution function to calculate the probability that a random variable will fall below a given value. The test calculates tail areas for up to five critical values.



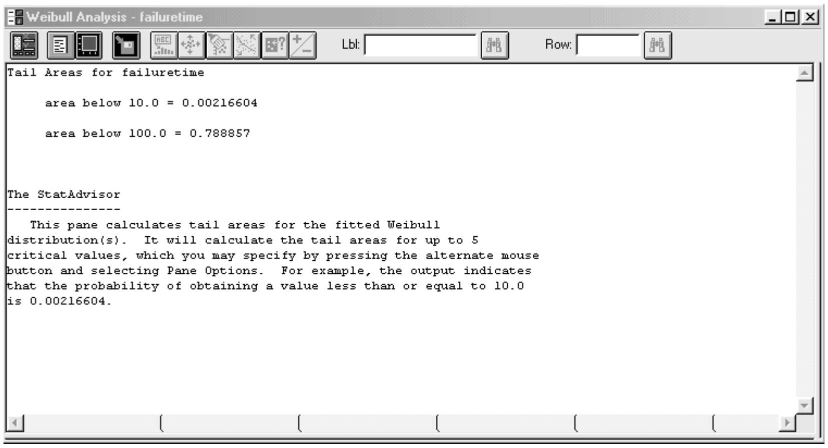
EDF Statistics

☑ Kolmogorov-Smirnov D
□ Kuiper V
□ Cramer-Von Mises W^2
□ Watson U^2
☑ Anderson-Darling A^2

Censoring

○ Random
● Type I
○ Type II

Figure 12-20. *Goodness-of-Fit Tests Options Dialog Box*



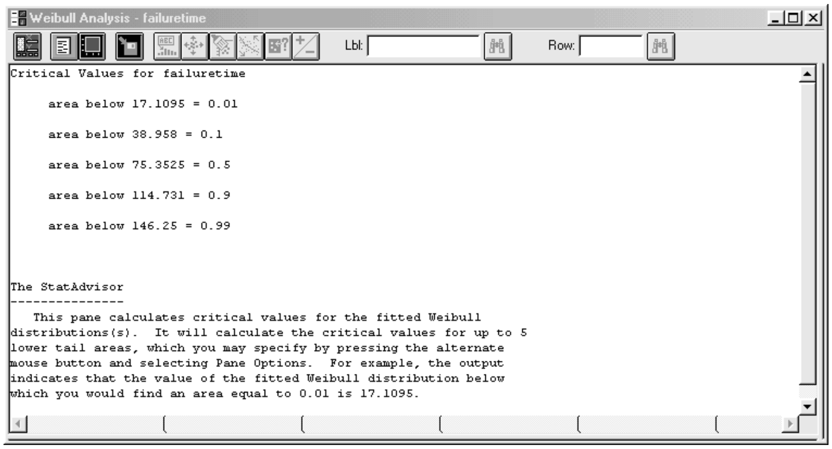
```
Tail Areas for failuretime

     area below 10.0 = 0.00216604

     area below 100.0 = 0.788857



The StatAdvisor
---------------
   This pane calculates tail areas for the fitted Weibull
distribution(s).  It will calculate the tail areas for up to 5
critical values, which you may specify by pressing the alternate mouse
button and selecting Pane Options.  For example, the output indicates
that the probability of obtaining a value less than or equal to 10.0
is 0.00216604.
```

Figure 12-21. *Tail Areas*

Use the *Tail Areas Options* dialog box to enter values that will be used to calculate the area under the curve.

Critical Values

The Critical Values option calculates critical values for the fitted Weibull distribution (see Figure 12-22). Critical values are the values below which you will find specified tail areas. The test calculates critical values for up to five lower tail areas.



```
Critical Values for failuretime

     area below 17.1095 = 0.01

     area below 38.958 = 0.1

     area below 75.3525 = 0.5

     area below 114.731 = 0.9

     area below 146.25 = 0.99



The StatAdvisor
---------------
   This pane calculates critical values for the fitted Weibull
distributions(s).  It will calculate the critical values for up to 5
lower tail areas, which you may specify by pressing the alternate
mouse button and selecting Pane Options.  For example, the output
indicates that the value of the fitted Weibull distribution below
which you would find an area equal to 0.01 is 17.1095.
```

Figure 12-22. Critical Values

Use the *Critical Values Options* dialog box to enter values for up to five tail areas.

Using Graphical Options

Weibull Plot

The Weibull Plot option creates a plot that helps determine if the data can be reasonably modeled using a Weibull distribution (see Figure 12-23). The values are plotted along the horizontal axis using a logarithmic scale. The vertical positions that correspond to each value are determined by the method you chose for the plotting position on the Weibull Analysis Options dialog box.

If the values are well described by a Weibull distribution, the plotted points should fall approximately along a straight line. The intercept and slope of the line are based on the current estimates of the Weibull parameters and the estimation method you selected.

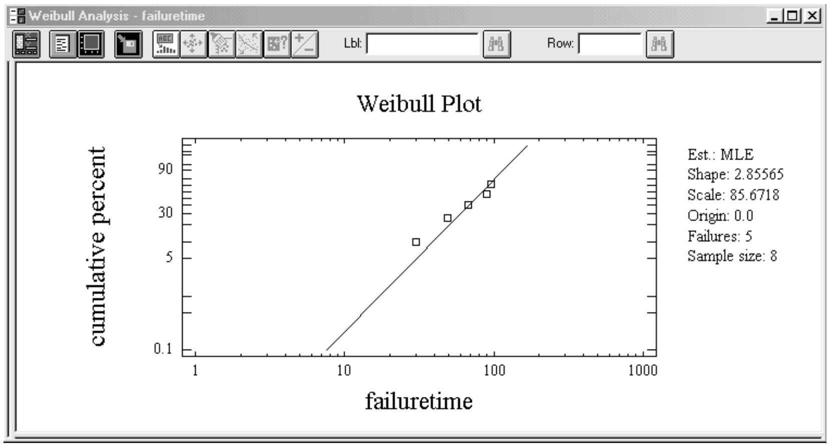


Figure 12-23. Weibull Plot

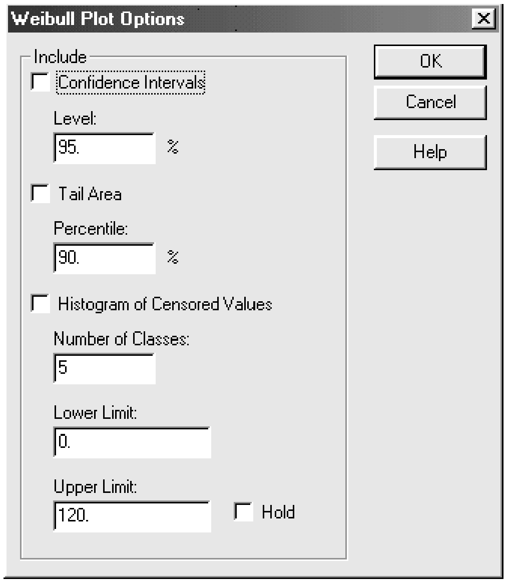
Use the *Weibull Plot Options* dialog box to indicate if you want to calculate confidence intervals and tail areas and, if so, to enter values that will be used to calculate them (see Figure 12-24). You can also indicate if you want to create a histogram of the censored values and, if so, to enter values for the number of classes, and the lower and upper limits.

Frequency Histogram

The Frequency Histogram option creates a histogram of the data that shows the intervals that were formed, and the values in each interval (the frequencies) that were calculated. A probability density function for the fitted Weibull distribution is superimposed on the histogram. If the distribution fits well, the top of the bars should be relatively close to the line (see Figure 12-25).

The plot also includes an overlay of the probability density function for the fitted Weibull distribution. If the distribution fits well, the top of the bars for the uncensored data should be relatively close to the line. The height of each bar is proportional to the number of observations that fall within that interval.

Use the *Frequency Tabulation Options* dialog box to enter values for the number of classes into which the data will be grouped, as well as for the



Weibull Plot Options

Include

□ Confidence Intervals

Level:

95. %

□ Tail Area

Percentile:

90. %

□ Histogram of Censored Values

Number of Classes:

5

Lower Limit:

0.

Upper Limit:

120. □ Hold

Figure 12-24. Weibull Plot Options Dialog Box

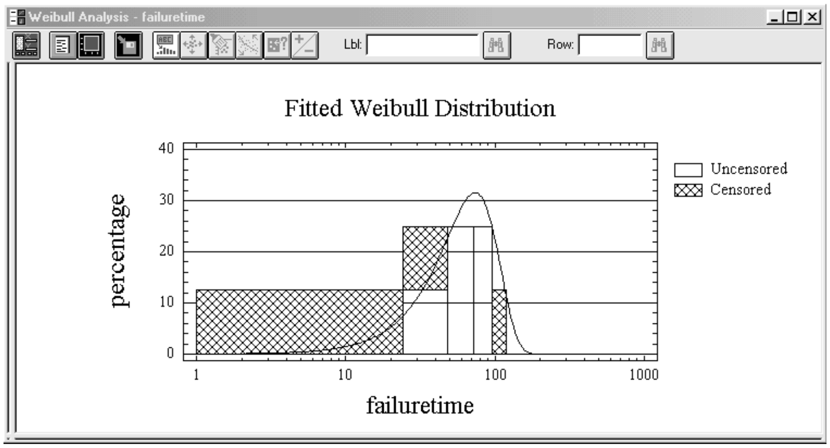
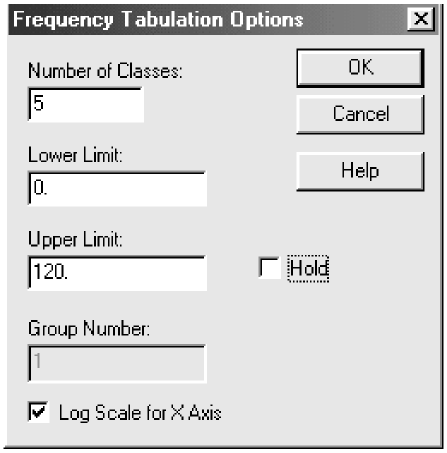


Figure 12-25. Frequency Histogram

Lower and Upper limits, and the group number. You can also indicate if a log scale should be used for the X-axis and if the scaling should be retained if you change values on the Analysis dialog box (see Figure 12-26).



Frequency Tabulation Options

Number of Classes: 5

Lower Limit: 0.

Upper Limit: 120.

□ Hold

Group Number: 1

☑ Log Scale for X Axis

Figure 12-26. Frequency Tabulation Options Dialog Box

Density Function

The Density Function Plot option creates a plot that shows the probability density function for the fitted Weibull distribution (see Figure 12-27). The program calculates the probability of attaining values within specified intervals by computing the area under this function.

Use the *Density Function Options* dialog box to indicate if a log scale should be used for the X-Axis.

CDF

The CDF option creates a plot that shows the cumulative distribution function for the Fitted Weibull distribution (see Figure 12-28). The height of the function indicates the probability of attaining a value less than or equal to the values on the X-axis.

Use the *CDF Function Options* dialog box to indicate if a log scale should be used for the X-Axis.

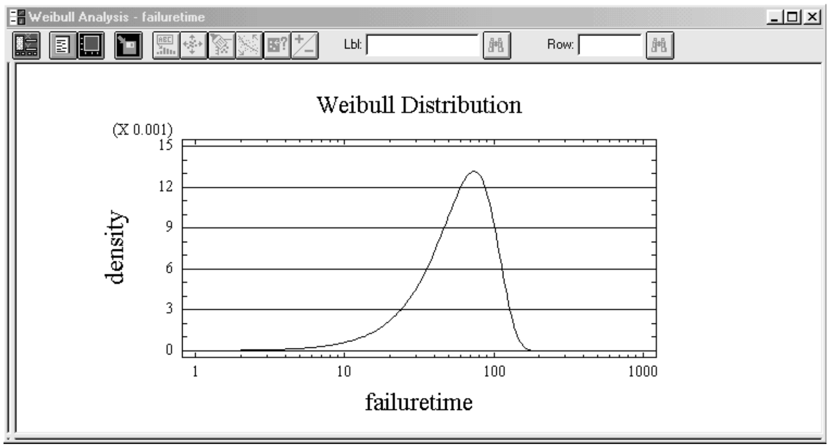


Figure 12-27.   Density Function Plot

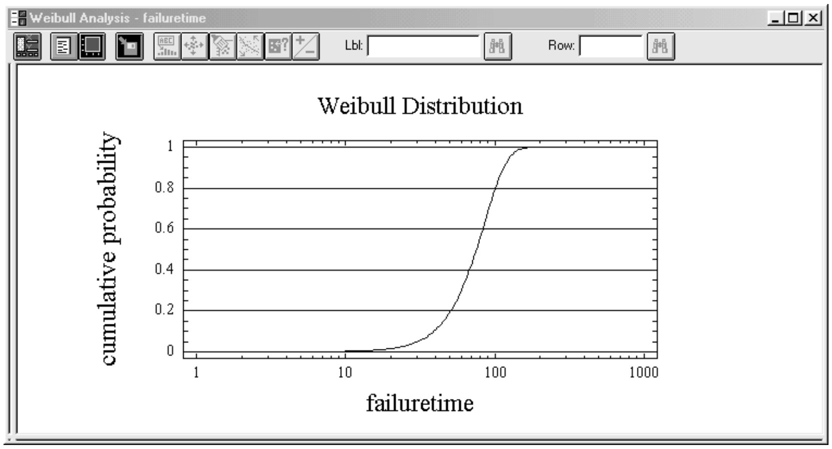


Figure 12-28.   CDF Plot

Survival Function

The Survival Function option creates a plot that shows the survival function for the fitted Weibull distribution (see Figure 12-29). The survival function indicates the probability of attaining a value greater than or equal to the values on the X-axis.

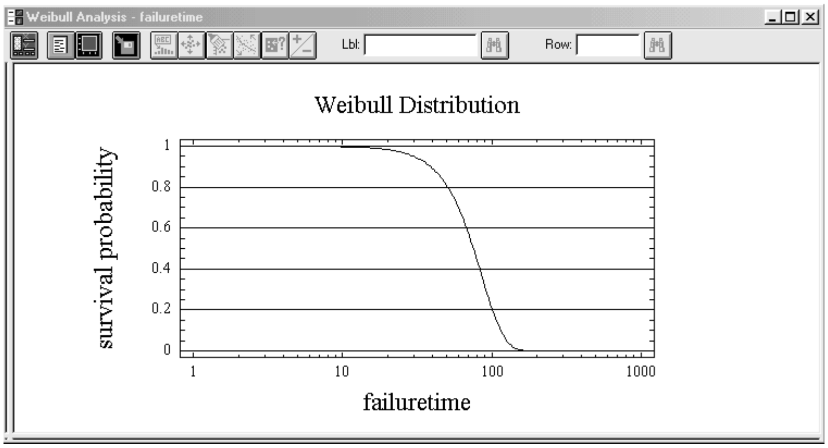


Figure 12-29. Survival Function Plot

Use the *Survival Function Options* dialog box to indicate if a log scale should be used for the X-Axis.

Log Survival Function

The Log Survival Function option creates a plot of the estimated log survival function, which is the probability that an item will not fail before a specified length of time (see Figure 12-30).

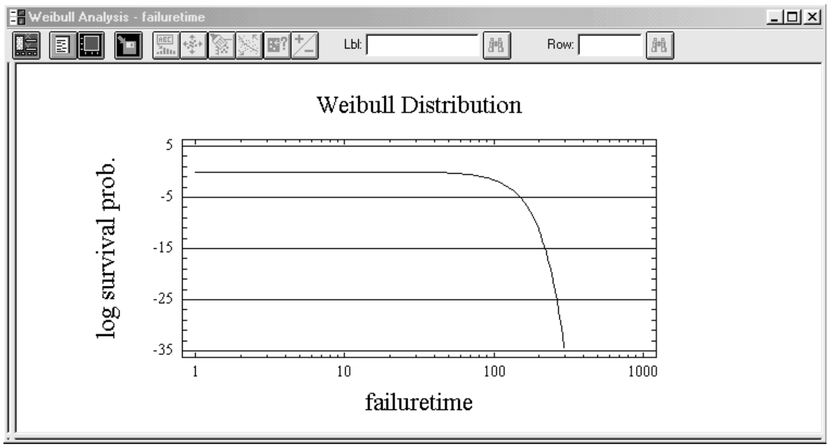


Figure 12-30. Log Survival Function Plot

Use the *Log Survival Function Options* dialog box to indicate if a log scale should be used for the X-Axis.

Hazard Function

The Hazard Function Plot option creates a plot of the estimated hazard function, which shows the hazard function for the fitted Weibull distribution (see Figure 12-31). The hazard function is equal to the probability density function divided by the survival function. When you model lifetime data, the hazard function represents the instantaneous failure rate.

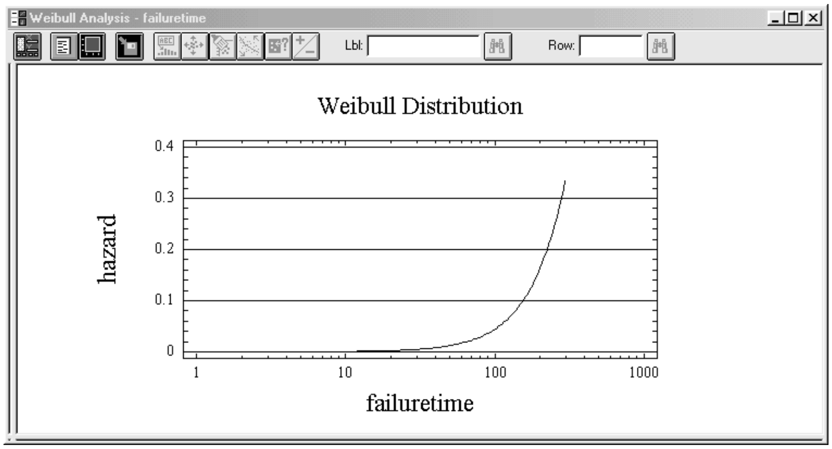


Figure 12-31. Hazard Function Plot

Use the *Hazard Function Options* dialog box to indicate if a log scale should be used for the X-Axis.

References

Abernethy, R. B., Breneman, J. E., Medlin, C. H., and Reinman, G. L. 1983. *Weibull Analysis Handbook*. Final Report for Period 1 July 1982 to 31 August 1983. Wright-Patterson AFB, Ohio: United States Air Force.

Collett, D. 1996. *Modelling Survival Data in Medical Research*. London: Chapman & Hall.

Kalbfleisch, J. D. and Prentice, R. L. 1980. *Statistical Analysis of Failure Time Data*. New York: Wiley.

D'Agostino, R. B. and Stephens, M. A. *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data.* New York: Wiley.

Nelson, W. B. 1982. *Applied Life Data Analysis*. New York: Wiley.

Tobias, P. and Trindade, D. 1995. *Applied Reliability*, second edition. New York: Chapman & Hall.

Using the Arrhenius Plot Analysis

When operating under normal conditions, it is often not possible to obtain sufficient amounts of data for fitting life distributions. However, by increasing stress levels, you can fail an adequate number of items, collect the data you need, and attempt to fit the distribution.

The Arrhenius Plot Analysis in STATGRAPHICS *Plus* fits an Arrhenius model to a set of percentiles taken at two or more study temperatures. The fitted model is then used to predict the corresponding percentile at a normal temperature. The analysis models the relationships between failure rates at high stress levels and normal operating levels. This analysis is a Physical Acceleration model that is useful when thermal stresses are significant. You can use this analysis to estimate the constants in an Arrhenius model.

The form of the Arrhenius model is:

$$P = Ae^{\text{Delta}/kT}$$

where

P = the estimated percentile

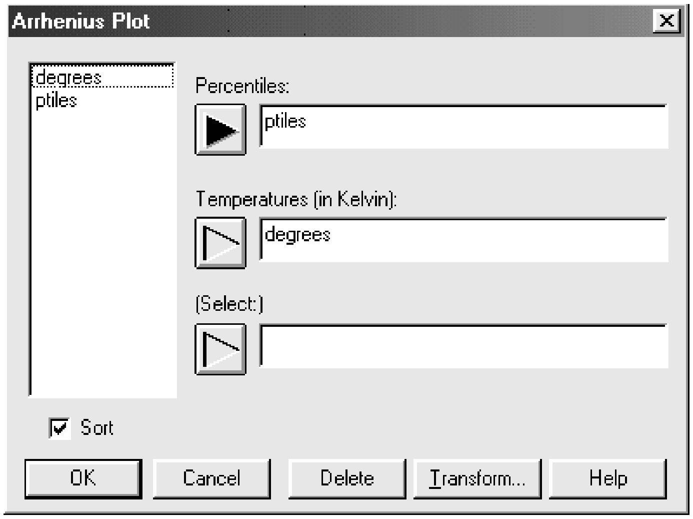
A and Delta = the unknown constants to be estimated

k = Boltzmann's Constant (8.617 10^{-5} EV/degrees Kelvin)

T = the temperature in degrees (Kelvin)

Using the estimated percentiles at the temperature at which it carries out the tests, the program calculates the unknown constants in the Arrhenius model.

To access the analysis, from the menus, choose: DESCRIBE... LIFE DATA... ARRHENIUS PLOTS... (see Figure 12-32).



Percentiles: ptiles

Temperatures (in Kelvin): degrees

(Select:)

☑ Sort

Figure 12-32. The Arrhenius Plot Analysis Dialog Box

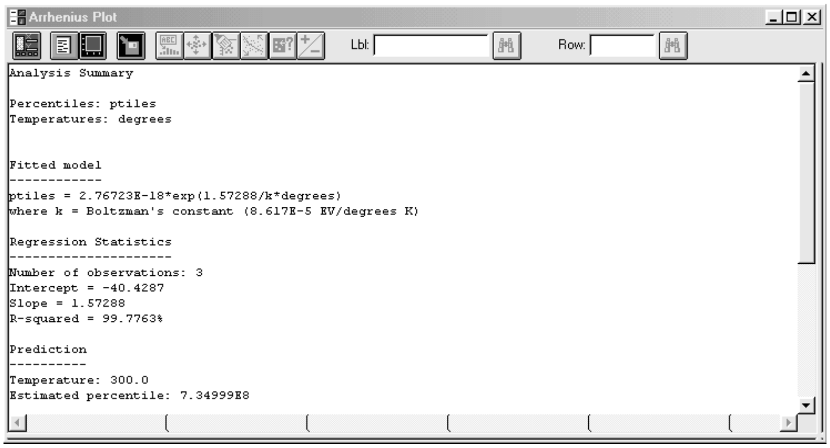
Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the names of the variables; the equation for the fitted model; the regression statistics, including the number of observations and values for the intercept, slope, and R-Squared; and the predictions, including values for the temperature, the estimated percentile, and the upper and lower limits (see Figure 12-33).

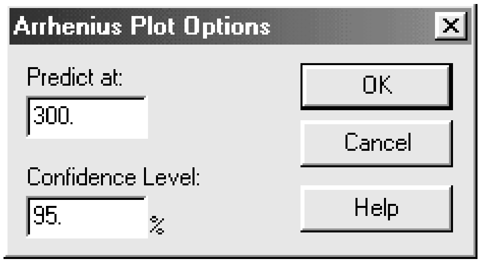
The analysis fits an Arrhenius model to a set of percentiles taken at two or more study temperatures. You can then use the fitted model to predict the corresponding percentile at a normal temperature.

Use the *Arrhenius Plot Options* dialog box to enter the temperature at which the prediction will be made and a confidence level (see Figure 12-34).



```
Analysis Summary

Percentiles: ptiles
Temperatures: degrees


Fitted model
------------
ptiles = 2.76723E-18*exp(1.57288/k*degrees)
where k = Boltzman's constant (8.617E-5 EV/degrees K)

Regression Statistics
---------------------
Number of observations: 3
Intercept = -40.4287
Slope = 1.57288
R-squared = 99.7763%

Prediction
----------
Temperature: 300.0
Estimated percentile: 7.34999E8
```

Figure 12-33.   Analysis Summary



Predict at: 300.

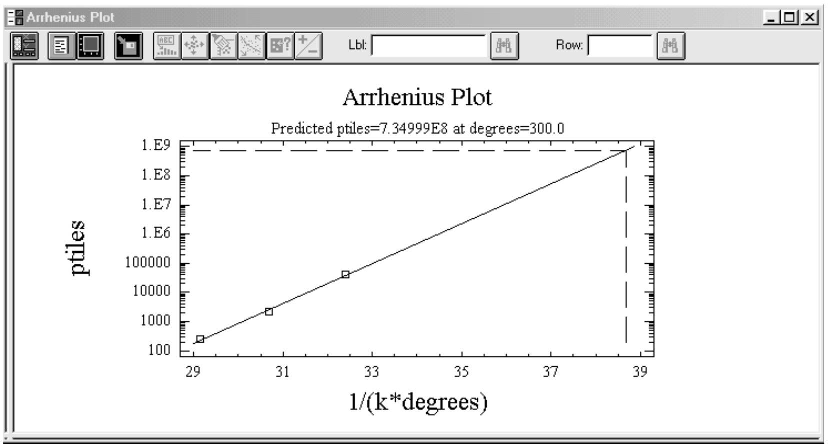
Confidence Level: 95. %

Figure 12-34.   Arrhenius Plot Options Dialog Box

Graphics Options

Arrhenius Plot

The Arrhenius Plot option creates a plot that shows the predicted percentile with the data and the fitted model (see Figure 12-35).



Predicted ptiles=7.34999E8 at degrees=300.0

Figure 12-35. Arrhenius Plot

References

Tobias, P. and Trindade, D. 1986. *Applied Reliability*. New York: Van Nostrand Reinhold, Inc.

13 Performing Hypothesis Tests and Determining Sample Size

This chapter discusses two analyses: Hypothesis Tests (Describe) and Sample Size Determination (Describe).

- **Hypothesis Tests (Describe) Analysis**
  This analysis performs hypothesis tests for four parameters: Normal Mean, Normal Sigma, Binomial Proportion, and Poisson Rate.

- **Sample Size Determination (Describe) Analysis**
  This analysis calculates sample sizes for four parameters: Normal Mean, Normal Sigma, Binomial Proportion, and Poisson Rate.

Using the Hypothesis Tests (Describe) Analysis

Hypothesis tests are based on a sample of data that determine which of two different states is true. The two states are commonly called the *null hypothesis* and the *alternative hypothesis*.

Hypothesis tests are a classical approach to assessing the statistical significance of findings, a technique that involves comparing empirically observed sample findings with theoretically expected findings — expected if the null hypothesis is true. This comparison allows you to compute the probability that the observed outcomes might be due to chance alone.

A simple example is testing a coin toss for fairness. Toss a coin a number of times, and note the number of heads. The sample data are the sequence of heads and tails, while the statistic is the number of heads divided by the number of tosses (10). If the proportion of heads is around 0.5, you can conclude that the evidence agrees with the hypothesis that the true probability of heads is 0.5.

You can use the hypothesis tests explained in this chapter to estimate and test hypotheses on four parameters that involve a single sample: normal mean, normal sigma, binomial proportion, or poisson rate for a single random sample. This means you can enter sample statistics directly into the analysis instead of first using the One-Variable Analysis. *See Chapter 15, Comparing Two Data Samples, for a discussion about hypothesis tests for two independent samples.*

When you are testing the normal mean and normal sigma parameters, the program calculates confidence intervals for the mean and variance using the *t*- and chi-square distributions, respectively, and assumes that the data come from a normal distribution. You can set the percentage level for the confidence intervals.

For example, suppose you want to study how working in groups instead of working alone affects performance. The hypothetical average is based on working alone. Eighty subjects volunteered for the study and worked in teams of four. You provided the subjects with a large number of problems to solve within a 20-minute time period, then recorded the number of correct answers. Your alternative or research hypothesis was that those who worked in teams were more efficient than those who worked individually. To examine this hypothesis, you tried to find evidence that allowed you to reject the null hypothesis, which was stated as: There is no difference between the average score of those who worked individually and those who worked in teams.

STATGRAPHICS *Plus* contains three alternative hypotheses:

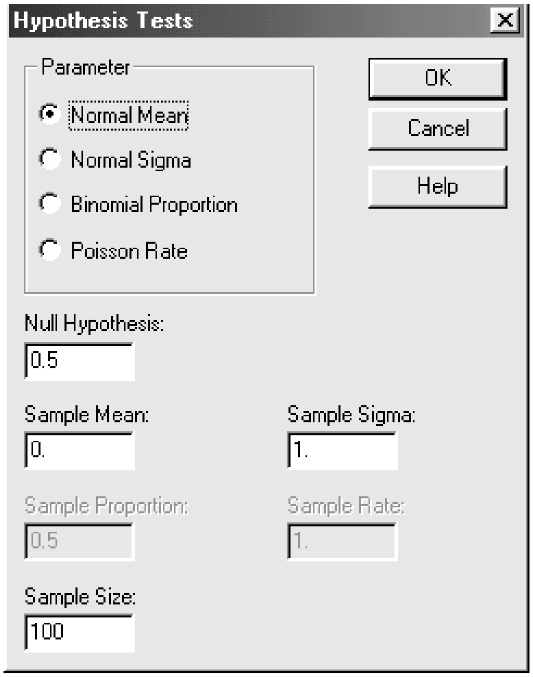
- Not Equal
- Less Than
- Greater Than.

To access the analysis, from the menus, choose: DESCRIBE... HYPOTHESIS TESTS... (see Figure 13-1).

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the test results for the parameter you selected.



Hypothesis Tests

Parameter

● Normal Mean

○ Normal Sigma

○ Binomial Proportion

○ Poisson Rate

Null Hypothesis: 0.5

Sample Mean: 0.

Sample Sigma: 1.

Sample Proportion: 0.5

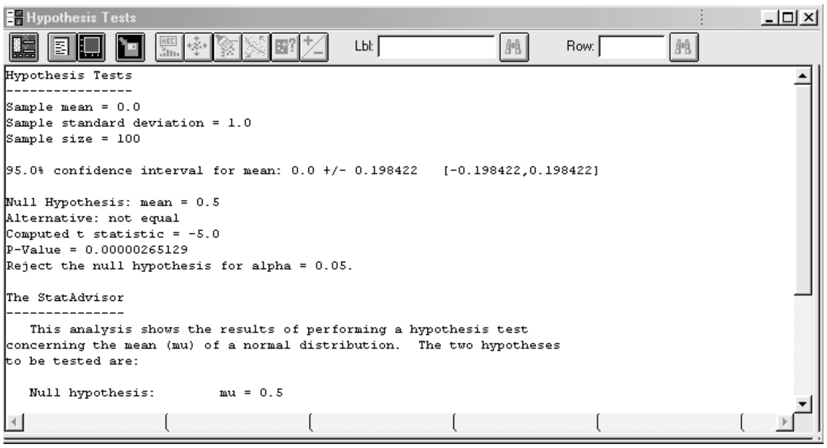
Sample Rate: 1.

Sample Size: 100

Figure 13-1. The Hypothesis Tests (Describe) Analysis Dialog Box

- **Analysis Summary for the Normal Mean Option**
  The summary for the Normal Mean Test displays the sample mean, sample standard deviation, and sample size as well as the given confidence interval for the mean. Information in the lower portion of the text pane displays the theoretical mean for the null hypothesis, the type of alternative hypothesis, the computed *t*-statistic, the *p*-value, and the results — either reject or do not reject the null hypothesis (see Figure 13-2).

- **Analysis Summary for the Normal Sigma Option**
  The summary for the Normal Sigma Test displays the sample standard deviation and sample size as well as the given confidence interval for sigma. Information in the lower portion of the text pane displays the theoretical standard deviation for the null hypothesis, the type of alternative hypothesis test, the computed chi-square statistic, the *p*-value, and the results — either reject or do not reject the null hypothesis for the given alpha.



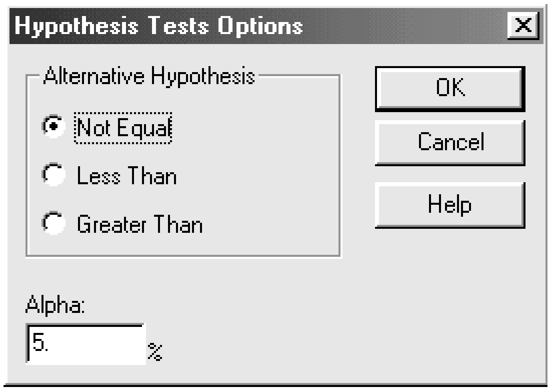
```
Hypothesis Tests
----------------
Sample mean = 0.0
Sample standard deviation = 1.0
Sample size = 100

95.0% confidence interval for mean: 0.0 +/- 0.198422   [-0.198422,0.198422]

Null Hypothesis: mean = 0.5
Alternative: not equal
Computed t statistic = -5.0
P-Value = 0.00000265129
Reject the null hypothesis for alpha = 0.05.

The StatAdvisor
---------------
   This analysis shows the results of performing a hypothesis test
concerning the mean (mu) of a normal distribution.  The two hypotheses
to be tested are:

   Null hypothesis:        mu = 0.5
```

Figure 13-2. Analysis Summary for the Normal Mean Option

- **Analysis Summary for the Binomial Proportion Option**
  The summary for the Binomial Proportion Test displays the sample proportion and sample size as well as the given confidence interval for *p*. Information in the lower portion of the text pane displays the hypothetical value for the proportion, the name of the alternative hypothesis test, the *p*-value, and the results — either reject or do not reject the null hypothesis for the given alpha.

- **Analysis Summary for the Poisson Rate Option**
  The summary for the Poisson Rate Test displays the sample rate and sample size as well as the given confidence interval for rate. Information in the lower portion of the text pane displays the hypothetical value for the rate, the type of alternative test, the *p*-value, and the results — either reject or do not reject the null hypothesis for the given alpha.

Use the *Hypothesis Tests Options* dialog box to choose an alternative hypothesis test and to enter a value for alpha, which represents the risk of a Type I error (see Figure 13-3).



Hypothesis Tests Options

Alternative Hypothesis

● Not Equal

○ Less Than

○ Greater Than

Alpha: 5. %

Figure 13-3. Hypothesis Tests Options Dialog Box

Graphical Options

Power Curve

The Power Curve option plots the probability of the null hypothesis being rejected versus a range of values for the parameter you are testing. The vertical axis of the Power Curve shows the probability of rejecting the null hypothesis; the horizontal axis represents the size of the effect (see Figure 13-4).

Use the *Power Curve Options* dialog box to enter a value for the assumed sigma (standard deviation). This dialog box is available only when you use the Normal Mean option.

References

Guttman, I., Wilks, S. S., and Hunter, J. S. 1982. *Introductory Engineering Statistics*, third edition. New York: Wiley.

Snedecor, G. W. and Cochran, W. G. 1967. *Statistical Methods*, sixth edition. Ames, Iowa: Iowa State University Press.

Vogt, Paul. 1993. *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences.* Newbury Park, California: Sage Publications, Inc.

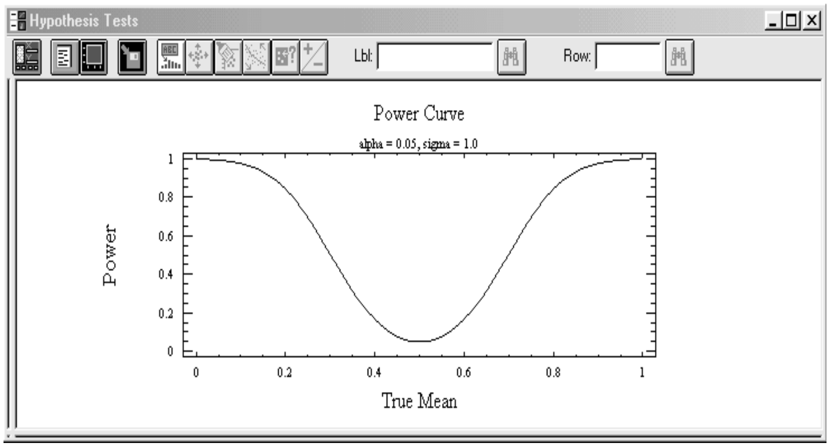


Figure 13-4. Power Curve

Using the Sample Size Determination (Describe) Analysis

In many statistical procedures, particularly in the planning stages and before data have been collected, it is important to determine the size of the sample required to adequately address the objectives of the study. Many research questions involve means or proportions where you must either

- estimate a population mean, proportion, or difference with a certain degree of accuracy, or
- test the statistical significance of the difference, either between a sample mean or proportion and some hypothesized value, or between means or proportions from two different samples.

An appropriate sample size should provide the degree of accuracy you need to estimate population means, proportions, or differences. It should also allow you to control the risk of reaching incorrect conclusions when you test for statistical significance.

When the estimates of means and proportions come from large samples, there is less sample-to-sample variation than in estimates that come from smaller samples. Large-sample estimates tend to be more accurate than estimates from smaller samples.

One factor that affects sample size requirements is the degree of accuracy you want to achieve — the maximum tolerable difference between the population parameter you are estimating and its true value.

The degree of accuracy attained by any sample size also depends on the spread (variability) of the variable you are testing. Variability is typically measured by the variance or standard deviation. In general, the smaller the variance, the smaller the number of subjects necessary to achieve the degree of accuracy you desire. It is better to err on the "high" side because this provides an estimated sample size that has, at least, the accuracy you desire.

Another factor that affects sample size is the degree of confidence that will actually achieve accuracy. Except in trivial cases, there is no sample size, short of a census, that allows you to be certain of achieving the accuracy you want.

The Sample Size Determination (Describe) Analysis in STATGRAPHICS *Plus* calculates sampling sizes for four parameters: Normal Mean, Normal Sigma, Binomial Proportion, and Poisson Rate.

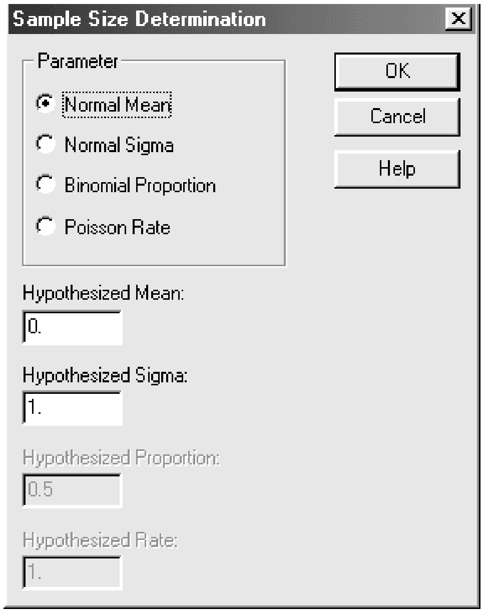
For a Normal distribution, STATGRAPHICS *Plus* calculates the sample size required to test hypotheses, while controlling for absolute or relative error (alpha). You can also use this analysis to determine the sample size necessary for a given power (beta) for a specific null hypothesis, alternative hypothesis, alpha level, and standard deviation, or you can use it to determine the power curve for given sample sizes.

Alpha represents the risk of a Type I error, rejecting the null hypothesis when in fact it is correct. Beta is the risk of a Type II error; for example, accepting a false null hypothesis. To estimate a normal sigma, STATGRAPHICS *Plus* detects certain differences when you specify an absolute or relative error, or when you specify a power.

The analysis for a Binomial distribution uses a normal approximation; therefore, you should not use this distribution for very small samples. For proportions near 0.5, the analysis provides useful approximations with sample sizes of 20 or more. For proportions less than 0.4, the analysis provides useful approximations with sample sizes of 50 or more.

For a Poisson distribution, the program estimates the mean of the distribution. Because it uses a normal approximation, you should not use it for very small samples. The approximation is reasonably close to normal when the mean of the Poisson distribution is 10 or larger (for example, when $n = 50$ and $p = 0.2$). Using a Poisson distribution is advantageous because it relies only on the mean frequency, which is applicable in situations where the probability of an event occurring is very small compared with the size of the sample.

To access the analysis, from the menus, choose: DESCRIBE... SAMPLE SIZE DETERMINATION... (see Figure 13-5).



Sample Size Determination

Parameter

● Normal Mean

○ Normal Sigma

○ Binomial Proportion

○ Poisson Rate

Hypothesized Mean: 0.

Hypothesized Sigma: 1.

Hypothesized Proportion: 0.5

Hypothesized Rate: 1.

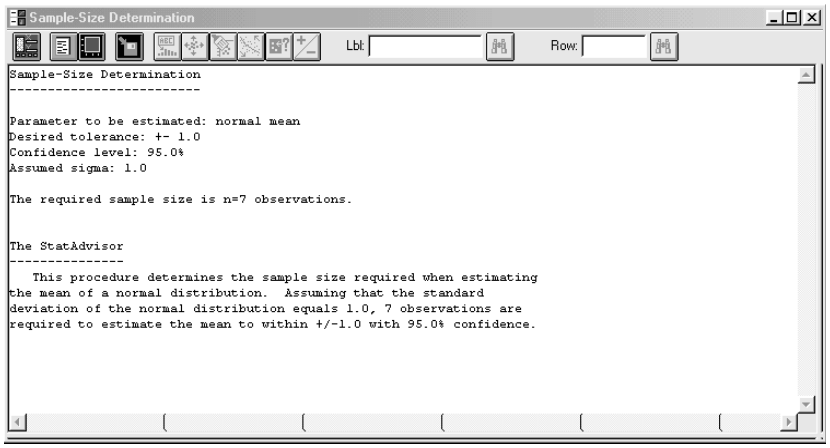
Figure 13-5. The Sample-Size Determination (Describe) Analysis Dialog Box

Tabular Options

Analysis Summary

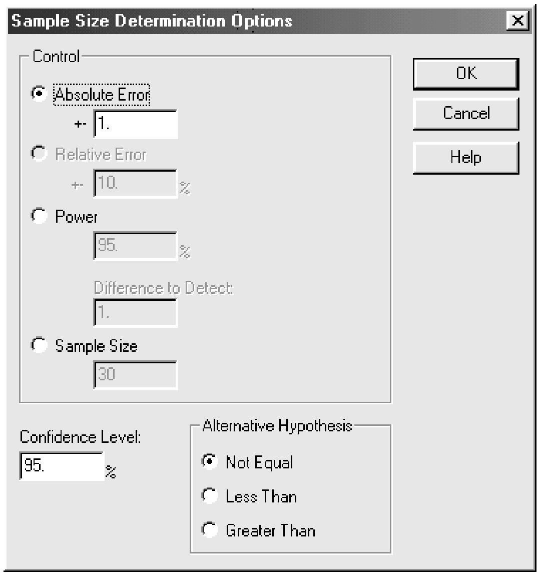
The Analysis Summary option creates a summary that shows the name of the parameter and the controls you selected (see Figure 13-6). The last line displays the calculated sample size. A separate summary is displayed, depending on the parameter you are using.

Use the *Sample Size Determination Options* dialog box to choose the type and quantity of control that will be used to determine the required sample size (see Figure 13-7). You can also enter a percentage that will be used as the significance level, and to choose an alternative hypothesis that will be used to determine which of two states is true. The option you choose depends on the type of analysis for which you are determining a sample size.



```
Sample-Size Determination
------------------------

Parameter to be estimated: normal mean
Desired tolerance: +- 1.0
Confidence level: 95.0%
Assumed sigma: 1.0

The required sample size is n=7 observations.


The StatAdvisor
---------------
   This procedure determines the sample size required when estimating
the mean of a normal distribution.  Assuming that the standard
deviation of the normal distribution equals 1.0, 7 observations are
required to estimate the mean to within +/-1.0 with 95.0% confidence.
```

Figure 13-6. Analysis Summary



Control

● Absolute Error
+- 1.

○ Relative Error
+- 10. %

○ Power
95. %

Difference to Detect:
1.

○ Sample Size
30

Confidence Level:
95. %

Alternative Hypothesis

● Not Equal

○ Less Than

○ Greater Than

Figure 13-7. Sample-Size Determination Options Dialog Box

Graphical Options

Power Curve

The Power Curve option creates a plot of the power versus the true value of the mean for a given sample size (see Figure 13-8).

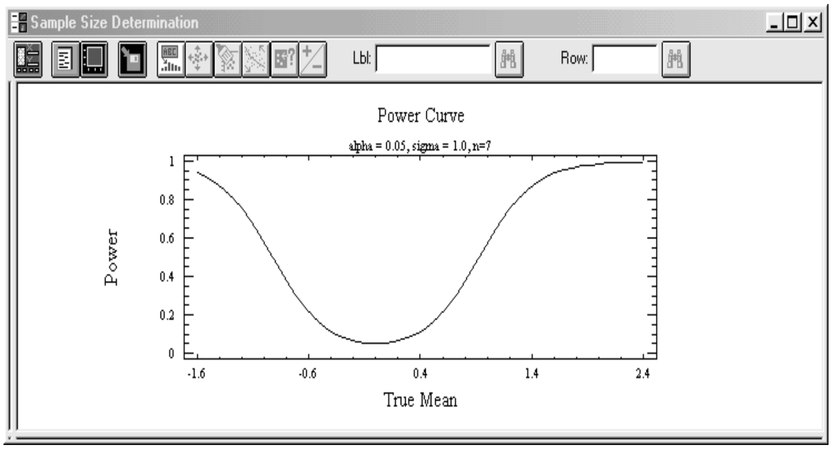


Figure 13-8. Power Curve

References

Haaland, P. 1989. *Experimental Design in Biotechnology*. New York: Marcel Dekker.

Hays, W. L. 1981. *Statistics,* third edition. New York: Holt, Rinehart, and Winston.

14 Comparing Two Data Samples

This chapter presents the analyses for which you can calculate descriptive statistics as well as parametric and nonparametric tests for comparing two samples. The Two-Sample Comparison Analysis compares two different samples of data using the two-sample *t*-test for dependent samples. The Paired-Sample Comparison Analysis parallels the Two-Sample Comparison Analysis except, it compares data collected in pairs, and the standard two-sample *t*-test is inappropriate because the independence assumption is not valid. The Hypothesis Test (Compare) and Sample Size Determination (Compare) analyses are used to perform hypothesis tests and to determine sample sizes for these types of data.

Assessing the Problem

Whether the problem you are solving is simple or complex, making a detailed examination of the data is required. Examining data involves investigating the distribution of the values, and testing for normality and measures of variation, including the mean, standard deviation, and confidence intervals. Looking at the distribution of the values is important for determining the appropriateness of the statistical methods you are planning to use for either testing hypotheses or building models.

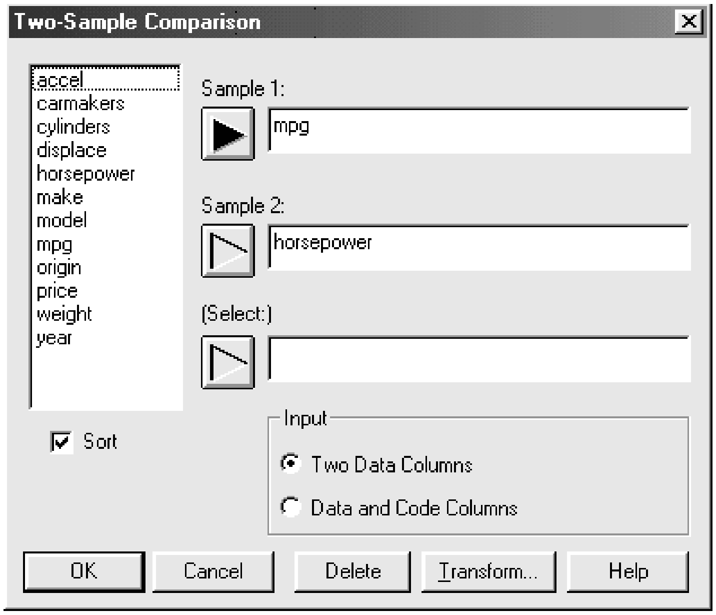
When data are normally distributed, calculating basic descriptive statistics may be an appropriate way to summarize the data. These types of statistics provide basic information such as the mean, minimum and maximum values, different measures of variation, as well as information about the shape of the distribution. The shape of the distribution is important because it provides information about the frequency of the values from different ranges of the variable, and researchers are typically interested in how well a distribution approximates a normal distribution.

What do you do when you know nothing about the assumptions underlying the population, or when the assumptions are questionable?   In these cases, nonparametric tests are often more powerful than statistical tests based on the assumption of normality. For example, when you compare two independent samples, the Wilcoxon Mann-Whitney test does not assume that the difference between the samples is normally distributed, whereas its parametric counterpart, the two-sample *t*-test does. At least one nonparametric test exists for each of its parametric counterpart.

Using the Two-Sample Comparison Analysis

The primary purposes of the Two-Sample Comparison Analysis are to: calculate confidence intervals for the difference between the population means and the ratio of the population variances; and to compare hypothesis tests of the means, variances, and medians. The analysis runs tests to determine if there are statistically significant differences between the two samples, then creates various reports and graphs that contain the results for each data sample.

To access the analysis, from the menus, choose: COMPARE... TWO-SAMPLE... TWO-SAMPLE COMPARISON... (see Figure 14-1).



Two-Sample Comparison

Sample 1: mpg

Sample 2: horsepower

(Select:)

☑ Sort

Input

● Two Data Columns

○ Data and Code Columns

Figure 14-1. The Two-Sample Comparison Analysis Dialog Box

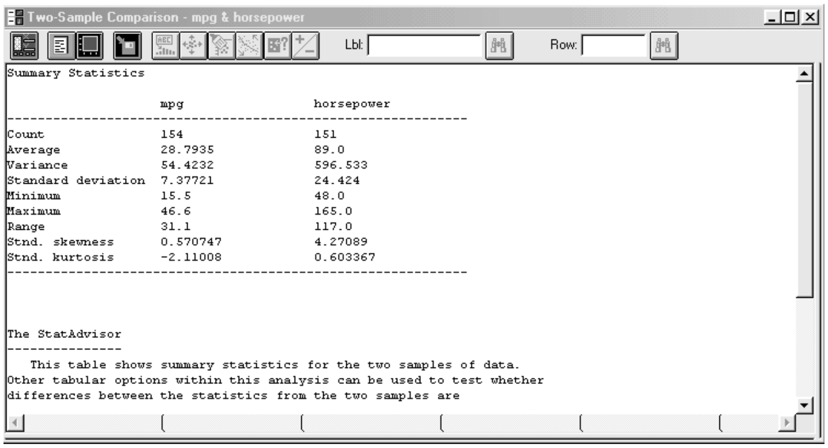
Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that includes the names of the variables for each of the two samples, and the number and range of the values in each sample.

Summary Statistics

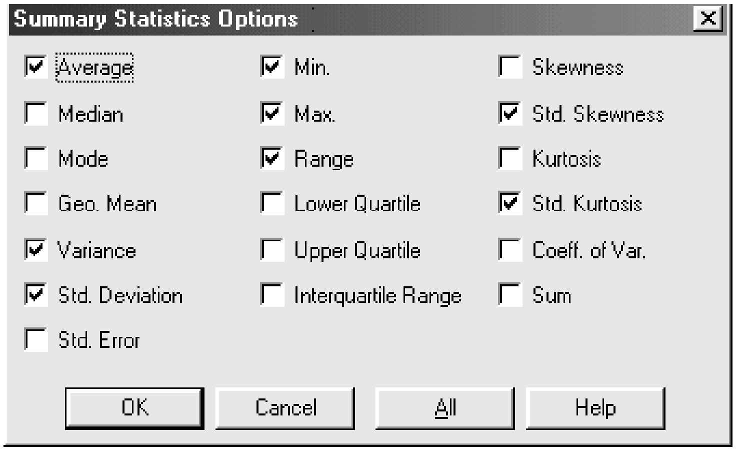
The Summary Statistics option creates statistical information about the statistics you select, and includes the number of values in each variable (Count); the average, variance, and standard deviation; as well as the minimum and maximum values. It also includes values for the range and the standardized skewness and standardized kurtosis (see Figure 14-2). The standard skewness and standardized kurtosis values are of particular interest because they help to determine if the samples are from normal distributions. Values outside the range of -2 to +2 indicate a significant departure from normality, which tends to invalidate the tests that compare the standard deviations.



```
Summary Statistics

                    mpg                 horsepower
------------------------------------------------------------
Count               154                 151
Average             28.7935             89.0
Variance            54.4232             596.533
Standard deviation  7.37721             24.424
Minimum             15.5                48.0
Maximum             46.6                165.0
Range               31.1                117.0
Stnd. skewness      0.570747            4.27089
Stnd. kurtosis      -2.11008            0.603367
------------------------------------------------------------



The StatAdvisor
---------------
   This table shows summary statistics for the two samples of data.
Other tabular options within this analysis can be used to test whether
differences between the statistics from the two samples are
```

Figure 14-2.   Summary Statistics

Use the *Summary Statistics Options* dialog box to select the statistics that will be calculated (see Figure 14-3).



☑ Average
☐ Median
☐ Mode
☐ Geo. Mean
☑ Variance
☑ Std. Deviation
☐ Std. Error
☑ Min.
☑ Max.
☑ Range
☐ Lower Quartile
☐ Upper Quartile
☐ Interquartile Range
☐ Skewness
☑ Std. Skewness
☐ Kurtosis
☑ Std. Kurtosis
☐ Coeff. of Var.
☐ Sum

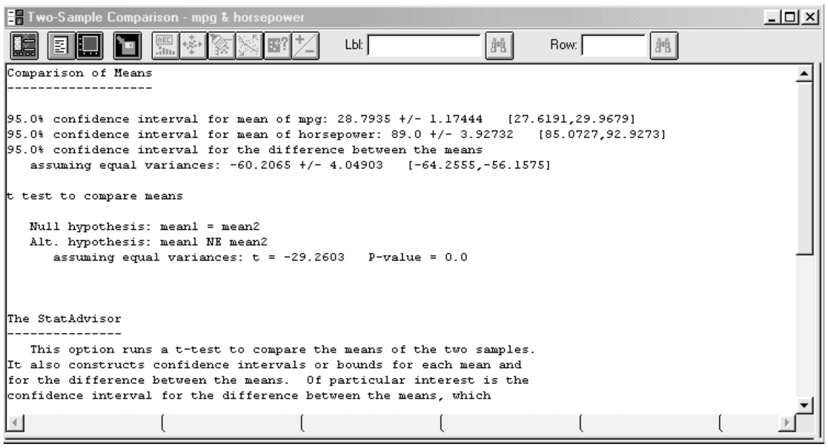
Figure 14-3. Summary Statistics Options Dialog Box

Comparison of Means

The Comparison of Means option runs a *t*-test to compare the means of the two samples (see Figure 14-4). It also constructs confidence intervals for each mean and for the difference between the means. The width of the confidence interval gives an idea about the certainty of the difference in the means; for instance, a very wide interval may indicate that you should collect more data before you make any definite conclusions.

A confidence interval for the difference between two means specifies a range of values within which the difference between the means of the two populations may lie. For example, a manufacturer who wants to estimate the difference in mean daily output from two machines, or a medical researcher who wants to estimate the difference in mean response by patients who are receiving two different drugs. If the confidence interval includes 0, there is no significant difference between the means of the two populations, at a given level of confidence.

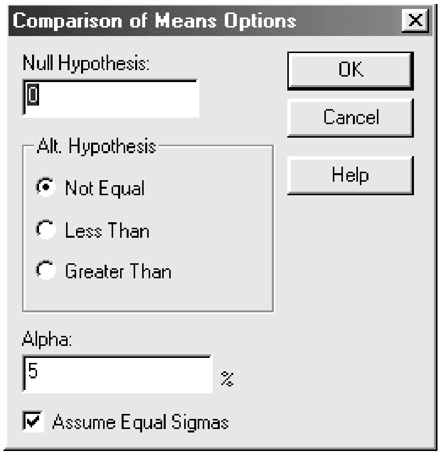
You can also test a specific hypothesis regarding the difference between the means of the populations from which the two samples originate. For example, if the *p*-value is less than 0.05, you can reject the null hypothesis in favor of the alternative hypothesis with 95 percent confidence.



```
Comparison of Means
-------------------

95.0% confidence interval for mean of mpg: 28.7935 +/- 1.17444   [27.6191,29.9679]
95.0% confidence interval for mean of horsepower: 89.0 +/- 3.92732   [85.0727,92.9273]
95.0% confidence interval for the difference between the means
   assuming equal variances: -60.2065 +/- 4.04903   [-64.2555,-56.1575]

t test to compare means

   Null hypothesis: mean1 = mean2
   Alt. hypothesis: mean1 NE mean2
      assuming equal variances: t = -29.2603   P-value = 0.0



The StatAdvisor
---------------
   This option runs a t-test to compare the means of the two samples.
It also constructs confidence intervals or bounds for each mean and
for the difference between the means.  Of particular interest is the
confidence interval for the difference between the means, which
```

Figure 14-4. Comparison of Means

Use the *Comparison of Means Options* dialog box to enter a value for the null hypothesis, to select an alternative hypothesis, to enter a percentage for alpha, and to indicate if the test should assume equal sigmas (see Figure 14-5).



Null Hypothesis: 0

Alt. Hypothesis

● Not Equal

○ Less Than

○ Greater Than

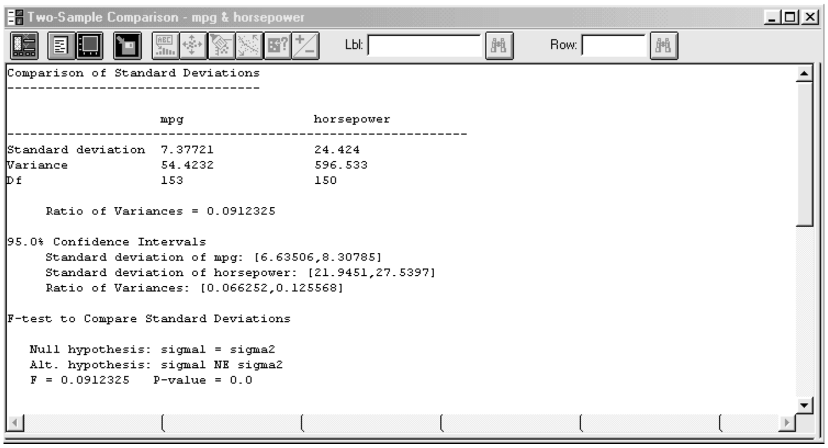
Alpha: 5 %

☑ Assume Equal Sigmas

Figure 14-5. Comparison of Means Options Dialog Box

Comparison of Standard Deviations

The Comparison of Standard Deviations option calculates confidence intervals and tests for the ratio of the variances of the populations by performing an F-test that compares the variances of two samples (see Figure 14-6). The option also constructs confidence intervals for each standard deviation, and for the ratio of the variances. If an interval does not contain a value of 1.0, there is a statistically significant difference between the standard deviations for the two samples at a given confidence level.



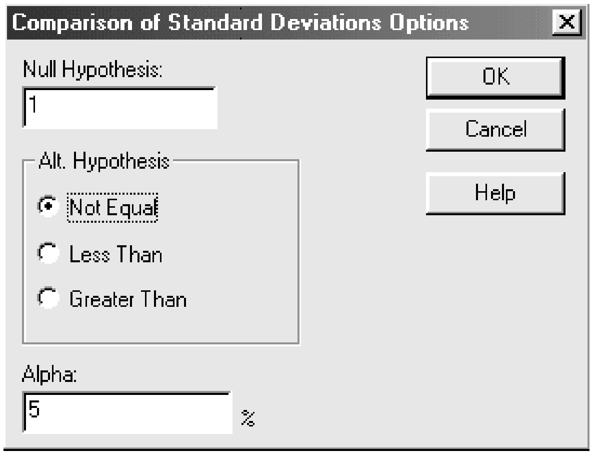
```
Comparison of Standard Deviations
---------------------------------


                    mpg                 horsepower
------------------------------------------------------------
Standard deviation  7.37721             24.424
Variance            54.4232             596.533
Df                  153                 150

     Ratio of Variances = 0.0912325

95.0% Confidence Intervals
     Standard deviation of mpg: [6.63506,8.30785]
     Standard deviation of horsepower: [21.9451,27.5397]
     Ratio of Variances: [0.066252,0.125568]

F-test to Compare Standard Deviations

   Null hypothesis: sigma1 = sigma2
   Alt. hypothesis: sigma1 NE sigma2
   F = 0.0912325   P-value = 0.0
```

Figure 14-6. Comparison of Standard Deviations

Use the *Comparison of Standard Deviations Options* dialog box to enter a value for the null hypothesis, to select an alternative hypothesis, and to enter a percentage for alpha (see Figure 14-7).

Comparison of Medians

The Comparison of Medians option performs a Mann-Whitney W test to compare the medians for the two samples (see Figure 14-8). The Mann-Whitney test, also known as the Wilcoxon test, does not require assumptions about the shape of the underlying distributions. It tests the hypothesis that two independent samples are from populations that have the same distribution. The test does not require that the variable be measured on an interval scale.



Null Hypothesis: 1

Alt. Hypothesis

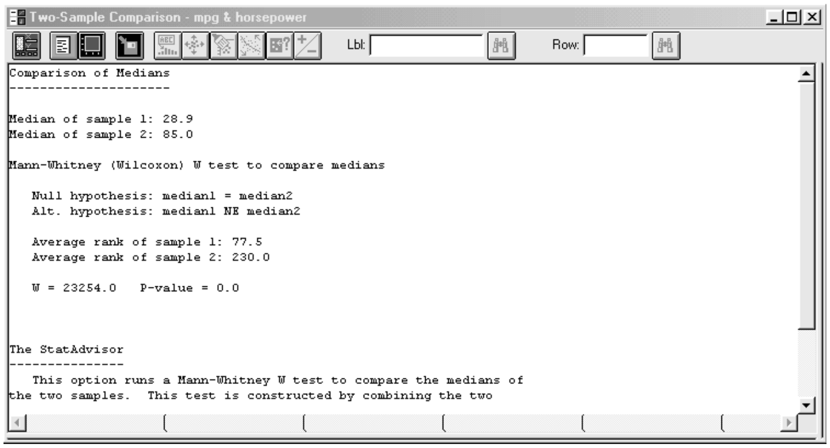
● Not Equal

○ Less Than

○ Greater Than

Alpha: 5 %

Figure 14-7. Comparison of Standard Deviations Options Dialog Box

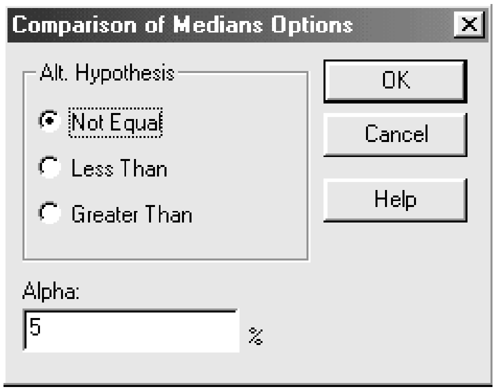


```
Comparison of Medians
---------------------

Median of sample 1: 28.9
Median of sample 2: 85.0

Mann-Whitney (Wilcoxon) W test to compare medians

   Null hypothesis: median1 = median2
   Alt. hypothesis: median1 NE median2

   Average rank of sample 1: 77.5
   Average rank of sample 2: 230.0

   W = 23254.0   P-value = 0.0



The StatAdvisor
---------------
   This option runs a Mann-Whitney W test to compare the medians of
the two samples.  This test is constructed by combining the two
```

Figure 14-8. Comparison of Medians

It is constructed by combining the two samples, sorting them from smallest to largest, then comparing the average ranks for the two samples of combined data. If the *p*-value is small (less than a desired alpha level, such as .05), there is a statistically significant difference between the medians at a given confidence level.

Use the *Comparison of Medians Options* dialog box to select an alternative hypothesis, and to enter a percentage for alpha (see Figure 14-9).



Comparison of Medians Options

Alt. Hypothesis

● Not Equal

○ Less Than

○ Greater Than

Alpha: 5 %

Figure 14-9.   Comparison of Medians Options Dialog Box

Kolmogorov-Smirnov Test

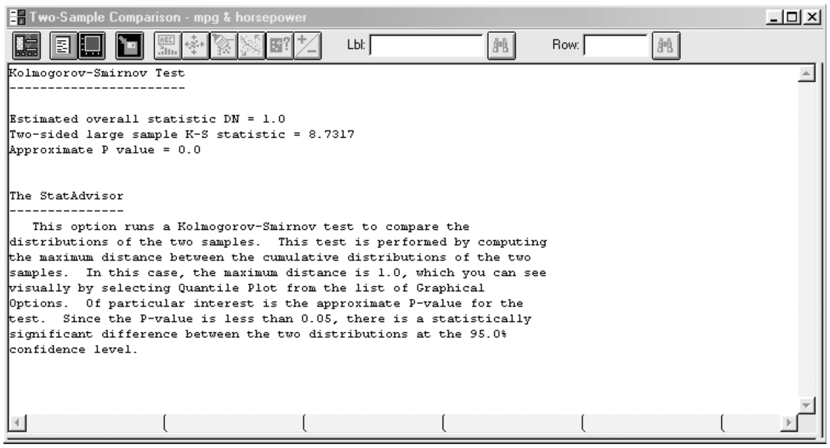
The Kolmogorov-Smirnov Test option computes the observed cumulative distributions for the two samples and the maximum positive, negative, and absolute differences (see Figure 14-10). If the distance is large enough, the *p*-value will be small (less than 0.05), indicating that there is a statistically significant difference between the two distributions at a given confidence level.

The method also tests the null hypothesis: two samples are from the same distribution.

Graphical Options

Frequency Histogram

The Frequency Histogram option creates a graph that displays the classes on the horizontal axis and the frequencies of the classes on the vertical axis. The frequency of each class is represented by a vertical bar whose height is equal to the frequency. The height of each bar is the number of observations that



```
Kolmogorov-Smirnov Test
-----------------------

Estimated overall statistic DN = 1.0
Two-sided large sample K-S statistic = 8.7317
Approximate P value = 0.0


The StatAdvisor
---------------
   This option runs a Kolmogorov-Smirnov test to compare the
distributions of the two samples.  This test is performed by computing
the maximum distance between the cumulative distributions of the two
samples.  In this case, the maximum distance is 1.0, which you can see
visually by selecting Quantile Plot from the list of Graphical
Options.  Of particular interest is the approximate P-value for the
test.  Since the P-value is less than 0.05, there is a statistically
significant difference between the two distributions at the 95.0%
confidence level.
```

Figure 14-10. Kolmogorov-Smirnov Test

fall within that interval. In the graph, two histograms are displayed; one for each sample, which is helpful when you need to visualize the shape of two distributions (see Figure 14-11).

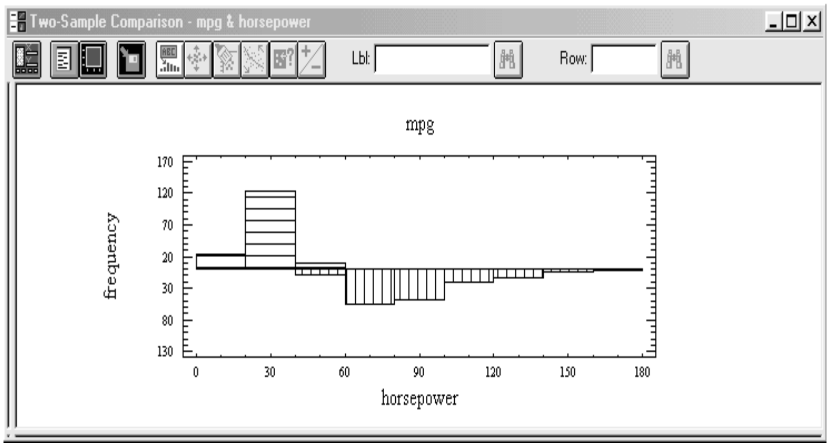
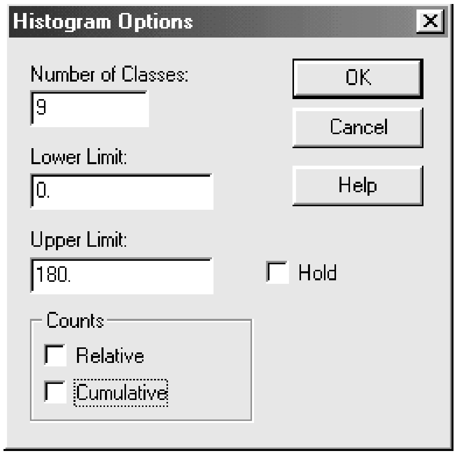


Figure 14-11. Frequency Histogram

In this analysis, the graph is divided horizontally; that is, the histogram for Sample 1 appears in the upper half of the plot, while the histogram for Sample 2 appears inverted in the lower half.

Use the *Histogram Options* dialog box to enter values for the number of classes, the Lower and Upper limits, and to indicate the type of data that will appear on the plot. You can also indicate if you want to retain the scaling on the plot even if you change the values in the Analysis dialog box (see Figure 14-12).



Number of Classes: 9

Lower Limit: 0.

Upper Limit: 180.

□ Hold

Counts

□ Relative

□ Cumulative

Figure 14-12. Histogram Options Dialog Box

Density Trace

The Density Trace option creates a graph that is essentially a smoothed histogram that shows the shape of each distribution (see Figure 14-13). The Density Trace provides a nonparametric estimate of the density function. The program estimates the density by counting (in a weighted manner) the number of observations in an interval of fixed length, which is moved through the data, then dividing that count by the width of the interval.

Unlike a histogram, a Density Trace uses overlapping intervals and a weight function to smooth the densities, resulting in a continuous line rather than a group of rectangles as in a histogram.

Use the *Density Trace Options* dialog box to indicate the method that will be used to shape the window that passes over the data, and to enter values for the interval width and the number of locations at which the traces will be calculated (see Figure 14-14).

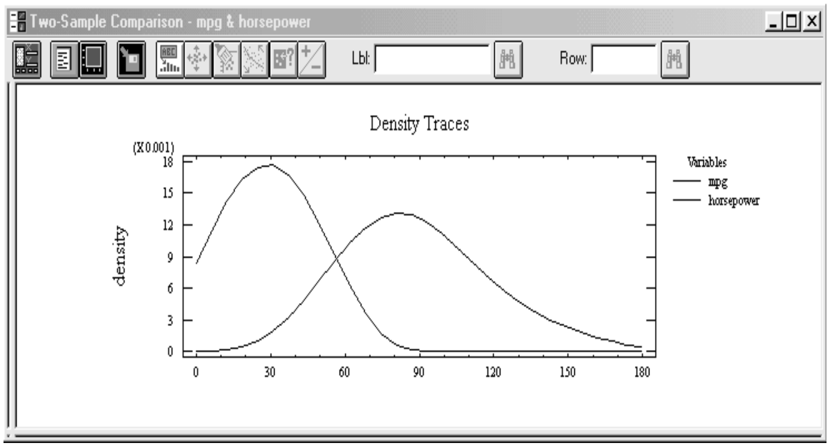
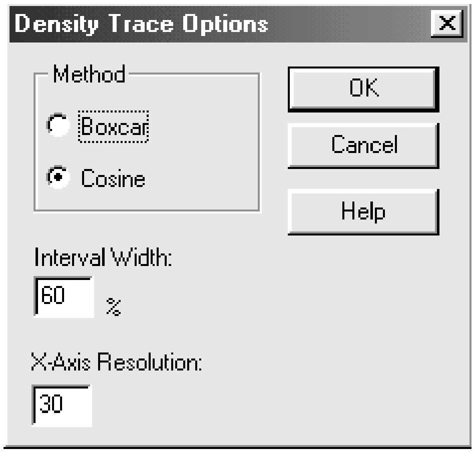


Figure 14-13. Density Trace



Method

○ Boxcar

● Cosine

Interval Width: 60 %

X-Axis Resolution: 30

Figure 14-14. Density Trace Options Dialog Box

Box-and-Whisker Plot

The Box-and-Whisker Plot option in this analysis creates two plots, one for each sample, which provide a way of summarizing a set of data measured on an interval scale. It is a type of graph that shows the shape of a distribution, its central value, and its variability (see Figure 14-15).

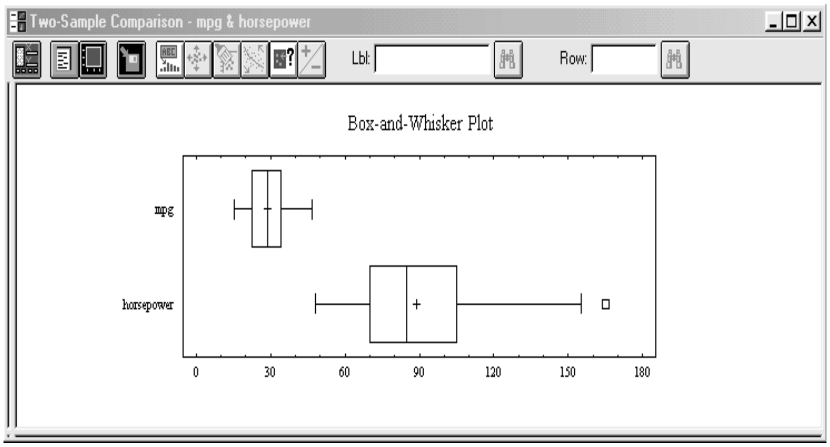


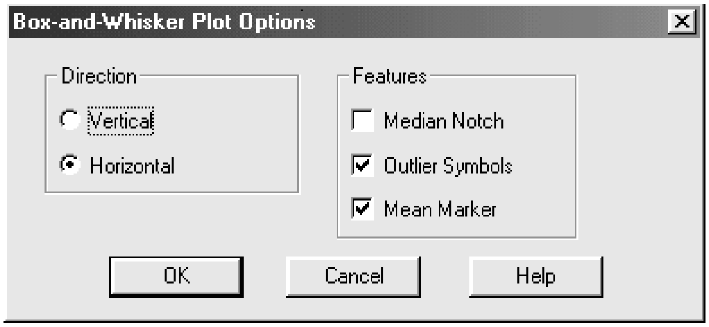
Figure 14-15. Box-and-Whisker Plot

The plot shows the most extreme values in the data (maximum and minimum values), the lower and upper quartiles, and the median. The rectangular part of the plot extends from the lower quartile to the upper quartile, and covers the center half of each sample. The center lines within each box show the location of the sample medians. Plus signs indicate the location of the sample means.

Whiskers extend from the box to the minimum and maximum values in each sample, except for the outside or far outside points, which are plotted separately. Outside points are points that lie more than 1.5 times the interquartile range above or below the box; they are shown as small squares. Far outside points are points that lie more than 3.0 times the interquartile range above or below the box; they are shown as small squares with plus signs through them.

This type of plot is helpful for indicating if a distribution is skewed and if there are any unusual observations (outliers). They are also helpful when large numbers of observations are involved and when you are comparing two or more datasets.

Use the *Box-and-Whisker Options* dialog box to indicate the direction of the plot and to select the features that will appear on it (see Figure 14-16).



Box-and-Whisker Plot Options

Direction

○ Vertical

● Horizontal

Features

□ Median Notch

☑ Outlier Symbols

☑ Mean Marker

Figure 14-16. Box-and-Whisker Plot Options Dialog Box

Quantile Plot

The Quantile Plot option creates a plot that shows the estimated quantiles for the two samples; that is, the Y-axis shows the fraction of data that are below a particular value (see Figure 14-17). Quantiles are a set of *cut points* that divide a sample of data into groups that contain (as far as possible), equal numbers of observations. Use the plot to compare the cumulative distributions for the two samples, or to estimate percentiles.

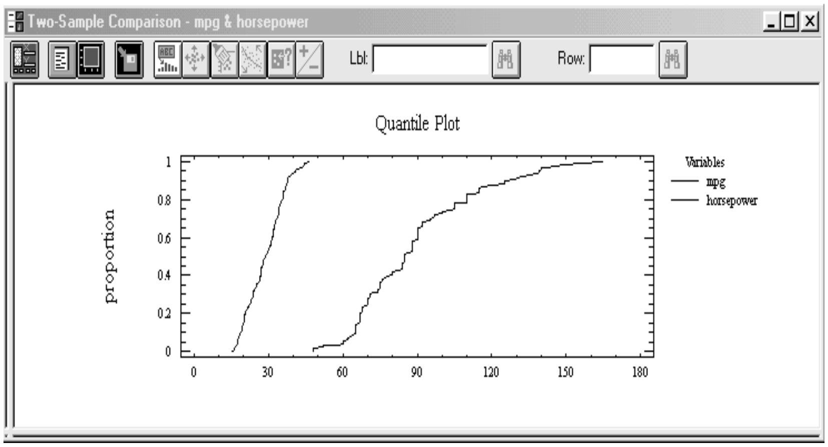


Figure 14-17. Quantile Plot

Quantile/Quantile Plot

The Quantile/Quantile Plot option produces a plot used to see if a given set of data follow some specified distribution. The distribution should be approximately linear if the specified distribution is the correct model.

The program constructs the plot using the theoretical cumulative distribution function, F(x) of the given model. The quantiles for the two samples are plotted versus each other (see Figure 14-18). If both samples come from the same distribution, the points should be close to the diagonal line. A general pattern of points on either side of the line indicates a difference between the two distributions.

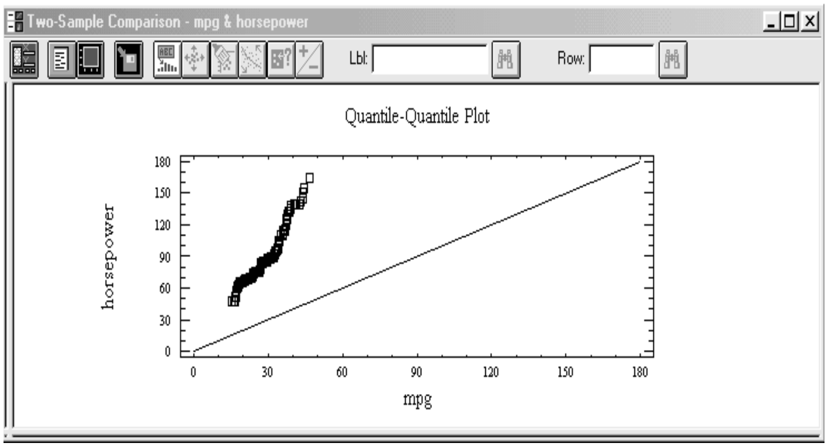


Figure 14-18. Quantile/Quantile Plot

References

Guttman, I., Wilks, S. S., and Hunter, J. S. 1982. *Introductory Engineering Statistics*, third edition. New York: Wiley.

Hollander, M. and Wolfe, D. A. 1973. *Nonparametric Statistical Methods*. New York: John Wiley and Sons, Inc.

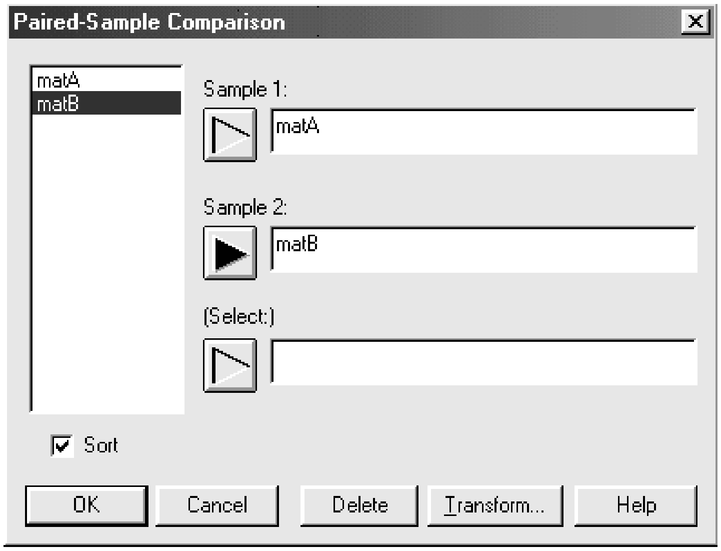
Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Using the Paired-Sample Comparison Analysis

The Paired-Sample Comparison Analysis parallels the Two-Sample Comparison Analysis except, it compares data collected in pairs; that is, you want to compare two sets of data because the observations in the samples are paired. For example, suppose you are analyzing data that contain information about the miles per gallon ratings achieved by the same automobile during highway driving and city driving. The data represent two different sets of values for each automobile: miles per gallon ratings for highway driving and miles per gallon ratings for city driving.

When data are from paired samples, you analyze it by calculating the difference between each value in each pair of observations. That is, you treat the paired differences as though they were a single sample from a population of differences.

To access the analysis, from the menus, choose: COMPARE... TWO SAMPLE... PAIRED-SAMPLE COMPARISON... (see Figure 14-19).



Paired-Sample Comparison

Sample 1: matA

Sample 2: matB

(Select:)

☑ Sort

Figure 14-19.   The Paired-Sample Comparison Analysis Dialog Box

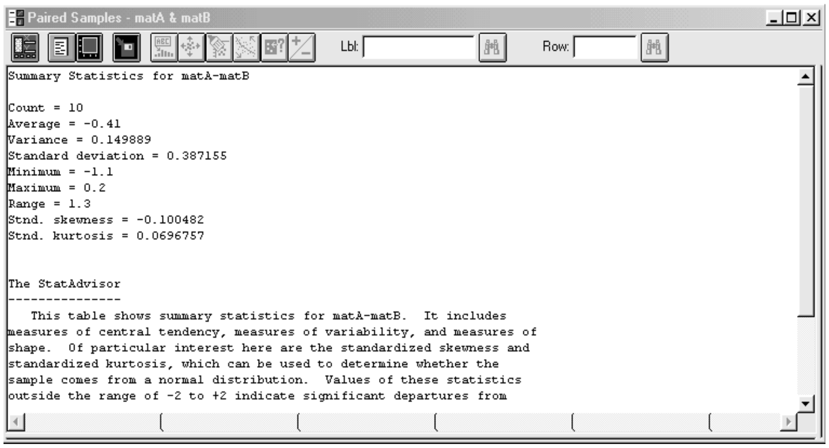
Tabular Options

Analysis Summary

The Analysis Summary option performs tests for significant differences between two data samples where the data are collected as pairs. The summary includes the results of tests that determine if the mean difference is equal to zero.

Summary Statistics

The Summary Statistics option creates summary statistics for two pairs of data. The statistical information includes measures of the central tendency, variability, and shape; as well as for the statistics you select, and includes the number of values in each variable (Count); the average, variance, and standard deviation; and the minimum and maximum values. It also includes values for the range, and standardized skewness and standardized kurtosis (see Figure 14-20).



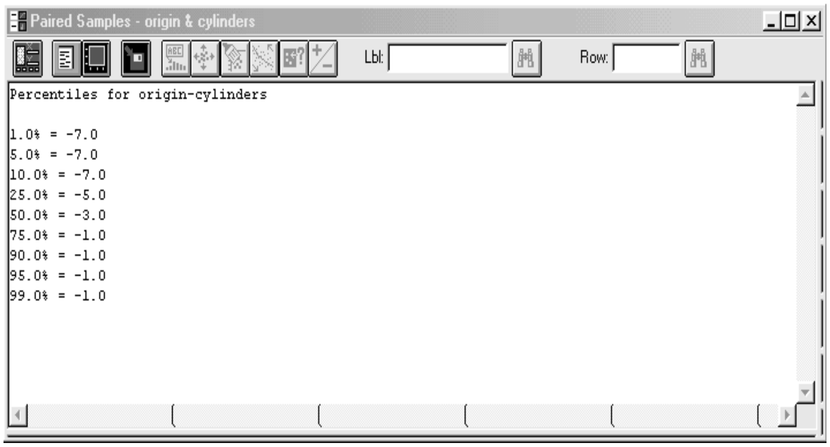
```
Summary Statistics for matA-matB

Count = 10
Average = -0.41
Variance = 0.149889
Standard deviation = 0.387155
Minimum = -1.1
Maximum = 0.2
Range = 1.3
Stnd. skewness = -0.100482
Stnd. kurtosis = 0.0696757


The StatAdvisor
---------------
   This table shows summary statistics for matA-matB.  It includes
measures of central tendency, measures of variability, and measures of
shape.  Of particular interest here are the standardized skewness and
standardized kurtosis, which can be used to determine whether the
sample comes from a normal distribution.  Values of these statistics
outside the range of -2 to +2 indicate significant departures from
```

Figure 14-20.   Summary Statistics

Use the *Summary Statistics Options* dialog box to select the statistics that will be calculated (see Figure 14-3 for an example of this dialog box).

Percentiles

The Percentiles option creates a table of percentages for the paired differences (see Figure 14-21). The results show the percentage of values that are equal to or less than a value you select. To see the percentiles in graphic form, use the Quantile Plot tabular option.



```
Percentiles for origin-cylinders

1.0% = -7.0
5.0% = -7.0
10.0% = -7.0
25.0% = -5.0
50.0% = -3.0
75.0% = -1.0
90.0% = -1.0
95.0% = -1.0
99.0% = -1.0
```

Figure 14-21. Percentiles

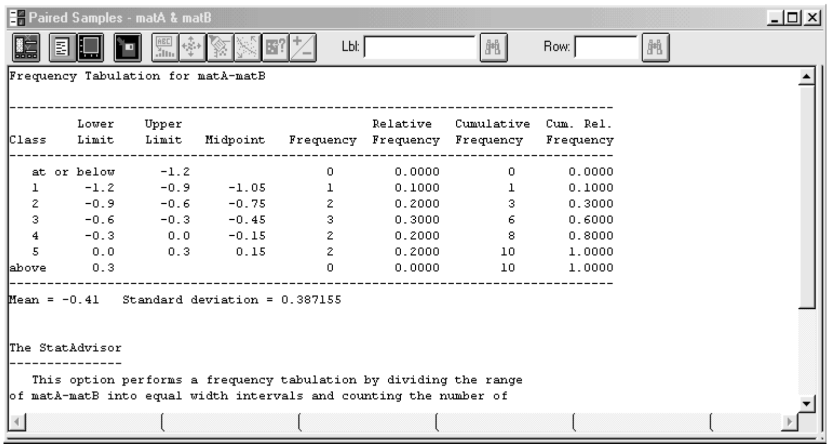
Use the *Percentiles Options* dialog box to enter values for other percentiles. If you do not want to calculate percentiles, enter 0.

Frequency Tabulation

The Frequency Tabulation option performs a frequency tabulation by dividing the range of the variables into intervals of equal width, then counting the number of values in each interval (see Figure 14-22).

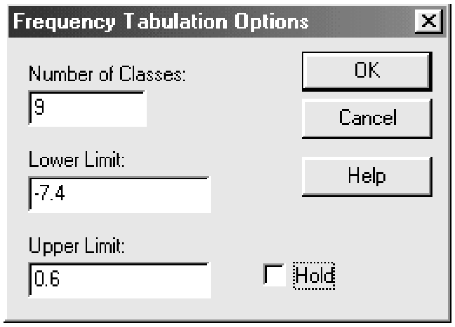
Frequencies are the number of values in each interval, while the relative frequencies are the proportions in each interval. The frequency table also displays values for the class, lower and upper limits, midpoint, cumulative frequencies, and the cumulative relative frequency.

Use the *Frequency Tabulation Options* dialog box to enter values for the number of classes into which the data will be grouped, as well as for the Lower limit for the first class and the Upper limit for the last class. You can also indicate if the scaling should be retained if you change values on the Analysis dialog box (see Figure 14-23).



```
Frequency Tabulation for matA-matB

--------------------------------------------------------------------------------
         Lower    Upper                          Relative   Cumulative  Cum. Rel.
Class    Limit    Limit   Midpoint   Frequency   Frequency  Frequency   Frequency
--------------------------------------------------------------------------------
   at or below     -1.2                  0        0.0000         0      0.0000
   1      -1.2     -0.9     -1.05        1        0.1000         1      0.1000
   2      -0.9     -0.6     -0.75        2        0.2000         3      0.3000
   3      -0.6     -0.3     -0.45        3        0.3000         6      0.6000
   4      -0.3      0.0     -0.15        2        0.2000         8      0.8000
   5       0.0      0.3      0.15        2        0.2000        10      1.0000
above      0.3                           0        0.0000        10      1.0000
--------------------------------------------------------------------------------
Mean = -0.41   Standard deviation = 0.387155


The StatAdvisor
---------------
   This option performs a frequency tabulation by dividing the range
of matA-matB into equal width intervals and counting the number of
```

Figure 14-22. Frequency Tabulation



Frequency Tabulation Options

Number of Classes: 9

Lower Limit: -7.4

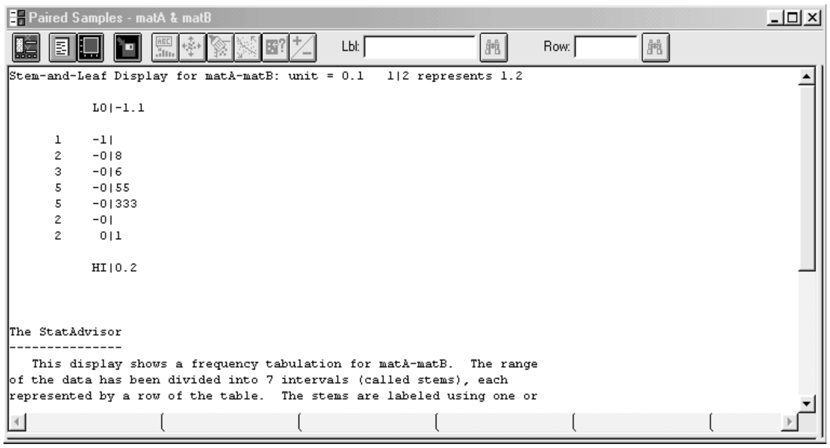
Upper Limit: 0.6

□ Hold

Figure 14-23. Frequency Tabulation Options Dialog Box

Stem-and-Leaf Display

The Stem-and-Leaf Display option creates a frequency tabulation for the two pairs of data (see Figure 14-24).



```
Stem-and-Leaf Display for matA-matB: unit = 0.1   1|2 represents 1.2

          LO|-1.1

     1    -1|
     2    -0|8
     3    -0|6
     5    -0|55
     5    -0|333
     2    -0|
     2     0|1

          HI|0.2



The StatAdvisor
---------------
   This display shows a frequency tabulation for matA-matB.  The range
of the data has been divided into 7 intervals (called stems), each
represented by a row of the table.  The stems are labeled using one or
```

Figure 14-24. Stem-and-Leaf Display

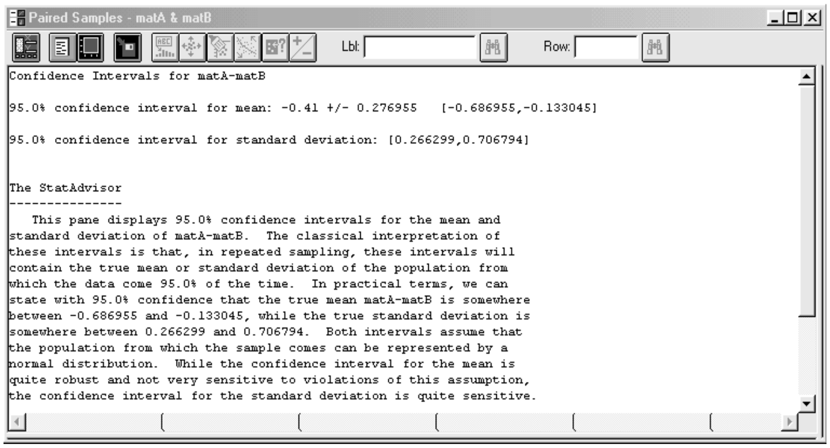
The program divides the range of the data into intervals (called *stems*), where each stem represents a row in the table. It then labels the stems, using one or more leading digits for the values that fall within that interval. A digit (called a *leaf*) to the right of the vertical line on the plot, represents the individual values in each row. The result is a histogram for which you can recover at least two significant digits for each value. Any points (called *outside* points) that appear a considerable distance away from most of the others, are placed on separate high and low stems. You can use the Box-and-Whisker Plot graphical option to see the outside or *outlier* points.

Use the *Stem-and-Leaf Display Options* dialog box to indicate if you want outliers placed on separate high or low stems. Flagged outliers is the default.

Confidence Intervals

The Confidence Intervals option displays the given confidence intervals for the mean and standard deviation of the paired differences (see Figure 14-25).

The classical interpretation is that in repeated sampling, the intervals contain the true mean or the standard deviation of the population from which the data originate a given percent of the time. The practical interpretation is shown by the results of this comparison. If the data are not normally distributed, the interval for the standard deviation is probably incorrect. The Normal Probability Plot graphical option can prove or disprove this assumption.



```
Confidence Intervals for matA-matB

95.0% confidence interval for mean: -0.41 +/- 0.276955   [-0.686955,-0.133045]

95.0% confidence interval for standard deviation: [0.266299,0.706794]


The StatAdvisor
---------------
   This pane displays 95.0% confidence intervals for the mean and
standard deviation of matA-matB.  The classical interpretation of
these intervals is that, in repeated sampling, these intervals will
contain the true mean or standard deviation of the population from
which the data come 95.0% of the time.  In practical terms, we can
state with 95.0% confidence that the true mean matA-matB is somewhere
between -0.686955 and -0.133045, while the true standard deviation is
somewhere between 0.266299 and 0.706794.  Both intervals assume that
the population from which the sample comes can be represented by a
normal distribution.  While the confidence interval for the mean is
quite robust and not very sensitive to violations of this assumption,
the confidence interval for the standard deviation is quite sensitive.
```

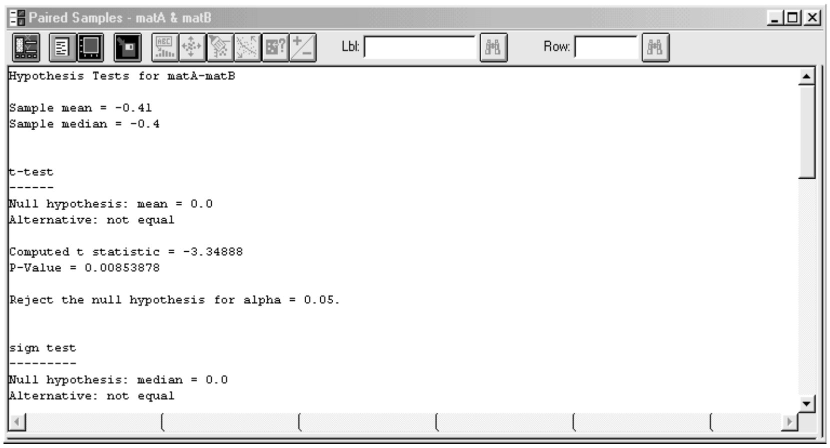
Figure 14-25.   Confidence Intervals

Use the *Confidence Intervals Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals for the mean and standard deviation. The default is 95 percent; other common percentages are 90 and 99.

Hypothesis Tests

The Hypothesis Tests option helps draw conclusions about population parameters based on results observed in a paired sample (see Figure 14-26). To formulate this type of test, some theory has usually been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proven. For example, claiming that a new drug is better than the current drug for treatment of the same symptoms. The option performs three tests for the null hypothesis: a *t*-test, a sign test, and a signed rank test.

- ***t*-Test**
  The *t*-test is a test of the null hypothesis that the mean equals a specific value versus the alternative hypothesis that the mean is not equal to a specific value. If the *p*-value is less than 0.05, the null hypothesis can be rejected at the 95 percent confidence level.



```
Hypothesis Tests for matA-matB

Sample mean = -0.41
Sample median = -0.4


t-test
------
Null hypothesis: mean = 0.0
Alternative: not equal

Computed t statistic = -3.34888
P-Value = 0.00853878

Reject the null hypothesis for alpha = 0.05.


sign test
---------
Null hypothesis: median = 0.0
Alternative: not equal
```

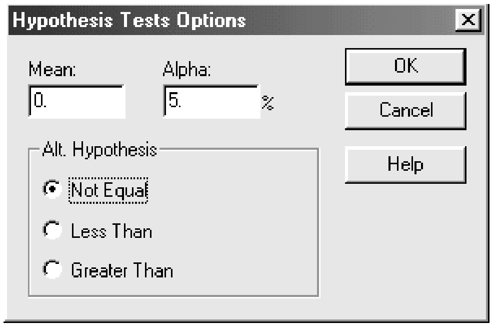
Figure 14-26. Hypothesis Tests

- **Sign Test**
  The Sign Test is a test of the null hypothesis about the population median that the mean equals a specific value versus the alternative hypothesis that the mean is not equal to a specific value, and involves the use of matched pairs. For example, before and after data, in which case it tests for a median difference of zero. The procedure is based on counting the number of values above and below the hypothesized median. If the *p*-value is less than 0.05, the null hypothesis can be rejected at the 95 percent confidence level. This test is less sensitive to the presence of outliers, but is somewhat less powerful than the *t*-test, if the data are all from a single normal distribution.

- **Signed Rank Test**
  The Signed Rank Test is a test of the null hypothesis about the population median that the mean equals a specific value versus the alternative hypothesis that the mean is not equal to a specific value, and involves the use of matched pairs. This test does not require the assumption that the population is normally distributed. In many applications it is used when the normality assumption is questionable. It is a more powerful alternative to the Sign test, but does assume that the population probability distribution is symmetric.

  The procedure is based on comparing the average ranks of values above and below the hypothesized median. If the *p*-value is less than 0.05, the null hypothesis can be rejected at the 95 percent confidence level. This test is less sensitive to the presence of outliers, but is somewhat less powerful than the *t*-test, if the data are all from a single normal distribution.

Use the *Hypothesis Tests Options* dialog box to enter values for the hypothetical mean of the paired differences and for the alpha level. You can also choose the type of alternative hypothesis test that will be performed (see Figure 14-27).



Mean: 0.

Alpha: 5. %

Alt. Hypothesis

● Not Equal

○ Less Than

○ Greater Than

Figure 14-27. Hypothesis Tests Options Dialog Box

Graphical Options

Scatterplot

The Scatterplot option creates a plot of the observed paired differences of the selected variables. The points are plotted along a single axis as point symbols with no connecting lines (see Figure 14-28). In the figure, the points have been jittered to prevent overplotting. Jittering adds a small random offset to each point before plotting it. To reduce the amount of overlapping, use the Jittering button on the Analysis toolbar.

Box-and-Whisker Plot

The Box-and-Whisker Plot option creates plot of the paired differences (see Figure 14-29). This type of plot is good to use because it shows various features in the sample of data. The program divides the paired differences into four areas of equal frequency (*quartiles*).

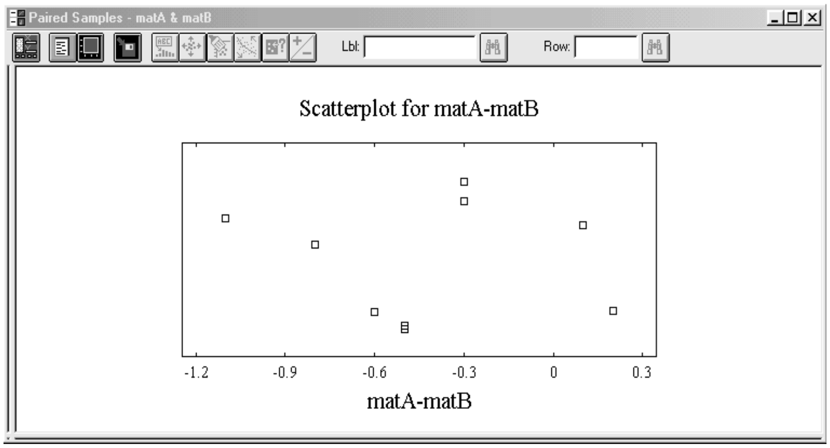


Figure 14-28. *Scatterplot*

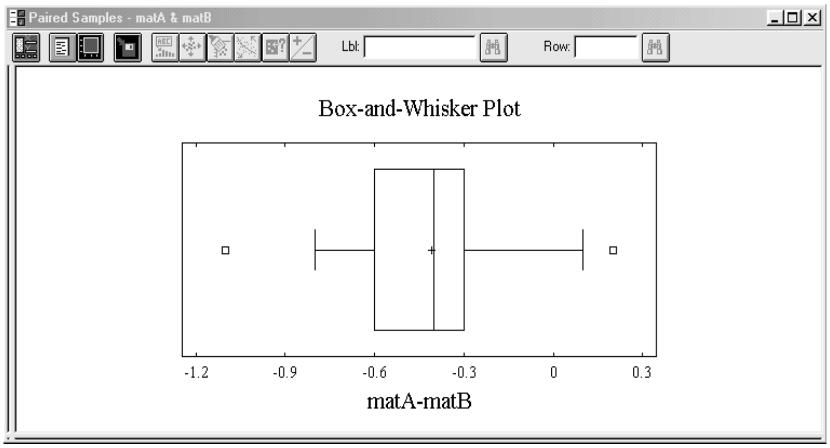


Figure 14-29. *Box-and-Whisker Plot*

A rectangular box extends from the lower quartile to the upper quartile, covering the center half of the sample. The center line within the box shows the location of the sample median. A plus (+) sign indicates the location of the sample mean. The whiskers extend from the box to the minimum and maximum values in the sample, except for any outside or far outside points (*outliers*), which will be plotted separately. Outside points are points that lie more than 1.5 times the interquartile range above or below the box and are shown as small squares. Far outside points are points that lie more than 3.0 times the interquartile range above or below the box, and are shown as small squares with a plus (+) sign through them.

Use the *Box-and-Whisker Options* dialog box to indicate the direction of the plot and to select the features that will appear on it (see Figure 14-16 for an example of this dialog box).

Frequency Histogram

The Frequency Histogram option performs a frequency tabulation by dividing the range of the variables into intervals of equal width, then counting the number of values in each interval (see Figure 14-30). Frequencies are the number of values in each interval, while the relative frequencies are the proportions in each interval.

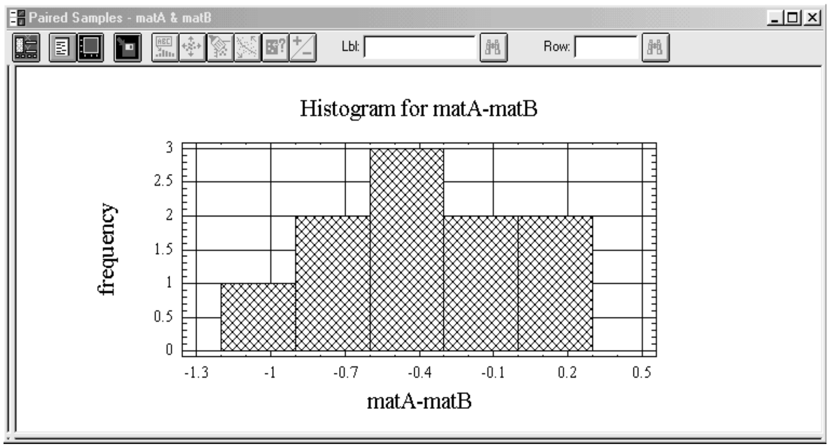
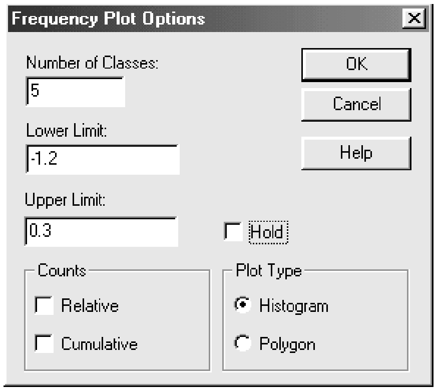


Figure 14-30. Frequency Tabulation

Use the *Frequency Plot Options* dialog box to enter values for the number of classes into which the data will be grouped, as well as for the Lower and Upper limits (see Figure 14-31). You can also indicate if the scaling should be retained if you change values on the Analysis dialog box; and to indicate the type of counts that will be included, and the type of plot that will be created.



Number of Classes: 5

Lower Limit: -1.2

Upper Limit: 0.3

□ Hold

Counts

□ Relative

□ Cumulative

Plot Type

● Histogram

○ Polygon

Figure 14-31. Frequency Plot Options Dialog Box

Quantile Plot

The Quantile Plot option creates a plot that shows the estimated quantiles for the paired samples; that is, the Y-axis shows the fraction of data that are below a particular value (see Figure 14-32). Quantiles are a set of *cut points* that divide a sample of data into groups that contain (as far as possible), equal numbers of observations. Use the plot to compare the cumulative distributions for the paired samples, or to estimate percentiles.

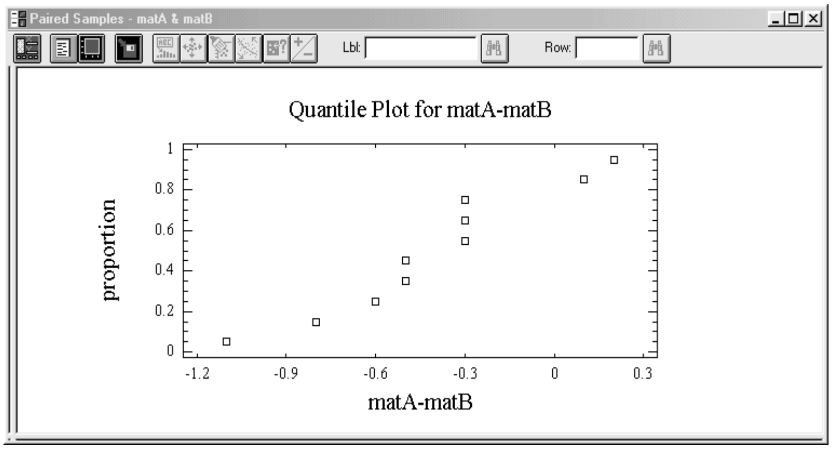


Figure 14-32. Quantile Plot

Normal Probability Plot

The Normal Probability Plot option creates a plot using values that are sorted from smallest to largest (see Figure 14-33). If the data come from a normal distribution, the points should fall approximately along a straight line. To help determine the closeness of the points to a straight line, a reference line is superimposed on the plot. The reference line passes through the median with slope determined by the interquartile range. Points showing significant curvature indicate skewness in the data.

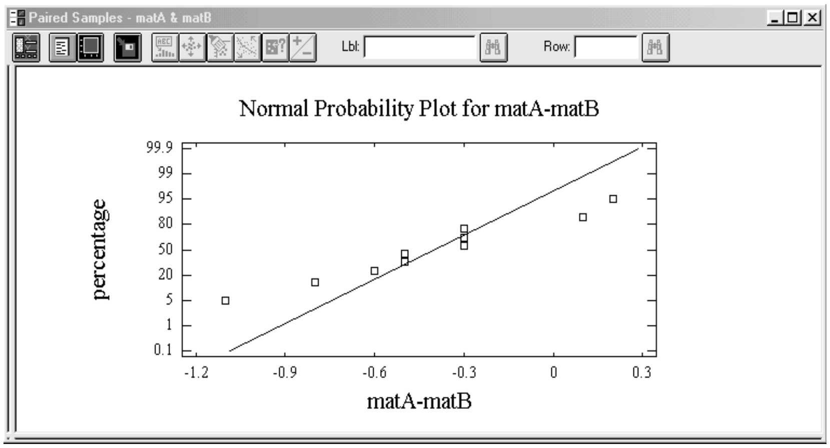
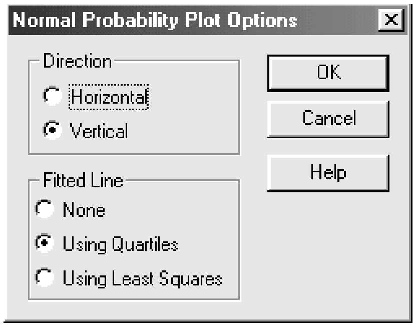


Figure 14-33. Normal Probability Plot

Use the *Normal Probability Plot Options* dialog box to indicate if the plot will display in a horizontal or vertical direction; and to indicate if a fitted line will appear on the plot; if so, whether quartiles or least squares will be the method used to fit the line (see Figure 14-34).

Density Trace

The Density Trace option creates a graph that is essentially a smoothed histogram that shows the shape of each distribution (see Figure 14-35). The Density Trace provides a nonparametric estimate of the density function. The program estimates the density by counting (in a weighted manner) the number of observations in an interval of fixed length, which is moved through the data, then dividing that count by the width of the interval.


Normal Probability Plot Options

Direction
○ Horizontal
● Vertical

Fitted Line
○ None
● Using Quartiles
○ Using Least Squares

Figure 14-34. Normal Probability Plot Options Dialog Box

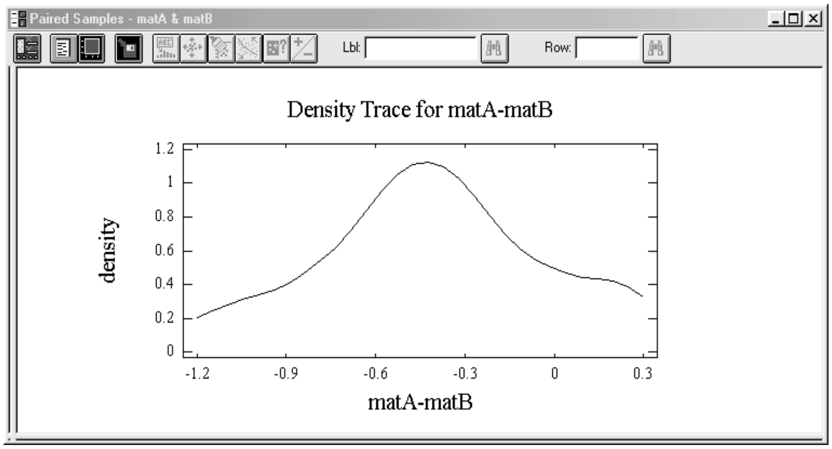


Figure 14-35. Density Trace

Unlike a histogram, a Density Trace uses overlapping intervals and a weight function to smooth the densities, resulting in a continuous line rather than a group of rectangles as in a histogram.

Use the *Density Trace Options* dialog box to indicate the method that will be used to shape the window that passes over the data, and to enter values for the interval width and the number of locations at which the traces will be calculated (see Figure 14-14 for an example of this dialog box).

Symmetry Plot

The Symmetry Plot option creates a plot that is helpful in determining the symmetry of the data by first sorting the values from smallest to largest, then plotting the points that are immediately to the left and right of median, which shows their respective distance from the median (see Figure 14-36). The process is repeated for the pair of points second closest to the median, third closest, and so on.

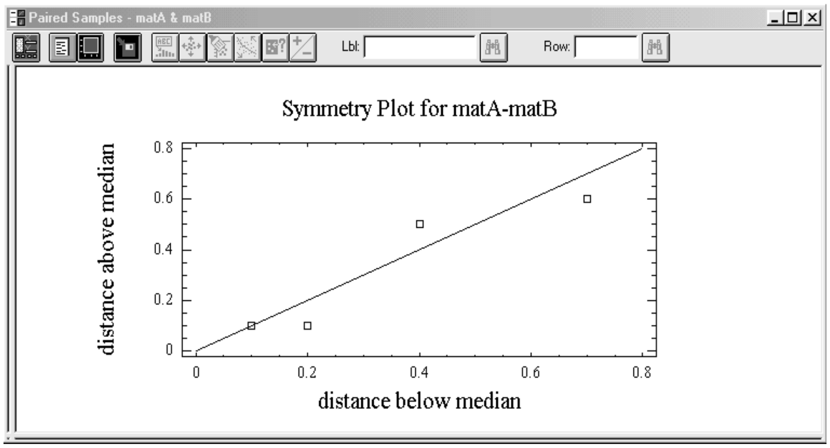


Figure 14-36. Symmetry Plot

If the distribution is symmetric, the points will lie close to the diagonal reference line. If the distribution is positively skewed the points will deviate above the line. If the distribution is negatively skewed, the points will deviate below the line.

Saving the Results

The Save Results Options dialog box allows you to select the results you want to save. There are six selections: Summary Statistics, Percentiles, Frequencies, Cumulative Frequencies, Relative Frequencies, and Cumulative Relative Frequencies.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. Belmont, California: Wadsworth International Group.

Frigge, M., Hoagland, D. C., and Iglewicz, B. 1989. "Some Implementations of the Boxplot," *American Statistician*, **43**:50-54.

Guttman, I., Wilks, S. S., and Hunter, J. S. 1982. *Introductory Engineering Statistics*, third edition. New York: Wiley.

Lapin, L. L. 1987. *Statistics for Modern Business Decisions*. Orlando, Florida: Harcourt Brace Jovanovich, Inc.

McGill, R., Tukey, J. W., and Larsen, W. A. 1978. "Variation of Box Plots," *American Statistician*, **32**:12-16.

Snedecor, G. W. and Cochran, W. G. 1967. *Statistical Methods*, sixth edition. Ames, Iowa: Iowa State University Press.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison Wesley.

Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Belmont, California: Duxbury Press.

Using the Hypothesis Tests (Compare) Analysis

Hypothesis tests are based on a sample of data to determine which of two different states is true. The two states are commonly called the *null hypothesis* and the *alternative hypothesis*. Hypothesis tests are a classical approach to assessing the statistical significance of findings, a technique that involves comparing empirically observed sample findings with theoretically expected findings — expected if the null hypothesis is true. This comparison allows you to compute the probability that the observed outcomes might be due to chance alone.

Hypothesis tests are a classical approach for assessing the statistical significance of findings. Basically, the technique involves comparing empirically observed sample findings with theoretically expected findings; expected if the null hypothesis is true. This comparison allows you to compute the probability that the observed outcomes might be due to chance alone. The data you use in this analysis should be sample statistics and sample sizes rather than raw data.

The Hypothesis Tests (Compare) Analysis in STATGRAPHICS *Plus* tests hypotheses about the sample means and variances for two or more random samples. *See Chapter 13, Performing Hypothesis Tests and Determining Sample Size, for a discussion about hypothesis tests for a single random sample.*

The analysis allows you to test four types of parameters: Normal Mean, Normal Sigma, Binomial Proportions, and Poisson Rate. In addition, you can control for error tolerance, power, and sample size; set confidence intervals; and choose from three alternative hypotheses: not equal, less than, or greater than.

To access the analysis, from the menus, choose: COMPARE... TWO SAMPLES... HYPOTHESIS TESTS... (see Figure 14-37).

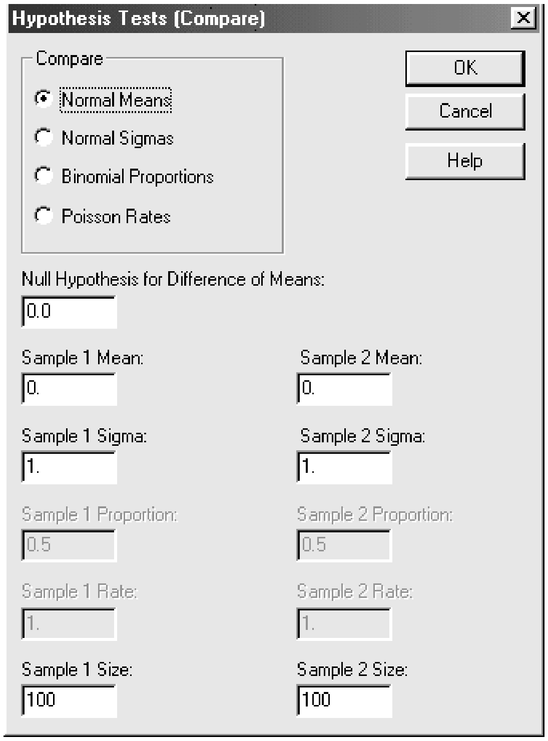
Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the estimated test statistics for the parameter you select.

- **Analysis Summary for the Normal Means Option**
  The analysis shows the results of performing a hypothesis test about the difference between the means (mu1/mu2) of two samples from normal distributions. Information in the lower portion of the text pane displays the difference between means for the null hypothesis, the name of the alternative hypothesis, the computed *t*-statistic, the *p*-value, and the recommendation — either reject or do not reject the null hypothesis.

- **Analysis Summary for the Normal Sigmas Option**
  The analysis shows the results of performing a hypothesis test about the ratio of the standard deviations (sigma1/sigma2) of two samples from normal distributions. Information in the lower portion of the text pane



Hypothesis Tests (Compare)

Compare

● Normal Means
○ Normal Sigmas
○ Binomial Proportions
○ Poisson Rates

Null Hypothesis for Difference of Means: 0.0

Sample 1 Mean: 0.
Sample 2 Mean: 0.

Sample 1 Sigma: 1.
Sample 2 Sigma: 1.

Sample 1 Proportion: 0.5
Sample 2 Proportion: 0.5

Sample 1 Rate: 1.
Sample 2 Rate: 1.

Sample 1 Size: 100
Sample 2 Size: 100

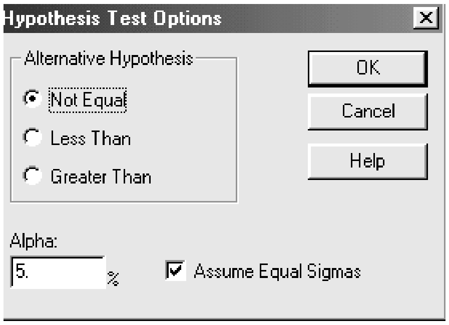
Figure 14-37. Hypothesis Tests (Compare) Analysis Dialog Box

displays the ratio of variances for the null hypothesis, the name of the alternative hypothesis test, the computed F-statistic, the *p*-value, and the recommendation — either reject or do not reject the null hypothesis for the given alpha.

- **Analysis Summary for the Binomial Proportions Option**
  The analysis shows the results of performing a hypothesis test about the difference between the proportions (theta1/theta2) of two samples from binomial distributions. Information in the lower portion of the text pane displays the difference between proportions for the null hypothesis, the name of the alternative hypothesis test, the computed *z*-statistic, the *p*-value, and the recommendation — either reject or do not reject the null hypothesis for the given alpha.

- **Analysis Summary for the Poisson Rates Option**
  The analysis shows the results of performing a hypothesis test about the difference between the rate parameters (lambda1/lambda2) of two Poisson distributions. Information in the lower portion of the text pane displays the difference between rates for the null hypothesis, the name of the alternative hypothesis test, the computed *z*-statistic, the *p*-value, and the recommendation — either reject or do not reject the null hypothesis for the given alpha.

Use the *Hypothesis Tests Options* dialog box to select an alternative hypothesis, enter a value for the confidence level, and indicate if equal standard deviations should be assumed for testing normal means (see Figure 14-38)



Hypothesis Test Options

Alternative Hypothesis

● Not Equal

○ Less Than

○ Greater Than

Alpha: 5. %

☑ Assume Equal Sigmas

Figure 14-38. Hypothesis Tests Options Dialog Box

Graphical Options

Power Curve

The Power Curve option displays a Power Curve for the test you are currently using.

- **Power Curve for Normal Means**
  The plot shows the power versus the true difference between the means (see Figure 14-39).

- **Power Curve for Normal Sigmas**
  The plot shows the power versus the true ratio of the variances.

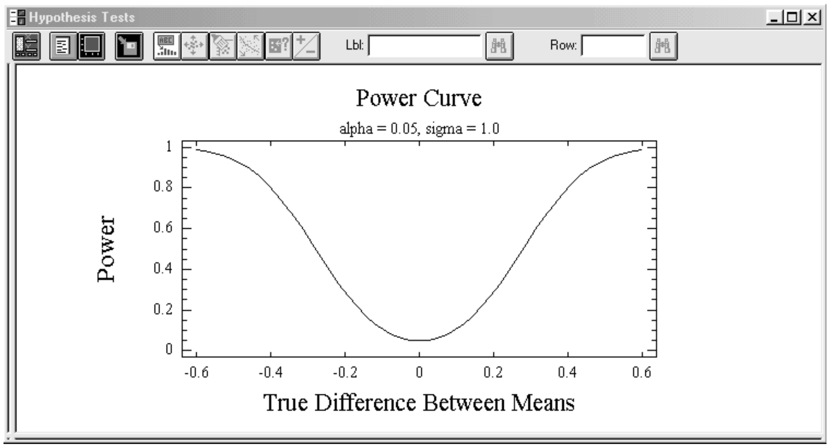


Figure 14-39. Power Curve for Normal Means

- **Power Curve for Binomial Proportions**
  The plot shows the power versus the true difference between the proportions.

- **Power Curve for Poisson Rates**
  The plot shows the power versus the true difference between the rates.

Use the *Power Curve Options* dialog box to enter a value for the assumed standard deviation/mean/proportion/mean rate, depending on the selected parameter.

References

Guttman, I., Wilks, S. S., and Hunter, J. S. 1982. *Introductory Engineering Statistics*, third edition. New York: Wiley.

Snedecor, G. W. and Cochran, W. G. 1967. *Statistical Methods*, sixth edition. Ames, Iowa: Iowa State University Press.

Vogt, P. 1993. *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences.* Newbury Park, California: Sage Publications, Inc.

Using the Sample-Size Determination (Compare) Analysis

In many statistical procedures, particularly in the planning stages and before data have been collected, it is important to determine the size of the sample required to adequately address the objectives of the study. Many research questions involve means or proportions where you must either

- estimate a population mean, proportion, or difference with a certain degree of accuracy, or
- test the statistical significance of the difference, either between a sample mean or proportion and some hypothesized value, or between means or proportions from two different samples.

An appropriate sample size should provide the degree of accuracy you need to estimate population means, proportions, or differences. It should also allow you to control the risk of reaching incorrect conclusions when you test for statistical significance.

When the estimates of means and proportions come from large samples, there is less sample-to-sample variation than in estimates that come from smaller samples. Large-sample estimates tend to be more accurate than estimates from smaller samples.

One factor that affects sample size requirements is the degree of accuracy you want to achieve — the maximum tolerable difference between the population parameter you are estimating and its true value.

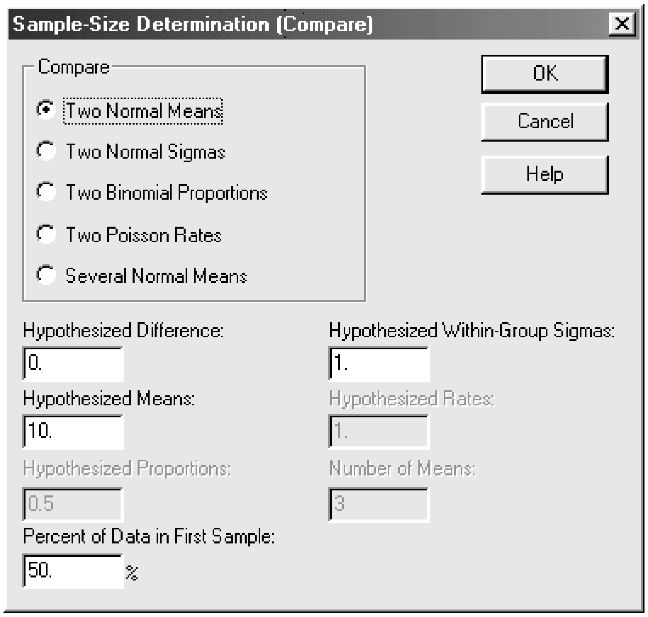
The degree of accuracy attained by any sample size also depends on the spread (variability) of the variable you are testing. Variability is typically measured by the variance or standard deviation. In general, the smaller the variance, the smaller the number of subjects necessary to achieve the degree of accuracy you desire. It is better to err on the "high" side because this provides an estimated sample size that has, at least, the accuracy you desire.

Another factor that affects sample size is the degree of confidence that will actually achieve accuracy. Except in trivial cases, there is no sample size, short of a census, that allows you to be certain of achieving the accuracy you want.

The Sample-Size Determination (Compare) Analysis in STATGRAPHICS *Plus* has two purposes: to help you determine the number of observations required to provide sufficiently powerful estimates; and to analyze the power curve for samples you have already drawn.

You can calculate sample sizes for testing five types of parameters: Two Normal Means, Two Normal Sigmas, Two Binomial Proportions, Two Poisson Rates, and Several Normal Means. In addition, you can control for error tolerance or power; set the confidence interval; and choose from three alternative hypotheses: not equal, less than, or greater than.

To access the analysis, from the menus, choose: COMPARE... TWO SAMPLES... SAMPLE SIZE DETERMINATION... (see Figure 14-40).



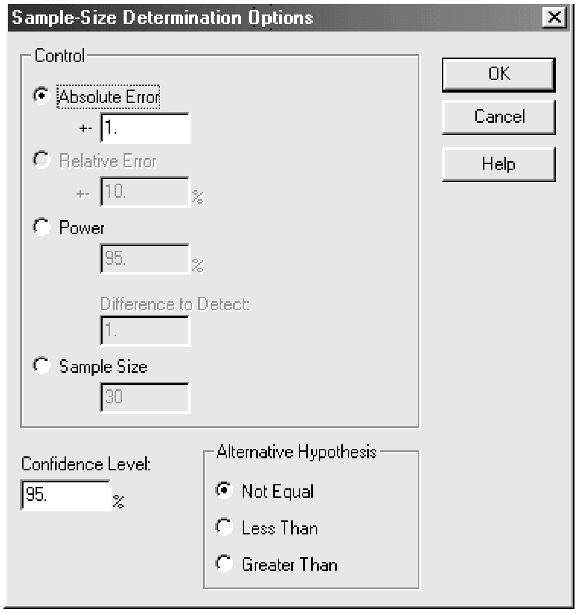
Sample-Size Determination (Compare)

Compare
● Two Normal Means
○ Two Normal Sigmas
○ Two Binomial Proportions
○ Two Poisson Rates
○ Several Normal Means

Hypothesized Difference: 0.
Hypothesized Within-Group Sigmas: 1.
Hypothesized Means: 10.
Hypothesized Rates: 1.
Hypothesized Proportions: 0.5
Number of Means: 3
Percent of Data in First Sample: 50. %

Figure 14-40. Sample-Size Determination (Compare) Analysis Dialog Box

Use the *Sample-Size Determination Options* dialog box to select the type of control you want to use and its precision; to enter a value for the percentage that will be used to calculate the confidence intervals; and to select the form of the alternative hypothesis (see Figure 14-41).



Sample-Size Determination Options

Control

● Absolute Error

+- 1.

○ Relative Error

+- 10. %

○ Power

95. %

Difference to Detect:

1.

○ Sample Size

30

Confidence Level:

95. %

Alternative Hypothesis

● Not Equal

○ Less Than

○ Greater Than

Figure 14-41. Sample-Size Determination Option Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis for the parameter you are testing.

- **Analysis Summary for Two Normal Means**
  The summary displays the results for determining the sample size required from each sample when you are comparing the difference between two normal means.

- **Analysis Summary for Two Normal Sigmas**
  The summary displays the results for determining the sample size required from each sample when you are comparing the standard deviations for two normal distributions.

- **Analysis Summary for Two Binomial Proportions**
  The summary displays the results for determining the sample size required from each sample when you are comparing the proportions for two binomial distributions.

- **Analysis Summary for Two Poisson Rates**
  The summary displays the results for determining the sample size required when you are comparing the rate parameters for two Poisson distributions.

- **Analysis Summary for Several Normal Means**
  This summary displays the results for determining the sample size required when you are comparing the difference between several normal means.

Graphical Options

Power Curve

The Power Curve option creates a graph of the power versus the true value of the parameter you are using (see Figure 14-42).

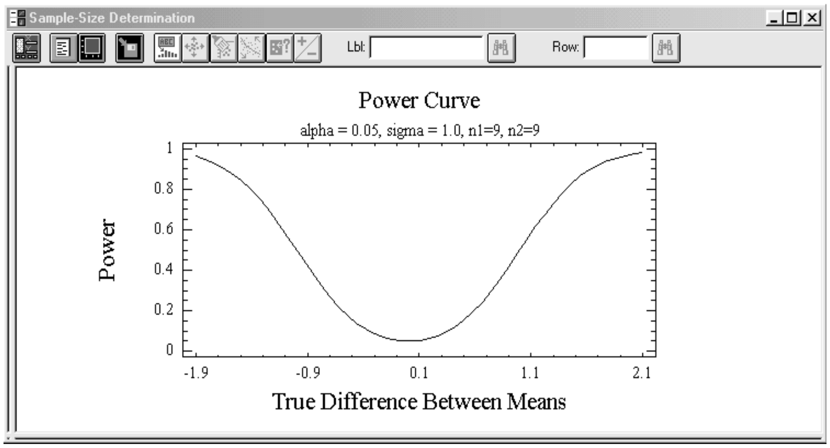

Sample-Size Determination
Lbl:
Row:
Power Curve
alpha = 0.05, sigma = 1.0, n1=9, n2=9
Power
1
0.8
0.6
0.4
0.2
0
-1.9
-0.9
0.1
1.1
2.1
True Difference Between Means

Figure 14-42. Power Curve

- **Power Curve for Two Normal Means**
  The plot displays the power of the hypothesis test performed. Power is defined as the probability that this statistical test would reject the null hypothesis as a function of the true population mu1/mu2.

- **Power Curve for Two Normal Sigmas**
  The plot displays the power of the hypothesis test performed. Power is defined as the probability that this statistical test would reject the null hypothesis as a function of the true population sigma1/sigma2.

- **Power Curve for Two Binomial Proportions**
  The plot displays the power of the hypothesis test performed. Power is defined as the probability that this statistical test would reject the null hypothesis as a function of the true population theta1/theta2

- **Power Curve for Two Poisson Rates**
  The plot displays the power of the hypothesis test performed. Power is defined as the probability that this statistical test would reject the null hypothesis as a function of the true population lambda1/lambda2.

- **Power Curve for Several Normal Means**
  The plot displays the power of the hypothesis test performed. Power is defined as the probability that this statistical test would reject the null hypothesis as a function of the true population mu1/mu2.

References

Desu, M. M. and Raghavarao, D. 1990. *Sample Size Methodology*. In Lieberman, G. J. and

Olkin, I., eds., *Statistical Modeling and Decision Science*. San Diego, California: Academic Press.

Haaland, P. 1989. *Experimental Design in Biotechnology*. New York: Marcel Dekker.

Hays, W. L. 1981. *Statistics,* third edition. New York: Holt, Rinehart, and Winston.

15 Comparing Multiple Samples

This chapter includes analyses for comparing two or more means (Multiple-Sample Comparison), two or more proportions (Comparison of Proportions), and two or more Poisson counts (Comparison of Counts). The analyses differ in how they adjust the observed significance level.

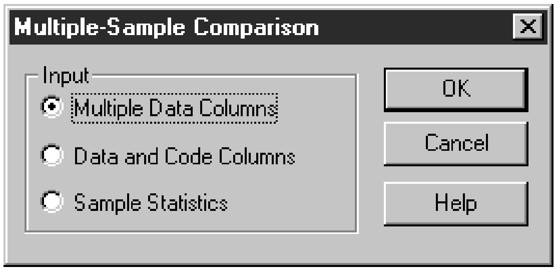
Using the Multiple-Sample Comparison Analysis

The Multiple-Sample Comparison Analysis compares two or more samples to test the probability that when the null hypothesis is true, at least one of the observed significance levels will be less than a specified number. The more comparisons you make, the more likely it is that one or more pairs will be statistically different, even if all the population means are equal. Special tests in the analysis determine the means that are different, and/or the means that are the smallest or largest.

To access the analysis, from the menus, choose: COMPARE... MULTIPLE SAMPLES... MULTIPLE-SAMPLE COMPARISON... .

Choose the type of input data from three options: Multiple Data Columns, Data and Code Columns, and Sample Statistics. Then click OK to display the appropriate dialog box (see Figure 15-1).

Depending on which input selection you choose, one of 3 input dialog boxes appears. See Figure 15-2 for an example of an input dialog box for multiple data columns.



Multiple-Sample Comparison

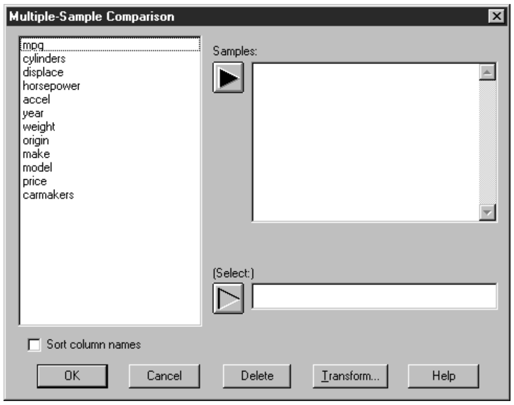
Input

● Multiple Data Columns

○ Data and Code Columns

○ Sample Statistics

Figure 15-1. Multiple-Sample Comparison Analysis Dialog Box



Multiple-Sample Comparison

mpg
cylinders
displace
horsepower
accel
year
weight
origin
make
model
price
carmakers

Samples:

(Select:)

□ Sort column names

Figure 15-2. The Multiple-Sample Comparison Input Dialog Box

Tabular Options

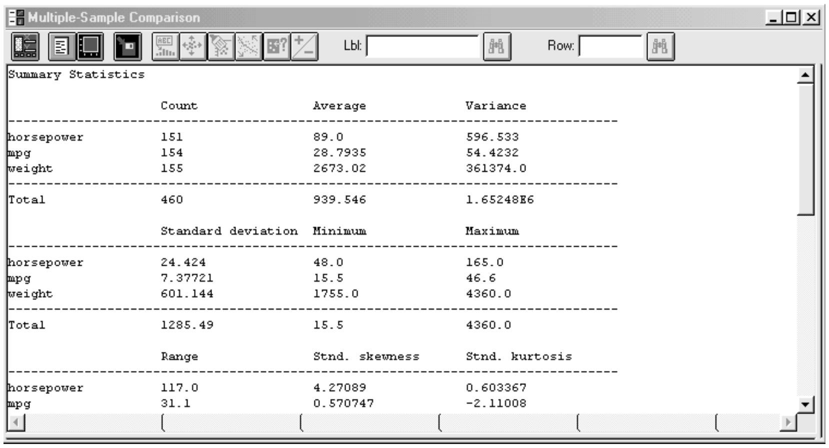
Analysis Summary

The Analysis Summary option creates a summary of the analysis, which shows the names of the variables in each sample, as well as the number and range of the values.

Summary Statistics

The Summary Statistics option creates statistical information about the statistics you choose. It includes the number of values in each variable (the

Count); the average, variance, and standard deviation; and the minimum and maximum values. It also includes values for the range, standardized skewness, and standardized kurtosis (see Figure 15-3).



```
Summary Statistics

                    Count              Average            Variance
------------------------------------------------------------------------------------
horsepower          151                89.0               596.533
mpg                 154                28.7935            54.4232
weight              155                2673.02            361374.0
------------------------------------------------------------------------------------
Total               460                939.546            1.65248E6

                    Standard deviation Minimum            Maximum
------------------------------------------------------------------------------------
horsepower          24.424             48.0               165.0
mpg                 7.37721            15.5               46.6
weight              601.144            1755.0             4360.0
------------------------------------------------------------------------------------
Total               1285.49            15.5               4360.0

                    Range              Stnd. skewness     Stnd. kurtosis
------------------------------------------------------------------------------------
horsepower          117.0              4.27089            0.603367
mpg                 31.1               0.570747           -2.11008
```

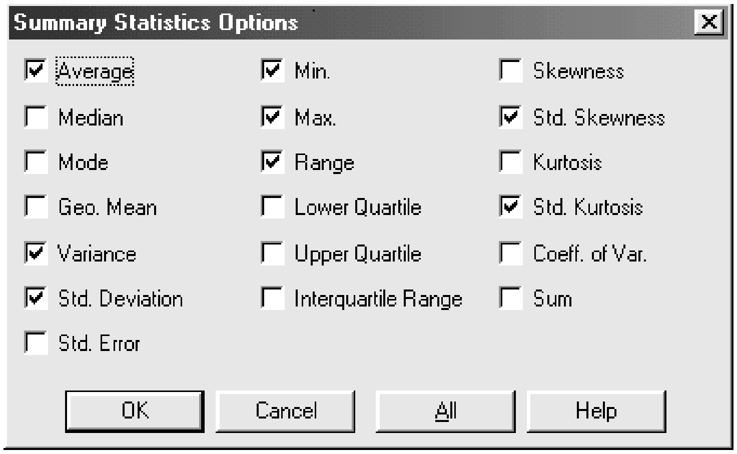
Figure 15-3. Summary Statistics

The values for the standard skewness and standardized kurtosis are of particular interest because they help determine if the samples are from normal distributions. Values outside the range of -2 to +2 indicate a significant departure from normality, which tends to invalidate the tests that compare the standard deviations.

Use the *Summary Statistics Options* dialog box to choose the statistics that will be calculated (see Figure 15-4).

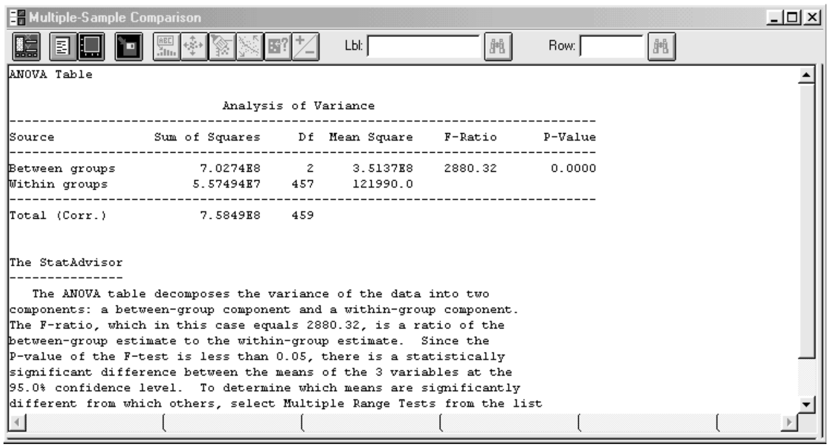
ANOVA Table

The ANOVA Table option creates a standard analysis of variance (ANOVA) table and decomposes the variance of the data into two components: Between-Groups and Within Groups. The F-ratio is the mean square value for Between Groups divided by the mean square value for Within Groups. If the *p*-value is less than a given value, there is a statistically significant difference between the means of the variables at a given confidence level (see Figure 15-5).



☑ Average
☐ Median
☐ Mode
☐ Geo. Mean
☑ Variance
☑ Std. Deviation
☐ Std. Error
☑ Min.
☑ Max.
☑ Range
☐ Lower Quartile
☐ Upper Quartile
☐ Interquartile Range
☐ Skewness
☑ Std. Skewness
☐ Kurtosis
☑ Std. Kurtosis
☐ Coeff. of Var.
☐ Sum

Figure 15-4.   Summary Statistics Options Dialog Box



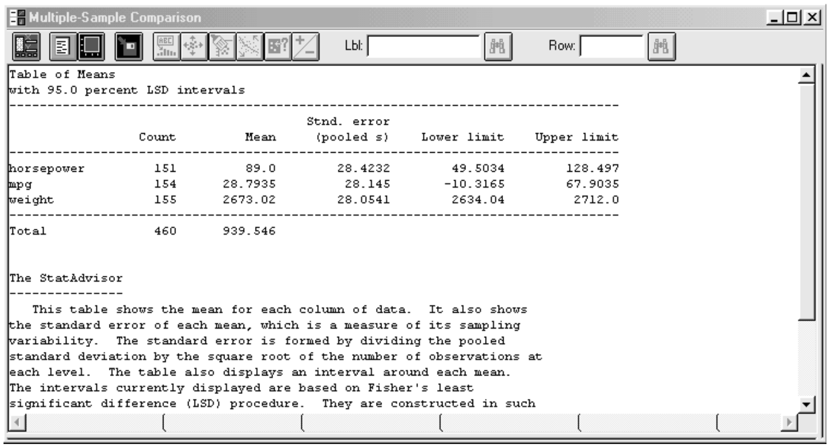
```
ANOVA Table

                            Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Between groups           7.0274E8      2     3.5137E8    2880.32       0.0000
Within groups          5.57494E7    457     121990.0
-----------------------------------------------------------------------------
Total (Corr.)            7.5849E8    459


The StatAdvisor
---------------
   The ANOVA table decomposes the variance of the data into two
components: a between-group component and a within-group component.
The F-ratio, which in this case equals 2880.32, is a ratio of the
between-group estimate to the within-group estimate.  Since the
P-value of the F-test is less than 0.05, there is a statistically
significant difference between the means of the 3 variables at the
95.0% confidence level.  To determine which means are significantly
different from which others, select Multiple Range Tests from the list
```

Figure 15-5.   ANOVA Table

The Sum of Squares - Between Groups statistic is the measure of variability among the different samples. The Sum of Squares - Within Groups statistic is the measure of variability within each of the samples. The Total is the measure of variability for all the data around the grand mean. Each Mean Square is the Sum of Squares for the source of the variation divided by the degrees of freedom (df). The *p*-values indicate the significance level. Small significance levels (less than .05 for most practical applications) indicate that the averages of the samples differ significantly.

Table of Means

The Table of Means option shows the mean for each column of data (see Figure 15-6). It also shows the standard error for each mean, which is a measure of its sampling variability. The standard error is calculated by dividing the pooled standard deviation by the square root of the number of observations at each level.

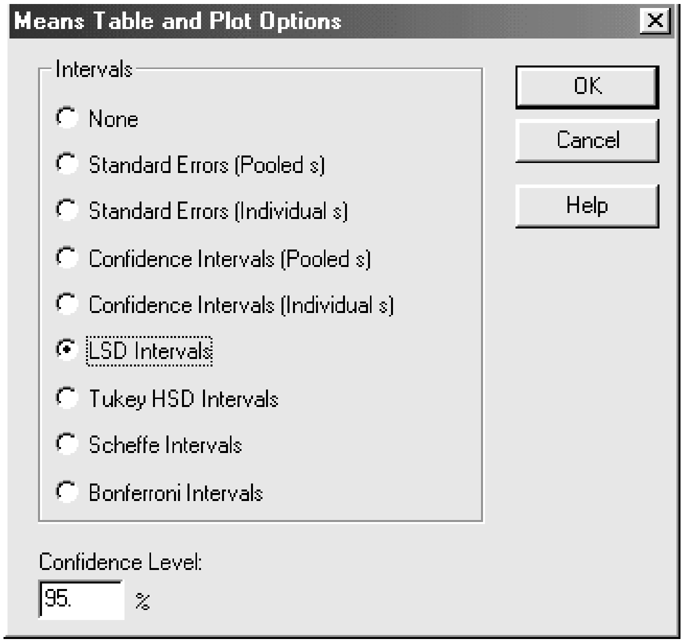


```
Table of Means
with 95.0 percent LSD intervals
--------------------------------------------------------------------------------
                                       Stnd. error
                 Count         Mean     (pooled s)    Lower limit    Upper limit
--------------------------------------------------------------------------------
horsepower         151         89.0        28.4232        49.5034        128.497
mpg                154      28.7935         28.145       -10.3165        67.9035
weight             155      2673.02        28.0541        2634.04         2712.0
--------------------------------------------------------------------------------
Total              460      939.546


The StatAdvisor
---------------
   This table shows the mean for each column of data.  It also shows
the standard error of each mean, which is a measure of its sampling
variability.  The standard error is formed by dividing the pooled
standard deviation by the square root of the number of observations at
each level.  The table also displays an interval around each mean.
The intervals currently displayed are based on Fisher's least
significant difference (LSD) procedure.  They are constructed in such
```

Figure 15-6. Table of Means

The table also displays an interval around each mean, which is based on Fisher's least significant difference (LSD). The intervals are constructed so that if two means are the same, their intervals will overlap a given percentage of time. You can choose not to calculate the intervals, or to calculate them using one of eight methods, which are discussed in detail in Online Help. The eight methods are Standard Errors (Pooled s), Standard Errors (Individual s), Confidence Intervals (Pooled s), Confidence Intervals (Individual s), LSD (Least Significant Differences) Intervals, Tukey HSD (Honestly Significant Differences) Intervals, Scheffe Intervals, or Bonferroni Intervals.

Use the *Means Table and Plot Options* dialog box to choose the method that will be used to calculate the values for the table, and to enter a value for the confidence level that will be used to calculate the confidence intervals (see Figure 15-7). You can use the intervals in the Multiple Range Tests option to determine which means are statistically different.



Means Table and Plot Options

Intervals

○ None

○ Standard Errors (Pooled s)

○ Standard Errors (Individual s)

○ Confidence Intervals (Pooled s)

○ Confidence Intervals (Individual s)

● LSD Intervals

○ Tukey HSD Intervals

○ Scheffe Intervals

○ Bonferroni Intervals

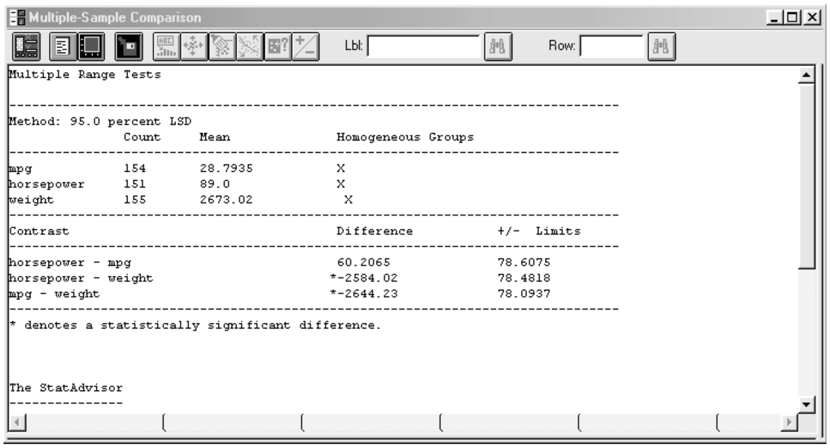
Confidence Level: 95. %

Figure 15-7. Means Table and Plot Options Dialog Box

Multiple Range Tests

The Multiple Range Tests option applies various methods to the data to determine which means are significantly different (see Figure 15-8). The methods are LSD (Least Significant Differences), Scheffe, Bonferroni, Newman-Keuls, and Duncan. *See Online Help for detailed descriptions of each of these methods.*

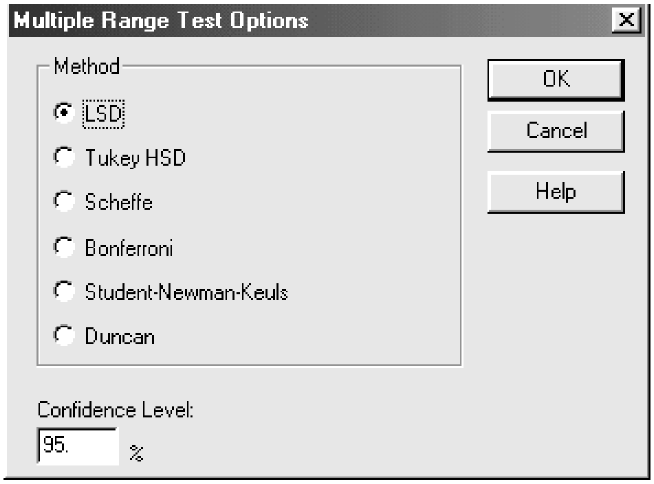
The top portion of the table identifies homogenous groups with columns of Xs that, within each column, indicate the group means for which there are no statistically significant differences. The bottom portion of the table shows the estimated difference between each pair of means. An asterisk indicates a statistically significant difference at a given confidence level.



```
Multiple Range Tests

--------------------------------------------------------------------------------
Method: 95.0 percent LSD
               Count     Mean              Homogeneous Groups
--------------------------------------------------------------------------------
mpg            154       28.7935           X
horsepower     151       89.0              X
weight         155       2673.02            X
--------------------------------------------------------------------------------
Contrast                                   Difference           +/-  Limits
--------------------------------------------------------------------------------
horsepower - mpg                           60.2065              78.6075
horsepower - weight                       *-2584.02             78.4818
mpg - weight                              *-2644.23             78.0937
--------------------------------------------------------------------------------
* denotes a statistically significant difference.



The StatAdvisor
---------------
```

Figure 15-8.   Multiple Range Tests

Use the *Multiple Range Tests Options* dialog box to choose the method that will be used to calculate the values for the table, and to enter a value for the confidence level that will be used to calculate the confidence intervals (see Figure 15-9).



Method

● LSD

○ Tukey HSD

○ Scheffe

○ Bonferroni

○ Student-Newman-Keuls

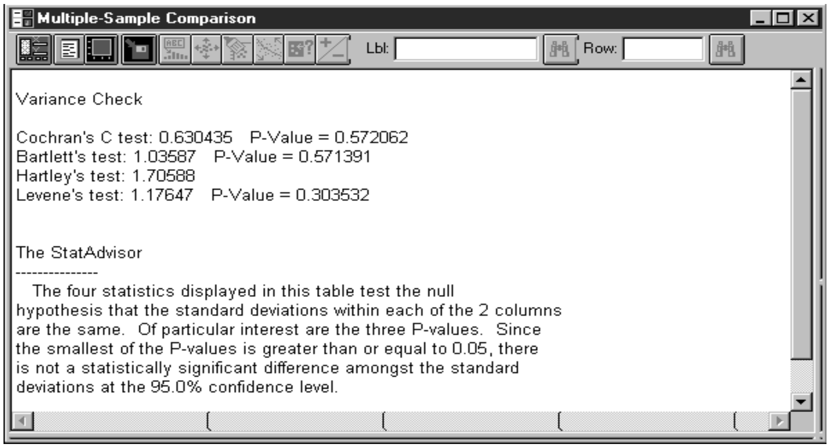
○ Duncan

Confidence Level: 95. %

Figure 15-9.   Multiple Range Tests Options Dialog Box

Variance Check

The Variance Check option performs four tests to test the null hypothesis that the standard deviations within each of the samples are the same: Cochran's C Test, Bartlett's Test, Hartley's Test and Levene's Test (see Figure 15-10). Each of these statistics tests the homogeneity of variances assumption that the variances are equal (homogeneous) across the cells of the between-groups design. If the *p*-values are less than a given value, there is a statistically significant difference among the standard deviations at a given confidence level.



```
Variance Check

Cochran's C test: 0.630435   P-Value = 0.572062
Bartlett's test: 1.03587   P-Value = 0.571391
Hartley's test: 1.70588
Levene's test: 1.17647   P-Value = 0.303532


The StatAdvisor
---------------
   The four statistics displayed in this table test the null
hypothesis that the standard deviations within each of the 2 columns
are the same.  Of particular interest are the three P-values.  Since
the smallest of the P-values is greater than or equal to 0.05, there
is not a statistically significant difference amongst the standard
deviations at the 95.0% confidence level.
```

Figure 15-10. Variance Check

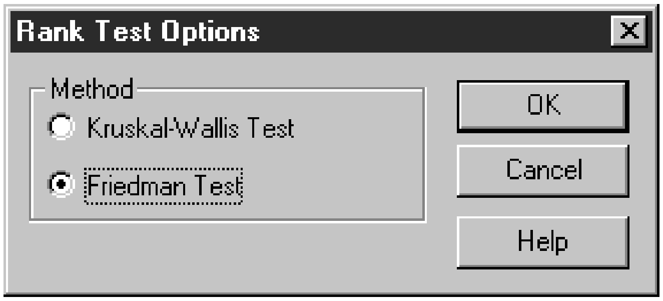
Levene's Test compares the sample variances by performing an analysis of variance on the absolute deviations of the data values from their respective sample menus. It is less sensitive than Bartlett's test to departures from normality of the underlying populations.

Kruskal-Wallis and FriedmanTests

The Kruskal-Wallis Test option is a nonparametric method that tests the assumption that the medians of the samples are equal.

For balanced data in which each row with any data has data for all of the columns, a Friedman Test may be performed instead of the Kruskal-Wallis Test. The hypothesis tested, that of equal medians in each column, is the same for both tests. However, the Friedman test is only meaningful if the data is blocked by row. The test will only be performed if the data is entered in multiple columns.

Use the *Rank Test Options* dialog box to select either the Kruskal-Wallis or Friedman test to determine differences in medians (see Figure 15-11).



Method
○ Kruskal-Wallis Test
● Friedman Test

Figure 15-11. Rank Test Options Dialog Box

The program combines and ranks the values in the entire set of observations from the lowest to the highest, calculates the mean of the ranks for each sample, and tests the hypothesis for the assumption that the medians are the same. If the *p*-value is less than .05, you can reject the hypothesis.

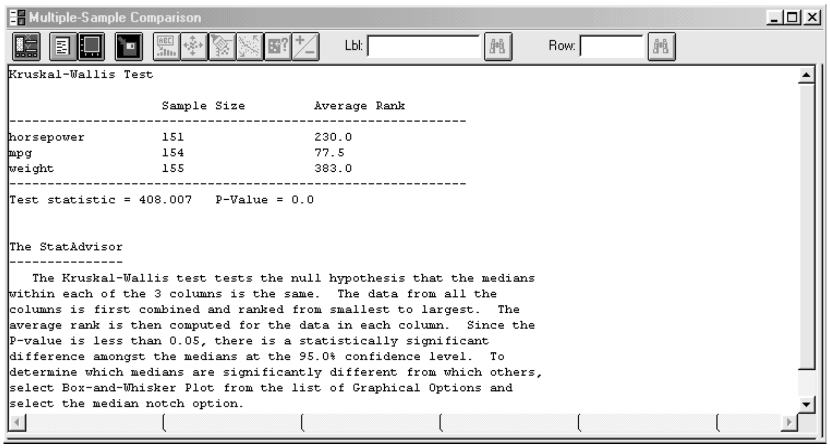
See Figures 15-12 and 15-13 for samples of each test.

Graphical Options

Scatterplot

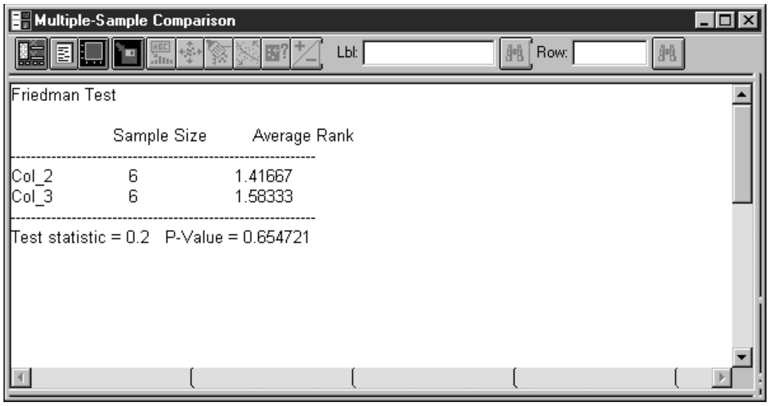
The Scatterplot option creates a plot of the values for each variable along a single axis as point symbols with no connecting lines (see Figure 15-14).

If the points overlap, use the *Jitter* button on the Analysis toolbar to add a small amount of horizontal offset. *For more information about jittering, see the section, "Jittering, Brushing, and Smoothing Points," in Chapter 5, Working with Graphs and Graphics Options.*



```
Kruskal-Wallis Test

                    Sample Size         Average Rank
------------------------------------------------------------
horsepower          151                 230.0
mpg                 154                 77.5
weight              155                 383.0
------------------------------------------------------------
Test statistic = 408.007   P-Value = 0.0


The StatAdvisor
---------------
   The Kruskal-Wallis test tests the null hypothesis that the medians
within each of the 3 columns is the same.  The data from all the
columns is first combined and ranked from smallest to largest.  The
average rank is then computed for the data in each column.  Since the
P-value is less than 0.05, there is a statistically significant
difference amongst the medians at the 95.0% confidence level.  To
determine which medians are significantly different from which others,
select Box-and-Whisker Plot from the list of Graphical Options and
select the median notch option.
```

Figure 15-12.   Kruskal-Wallis Test



```
Friedman Test

              Sample Size      Average Rank
------------------------------------------------------------
Col_2            6              1.41667
Col_3            6              1.58333
------------------------------------------------------------
Test statistic = 0.2   P-Value = 0.654721
```

Figure 15-13.   Friedman Test

Means Plot

The Means Plot option displays a plot of the means for each sample and the intervals around each mean (see Figure 15-15). The intervals are constructed so that if two means are the same, their intervals will overlap a given percentage.

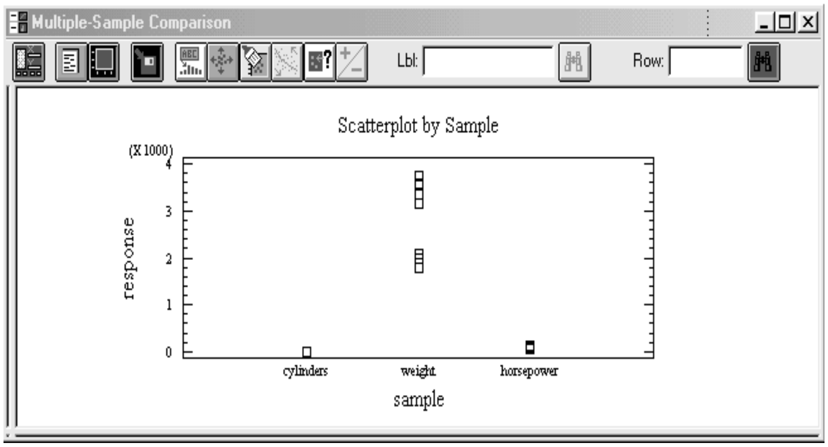


Figure 15-14. Scatterplot

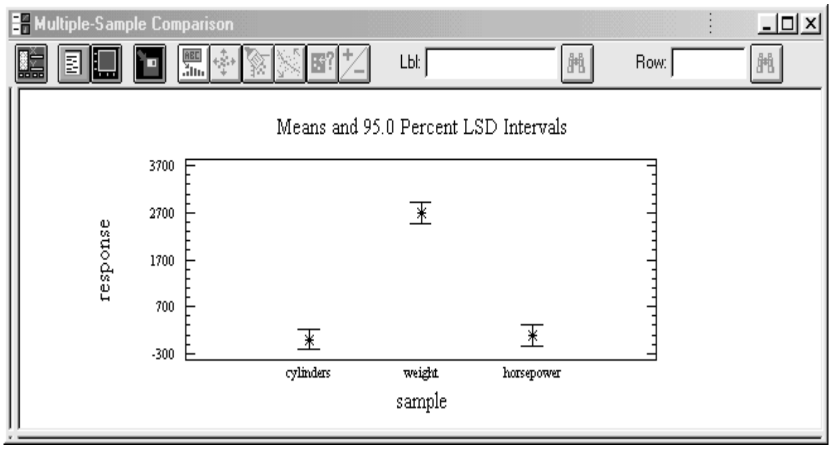


Figure 15-15. Means Plot

Use the *Means Table and Plot Options* dialog box to choose the method that will be used to calculate the values for the plot, and to enter a value for the confidence level that will be used to calculate the confidence intervals (see Figure 15-7 for an example of this dialog box).

Box-and-Whisker Plot

The Box-and-Whisker Plot option calculates a plot for each column of data (see Figure 15-16). The program divides the data into four areas of equal frequency (quartiles). The rectangular part of the plot extends from the lower quartile to the upper quartile, covering the center half of each sample. The center lines in each box show the location of the sample medians, while the plus (+) signs indicate the location of the sample means.

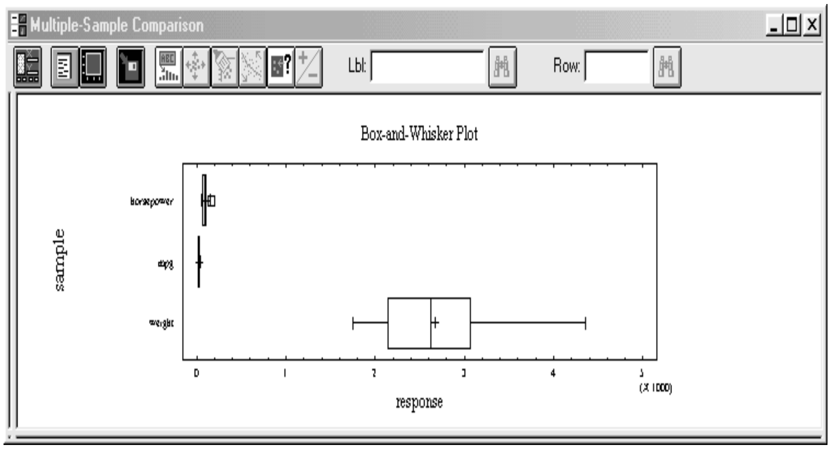
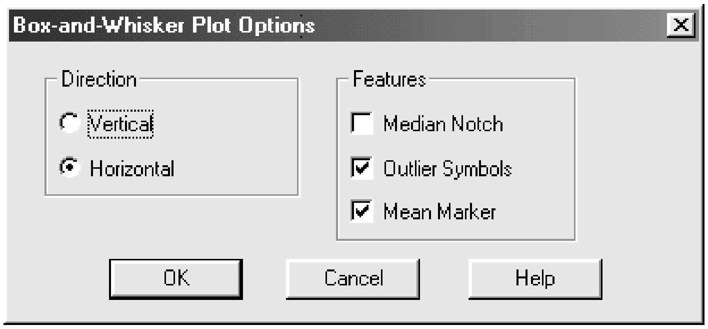


Figure 15-16. Box-and-Whisker Plot

Horizontal lines, known as whiskers, extend from the box to the minimum and maximum values in each sample, except for any outside or far outside points, which are plotted separately. Outside points lie more than 1.5 times the interquartile range above or below the box and are shown as small squares. Far outside points lie more than 3.0 times the interquartile range above or below the box and are shown as small squares with plus (+) signs through them.

Use the *Box-and-Whisker Options* dialog box to indicate the direction of the plot and to choose the features that will appear on it (see Figure 15-17).



Direction

○ Vertical

● Horizontal

Features

□ Median Notch

☑ Outlier Symbols

☑ Mean Marker

Figure 15-17. Box-and-Whisker Plot Options Dialog Box

Residuals versus Samples Graphical Option

The Residuals versus Samples option creates a plot of the residuals within each column (see Figure 15-18). The residuals are equal to the observed data minus the mean of the column from which they originate. The plot is helpful when you need to determine if the variance is approximately the same.

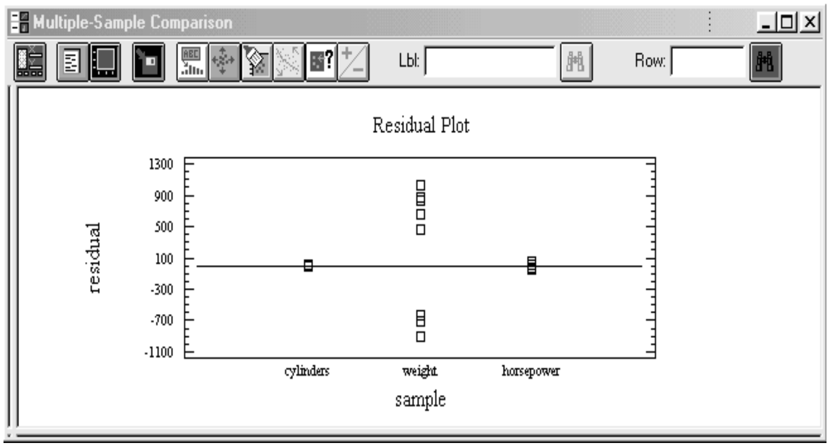


Figure 15-18. Residuals versus Samples Plot

Residuals versus Predicted

The Residuals versus Predicted option creates a plot of the residuals versus the predicted values. The residuals are equal to the observed values minus the mean of the column from which they originate. The predicted values are equal to the column means (see Figure 15-19).

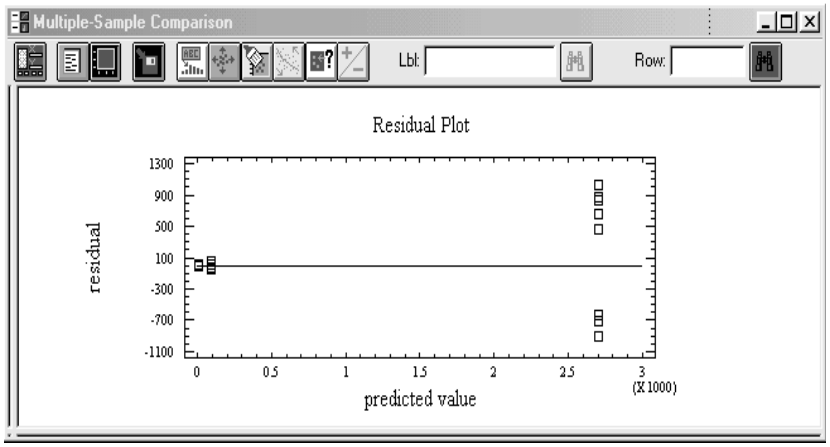


Figure 15-19. Residuals versus Predicted Plot

The plot is helpful for detecting heteroscedasticity, in which the variance changes together with the mean. If the points create a general funnel-shaped pattern, it indicates a type of heteroscedasticity that can often be corrected by transforming the data. In those cases, try entering LOG(Y) or 1/Y as the variable on the Analysis dialog box.

Residuals versus Observation

The Residuals versus Observation option creates a plot of the residuals versus the observation. The plot is helpful when you need to identify sequential correlations among the residuals (see Figure 15-20).

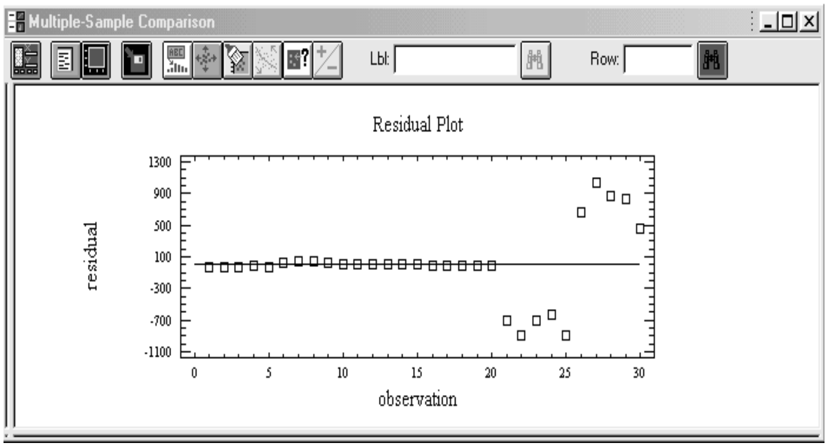


Figure 15-20. Residuals versus Observation Plot

Analysis of Means Plot

The Analysis of Means (ANOM) Plot option creates a plot that shows the mean for each of the samples, the grand mean, and the decision limits, which determine the groups that differ significantly at the grand mean (see Figure 15-21). If any points are outside the decision limits, it can be concluded that there is a statistically significant difference between the samples.

Use the *Analysis of Means Plot Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals and to enter the number of decimal places that will be used for the decision limits (see Figure 15-22).

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are nine selections: Counts, Means, Standard Deviations, Standard Errors, Labels, Residuals, Ranges, Data Column, and Code Column. You can save group ranges as well as other statistics. You can also save the data in a different format. If data is entered in multiple columns, you can save the data in a single column with row numbers.

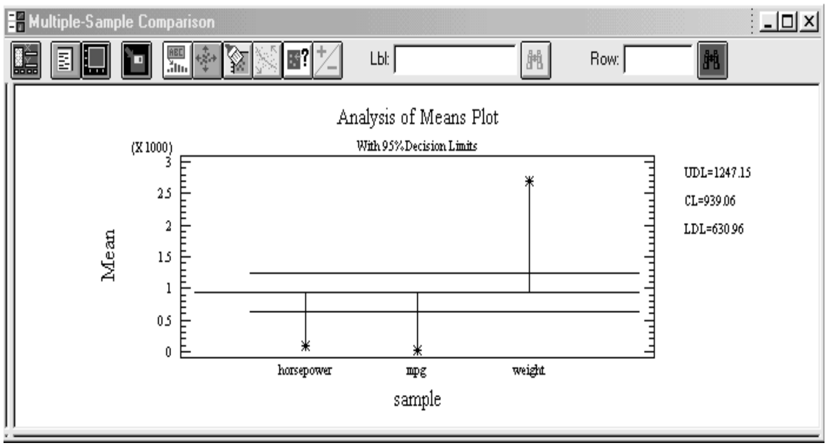
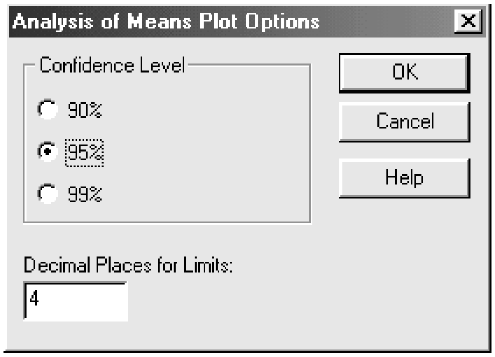


Figure 15-21. Analysis of Means Plot



Confidence Level

○ 90%

● 95%

○ 99%

Decimal Places for Limits: 4

Figure 15-22. Analysis of Means Plot Options Dialog Box

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. New York: Wiley.

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: Wiley.

Hollander, M. and Wolfe, D. A. 1973. *Nonparametric Statistical Methods*. New York: Wiley.

Levene, H. 1960. *Robust Tests for Equality of Variances*. In: *Contributions to Probability and Statistics*, ed. by Olkin, et. Al. Standord University Press.

Neter, J., Wasserman, W., and Kutner, M. 1985. *Applied Linear Statistical Models*. Homewood, Illinois: Richard E. Irwin, Inc.

Owen, D. B. 1962. *Handbook of Statistical Tables*. Redding, Massachusetts: Addison-Wesley.

Using the Comparison of Proportions Analysis

The Comparison of Proportions Analysis uses an ANOM (analysis of means) statistical technique, which is conceptually similar to a control chart. The plot shows the observed proportions plotted against the decision limits. This type of analysis uses critical values from a sampling distribution, which lend power that is comparable to an ANOVA under similar conditions. While an ANOM is not the best test from a mathematical perspective, it is easier to use and understand, which gives it an advantage over ANOVA.

The Comparison of Proportions Analysis compares the proportion of items that have a particular characteristic of an attribute. For example, at some time or another, manufacturers face the problem of having a variable that turns out to also be an attribute, such as a light bulb that will or will not light or a battery whose life is or is not below standard (Ott, 1979).

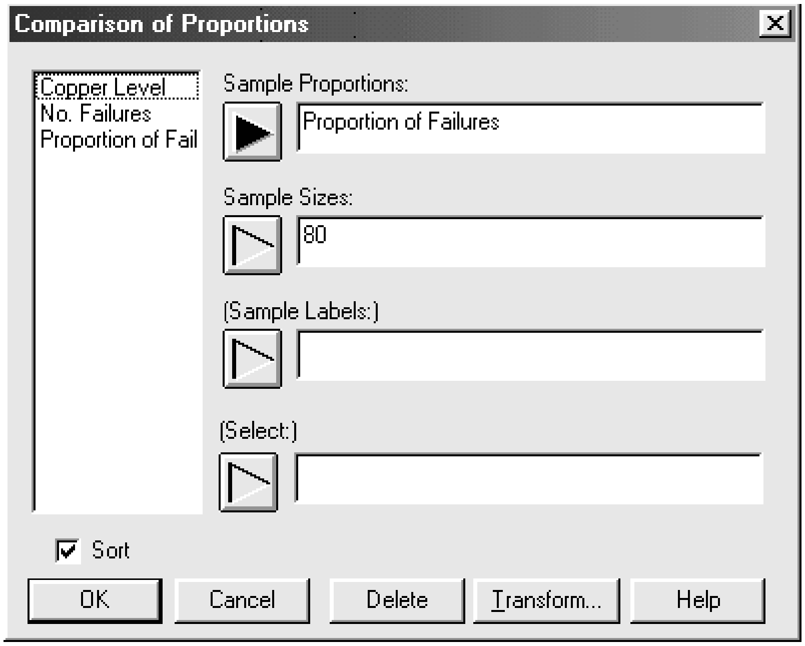
The data for this analysis should meet these criteria:

- the sampling distribution of the proportion should be a binomial distribution
- if *np* and *n* (1 - *p*) are both greater than 5, you should be able to use a normal distribution to approximate the sampling distribution of a portion.

If the variable you are using consists of attribute data, you can use an ANOM to compare the proportion of the items to test for differences between the proportions.

The analysis uses three or more proportions to test the hypothesis that all the proportions are equal. The results include several reports as well as an ANOM (analysis of means) plot.

To access the analysis, from the menus, choose: COMPARE... MULTIPLE SAMPLES... COMPARISON OF PROPORTIONS... (see Figure 15-23).



Sample Proportions: Proportion of Failures

Sample Sizes: 80

(Sample Labels:)

(Select:)

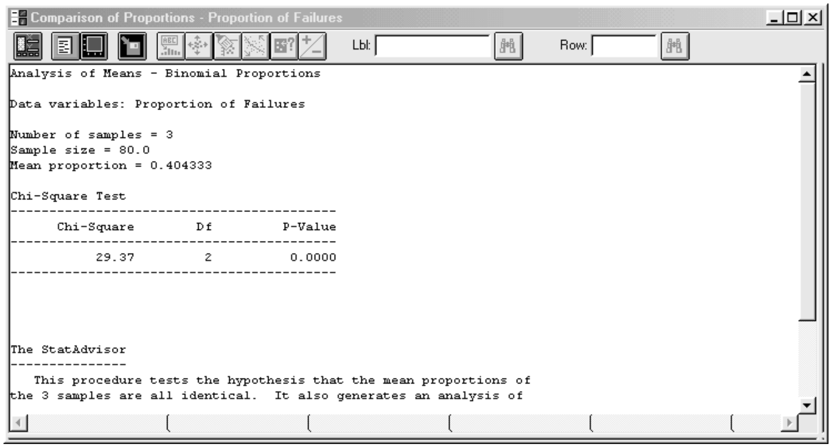
☑ Sort

Figure 15-23. Comparison of Proportions Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the calculated decision limits, the average proportion, and the observed proportions. Any proportions outside the decision limits are noted by an asterisk (see Figure 15-24). The lower portion of the table displays the results of a chi-square test, which compares each sample value with its grand mean. If the *p*-value is less than 0.05, there are significant differences between the samples at the 95 percent confidence level.



```
Analysis of Means - Binomial Proportions

Data variables: Proportion of Failures

Number of samples = 3
Sample size = 80.0
Mean proportion = 0.404333

Chi-Square Test
------------------------------------------
      Chi-Square        Df         P-Value
------------------------------------------
           29.37         2          0.0000
------------------------------------------




The StatAdvisor
---------------
   This procedure tests the hypothesis that the mean proportions of
the 3 samples are all identical.  It also generates an analysis of
```

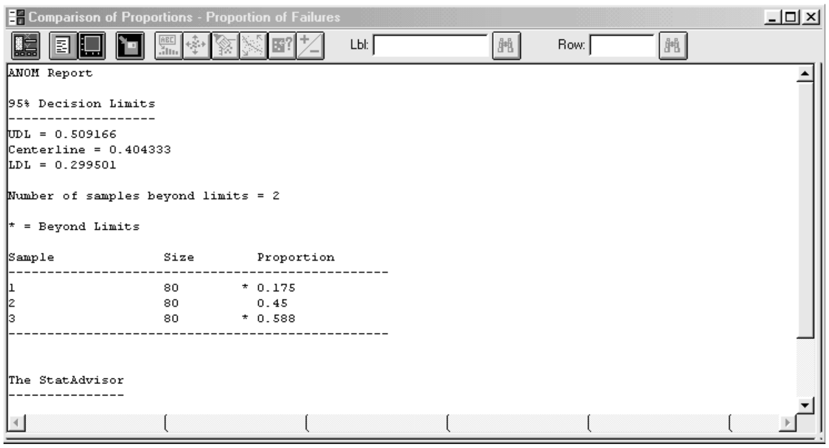
Figure 15-24. Analysis Summary

Analysis of Means (ANOM) Report

The Analysis of Means (ANOM) Report option creates a table of the observed proportions for each of the samples (see Figure 15-25). The ANOM for proportions is performed by first calculating the average proportion, then calculating an estimate of the variation, based on the average proportion.

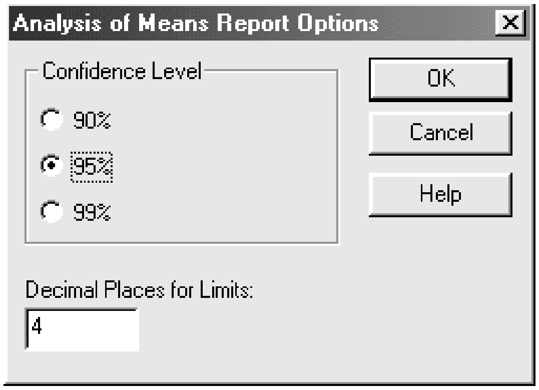
The top portion of the table shows the values for the decision limits at a given confidence level, and the number of samples beyond the limits. An asterisk indicates values that are beyond the limits. The bottom portion of the table displays the sample sizes and the observed proportions.

Use the *Analysis of Means Report Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals and to enter the number of decimal places that will be used for the decision limits (see Figure 15-26).



```
ANOM Report

95% Decision Limits
-------------------
UDL = 0.509166
Centerline = 0.404333
LDL = 0.299501

Number of samples beyond limits = 2

* = Beyond Limits

Sample              Size        Proportion
-------------------------------------------------
1                   80        * 0.175
2                   80          0.45
3                   80        * 0.588
-------------------------------------------------


The StatAdvisor
---------------
```

Figure 15-25.   Analysis of Means Report



Confidence Level

○ 90%

● 95%

○ 99%

Decimal Places for Limits: 4

Figure 15-26.   Analysis of Means Report Options Dialog Box

Graphical Options

Analysis of Means (ANOM) Plot

The Analysis of Means (ANOM) Plot option creates a plot that shows each group mean, the centerline at the grand mean, and the decision limits, which, determine the groups that differ significantly at the grand mean (see Figure 15-27). If any points are outside the decision limits, it can be concluded that there is a statistically significant difference between the proportions.

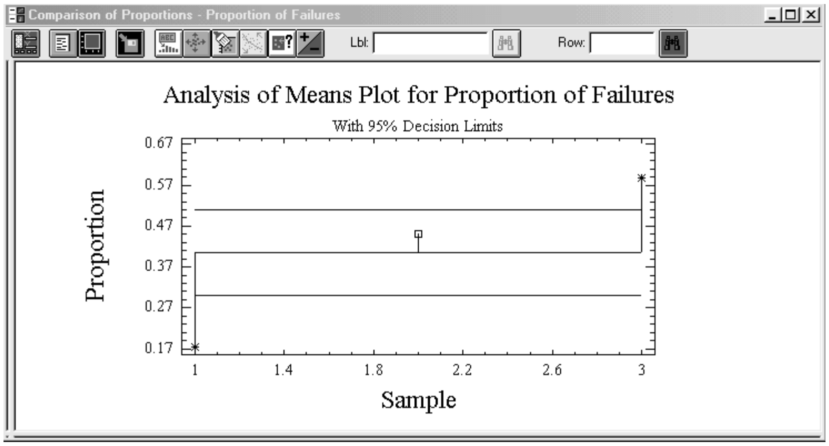


Figure 15-27. Analysis of Means Plot

Use the *Analysis of Means Plot Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals and to enter the number of decimal places that will be used for the decision limits (see Figure 15-22 for an example of this dialog box).

References

Ott, E. R. and E. G. Schilling. 1990. *Process Quality Control*, second edition. New York: McGraw-Hill Book Company.

Ott, E. R. 1983.   “Analysis of Means: A Graphical Procedure,” *Journal of Quality Technology*, **15**:10-18.

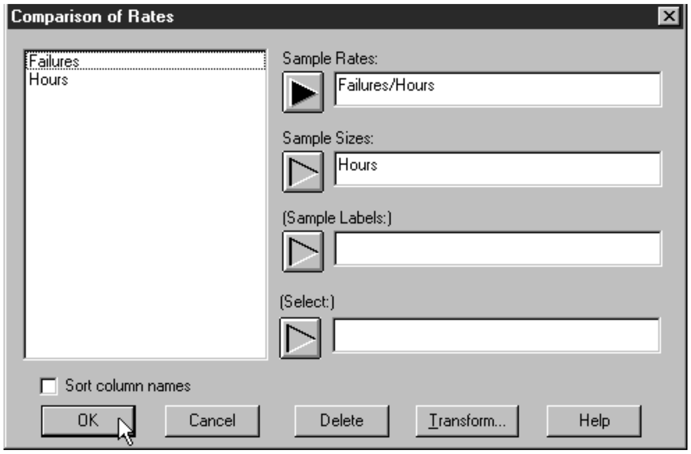
Using the Comparison of Rates Analysis

The Comparison of Rates Analysis uses an ANOM (analysis of means) technique for comparing several populations you use when the response is rate data. This analysis is essentially the same as the Comparison of Proportions Analysis, except that the sample size is such that the analysis uses a normal approximation to the Poisson distribution.

Ott (1979) provides this example of a quality control environment, where workers are monitoring the number of defective items from a production line. An inspection unit is first defined for each of the populations, which might be defined as a period of time, a fixed number of items, or a fixed unit of measurement; for example, an inspection unit of a square-foot section of material from a weaving loom.

Periodically, a square-foot section of the material is examined and the number of flaws is recorded. The number of items with the defect is noted as $c_1,...,k$. The results are then calculated and plotted to see if any count falls outside the decision lines, which means there is a statistically significant difference among the counts.

To access the analysis, from the menus, choose: COMPARE... MULTIPLE SAMPLES... COMPARISON OF RATES... (see Figure 15-28).



Comparison of Rates

Failures
Hours

Sample Rates: Failures/Hours

Sample Sizes: Hours

(Sample Labels:)

(Select:)

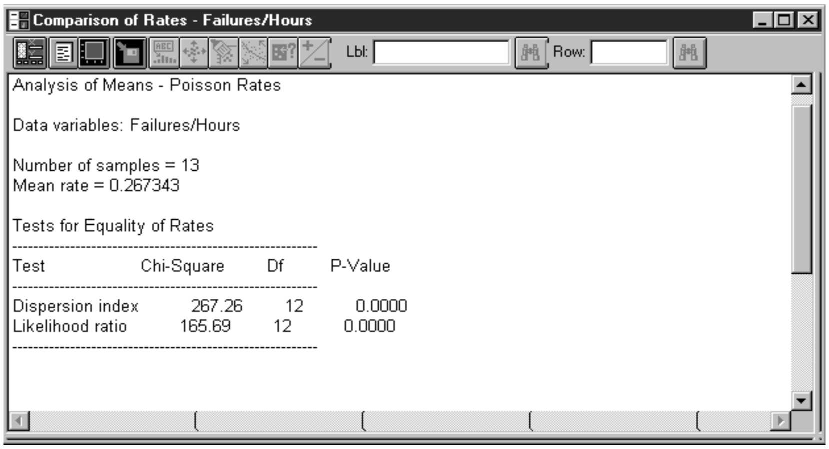
□ Sort column names

Figure 15-28. Comparison of Rates Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the name of the variable, the number of samples, and the mean count (see Figure 15-29). The lower portion of the table displays the results of a chi-square test, which compares each sample value with its grand mean. If the *p*-value is less than 0.05, there are significant differences between the samples at the 95 percent confidence level.



```
Analysis of Means - Poisson Rates

Data variables: Failures/Hours

Number of samples = 13
Mean rate = 0.267343

Tests for Equality of Rates
--------------------------------------------------------
Test                Chi-Square       Df         P-Value
--------------------------------------------------------
Dispersion index          267.26       12          0.0000
Likelihood ratio        165.69       12         0.0000
--------------------------------------------------------
```

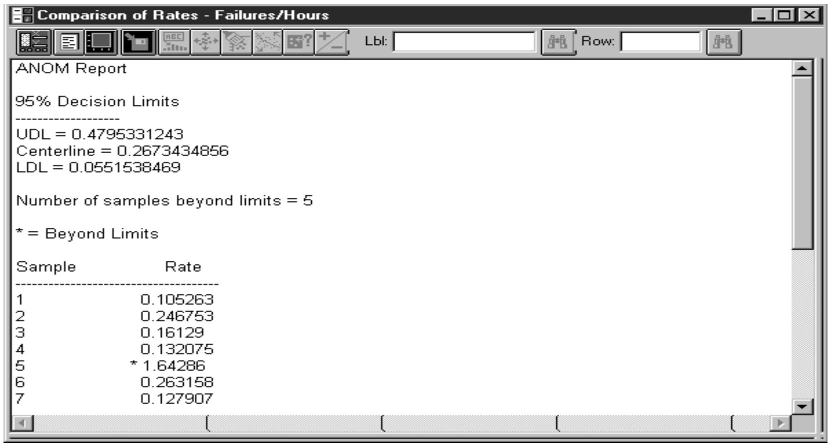
Figure 15-29.   Analysis Summary

Analysis of Means (ANOM) Report

The Analysis of Means (ANOM) Report option creates a table of the observed counts for each of the samples (see Figure 15-30). The ANOM for counts is performed by first calculating the average count, then calculating an estimate of the variation, based on the average count.

The top portion of the table shows the values for the decision limits at a given confidence level, and the number of samples beyond the limits. An asterisk indicates values that are beyond the limits. The bottom portion of the table displays the decision limits and the observed counts.

Use the *Analysis of Means Report Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals and to enter the number of decimal places that will be used for the decision limits (see Figure 15-26 for an example of this dialog box).



```
ANOM Report

95% Decision Limits
-------------------
UDL = 0.4795331243
Centerline = 0.2673434856
LDL = 0.0551538469

Number of samples beyond limits = 5

* = Beyond Limits

Sample              Rate
------------------------------------
1                   0.105263
2                   0.246753
3                   0.16129
4                   0.132075
5                 * 1.64286
6                   0.263158
7                   0.127907
```

Figure 15-30.   Analysis of Means Report

Graphical Options

Analysis of Means (ANOM) Plot

The Analysis of Means (ANOM) Plot option creates a plot that shows each group mean, the centerline at the grand mean, and the decision limits, which determine the groups that differ significantly at the grand mean (see Figure 15-31). If any points are outside the decision limits, it can be concluded that there is a statistically significant difference between the counts.

Use the *Analysis of Means Plot Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals and to enter the number of decimal places that will be used for the decision limits (see Figure 15-22 for an example dialog box).

References

Cox, D.R. And Lewis, P.A.W. 1966. *The Statistical Analysis of Series of Events.* London: Methuen and Company.

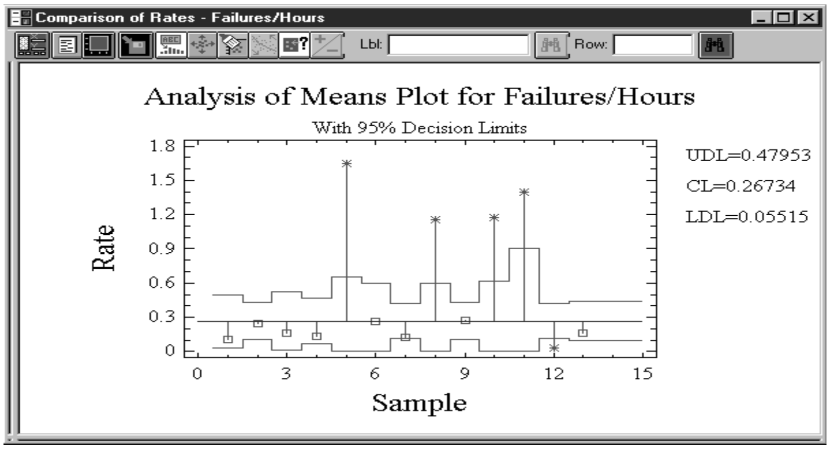


Figure 15-31. Analysis of Means (ANOM) Plot

Ott, E. R. and E. G. Schilling. 1990. *Process Quality Control*, second edition. New York: McGraw-Hill Book Company.

Ott, E. R. 1983. “Analysis of Means: A Graphical Procedure,” *Journal of Quality Technology*, **15:10-18.**

16 Performing Analysis of Variance Tests

This chapter describes analyses you use to perform four forms of Analysis of Variance (ANOVA): Factor Means Plot, One-Way ANOVA, Multifactor ANOVA, and Variance Components:

- **Factor Means Plot**
  Use this analysis with two or more experimental factors to plot the means at different factor levels.

- **One-Way ANOVA**
  Use this analysis with a single factor to analyze its effect on a response variable.

- **Multifactor ANOVA**
  Use this analysis with two or more experimental factors to analyze their effect on a response variable.

- **Variance Components**
  Use this analysis with models that have random effects to estimate the variance each random effect contributes to the dependent variable.

The four types of ANOVA allow you to:

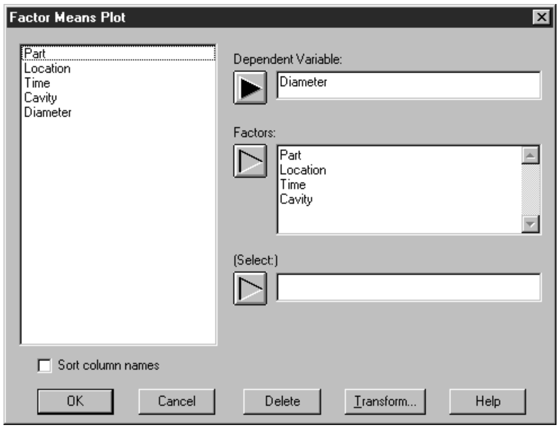
- divide the variability in the observed response into components, each attributable to a single experimental factor
- determine if the mean response varies at different levels of each experimental factor
- determine which factors interact in a multifactor experiment.

Using Factor Means Plot

Factor Means Plot is a statistical technique that produces an analysis of variance for an interval-level dependent variable by an independent variable.

The analysis examines the effect of one qualitative factor in one response variable where the design can be balanced or unbalanced.

To access the analysis, from the menus, choose: COMPARE... ANALYSIS OF VARIANCE... FACTOR MEANS PLOT... (see Figure 16-1).



Dependent Variable: Diameter

Factors: Part, Location, Time, Cavity

(Select:)

Figure 16-1. Factor Means Plot Analysis Dialog Box

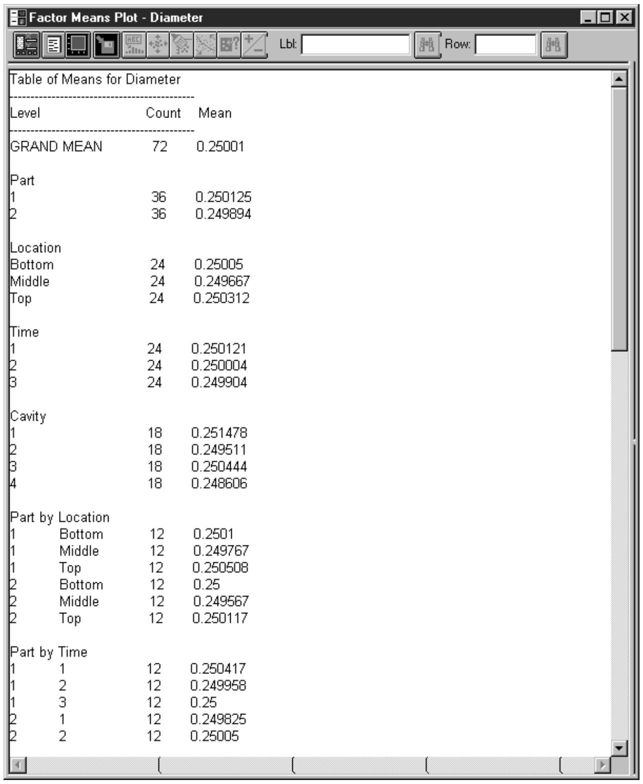
Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that contains the name of the dependent variable and the factors. It also displays the number of complete cases in the analysis.

Table of Means

The Table of Means option creates a table that shows the mean of the variable for each level of the factors and interations (see Figure 16-2). The table includes the number of observations (Count) at each level and the mean at each level.



```
Table of Means for Diameter
--------------------------------------------
Level                    Count    Mean
--------------------------------------------
GRAND MEAN                72      0.25001

Part
1                         36      0.250125
2                         36      0.249894

Location
Bottom                    24      0.25005
Middle                    24      0.249667
Top                       24      0.250312

Time
1                         24      0.250121
2                         24      0.250004
3                         24      0.249904

Cavity
1                         18      0.251478
2                         18      0.249511
3                         18      0.250444
4                         18      0.248606

Part by Location
1        Bottom           12      0.2501
1        Middle           12      0.249767
1        Top              12      0.250508
2        Bottom           12      0.25
2        Middle           12      0.249567
2        Top              12      0.250117

Part by Time
1        1                12      0.250417
1        2                12      0.249958
1        3                12      0.25
2        1                12      0.249825
2        2                12      0.25005
```

Figure 16-2. Table of Means

Graphical Options

Means Plot

The Means Plot option creates a plot of the means and various combinations of each factor (see Figure 16-3). The plots on the diagonal show the overall level means while the off diagonal plots show the mean for each pair of factor levels.

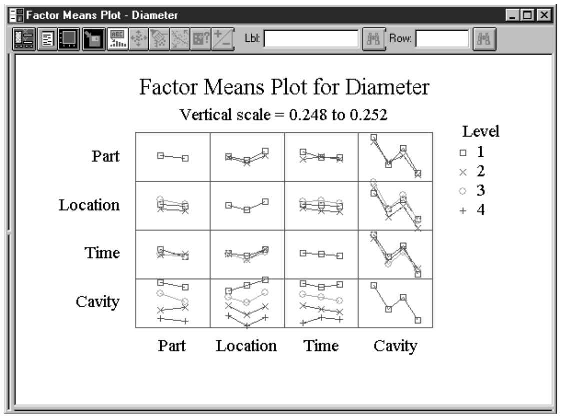

Factor Means Plot - Diameter
Lbl:
Row:
Factor Means Plot for Diameter
Vertical scale = 0.248 to 0.252
Level
□ 1
× 2
○ 3
+ 4
Part
Location
Time
Cavity
Part
Location
Time
Cavity

Figure 16-3. Factor Means Plot

Using One-Way ANOVA

One-Way ANOVA is a statistical technique that produces a one-way analysis of variance for an interval-level dependent variable by an independent variable. The analysis examines the effect of one qualitative factor on one response variable where the design can be balanced or unbalanced.

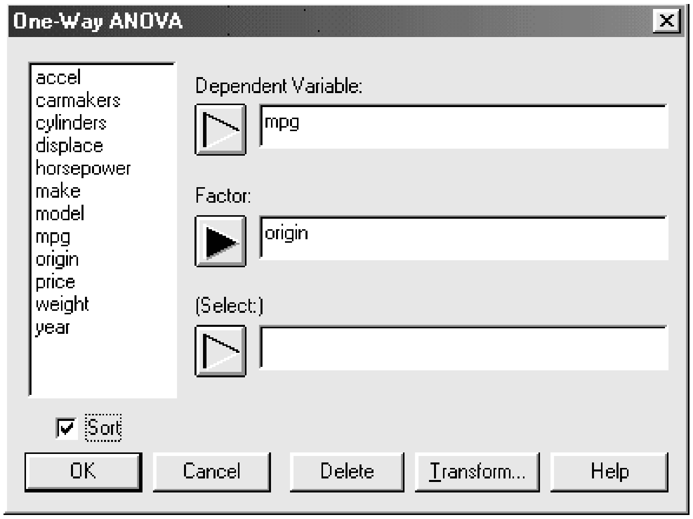
Assumptions of the analysis are that the:

- populations from which you collect the samples must be normally or approximately normally distributed
- samples must be independent
- variances of the populations must be equal.

The analysis tests the null hypothesis that all population means are equal, versus the alternative hypothesis, that at least one population mean is different.

This analysis requires all response data to be within a single column with a factor or code column that defines the sample that a particular observation is taken from. Each observation is classified into groups according to one factor. For example, you might examine how three different reading textbooks produce different reading levels for children, or how cylinders in an automobile engine affect its mileage rate.

To access the analysis, from the menus, choose: COMPARE... ANALYSIS OF VARIANCE... ONE-WAY ANOVA... (see Figure 16-4).



One-Way ANOVA

accel
carmakers
cylinders
displace
horsepower
make
model
mpg
origin
price
weight
year

Dependent Variable: mpg

Factor: origin

(Select:)

☑ Sort

Figure 16-4. The One-Way ANOVA Dialog Box

Tabular Options

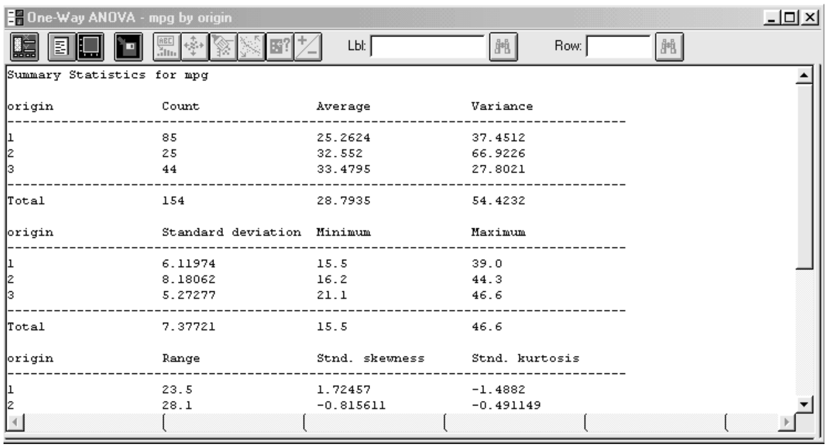
Analysis Summary

The Analysis Summary option creates a summary of the analysis that includes the name of the dependent variable and the factor, as well as the number of observations and levels.

Summary Statistics

The Summary Statistics option creates information about the statistics you choose, and includes the number of values in each variable (the Count); the average, variance, and standard deviation; as well as the minimum and maximum values. It also includes values for the range, standardized skewness, and standardized kurtosis (see Figure 16-5).

The values for the standard skewness and standardized kurtosis are of particular interest because they help determine if the samples are from normal distributions. Values outside the range of -2 to +2 indicate a

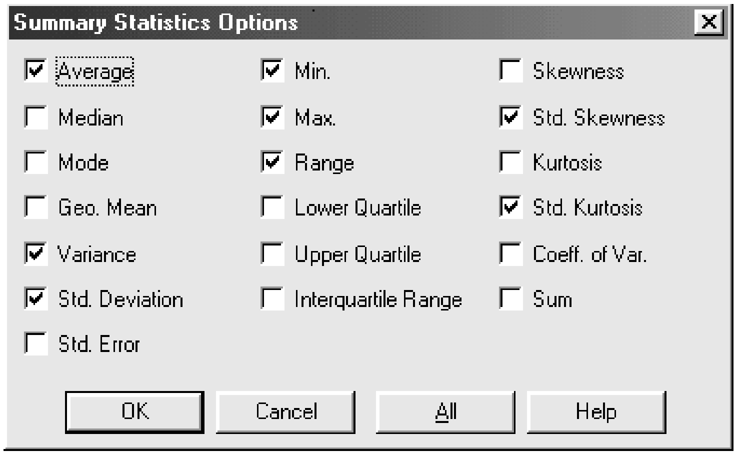


```
Summary Statistics for mpg

origin              Count               Average             Variance
--------------------------------------------------------------------------------
1                   85                  25.2624             37.4512
2                   25                  32.552              66.9226
3                   44                  33.4795             27.8021
--------------------------------------------------------------------------------
Total               154                 28.7935             54.4232

origin              Standard deviation  Minimum             Maximum
--------------------------------------------------------------------------------
1                   6.11974             15.5                39.0
2                   8.18062             16.2                44.3
3                   5.27277             21.1                46.6
--------------------------------------------------------------------------------
Total               7.37721             15.5                46.6

origin              Range               Stnd. skewness      Stnd. kurtosis
--------------------------------------------------------------------------------
1                   23.5                1.72457             -1.4882
2                   28.1                -0.815611           -0.491149
```

Figure 16-5. *Summary Statistics*

significant departure from normality, which tends to invalidate the tests for comparing the standard deviations.

Use the *Summary Statistics Options* dialog box to choose the statistics that



☑ Average
□ Median
□ Mode
□ Geo. Mean
☑ Variance
☑ Std. Deviation
□ Std. Error

☑ Min.
☑ Max.
☑ Range
□ Lower Quartile
□ Upper Quartile
□ Interquartile Range

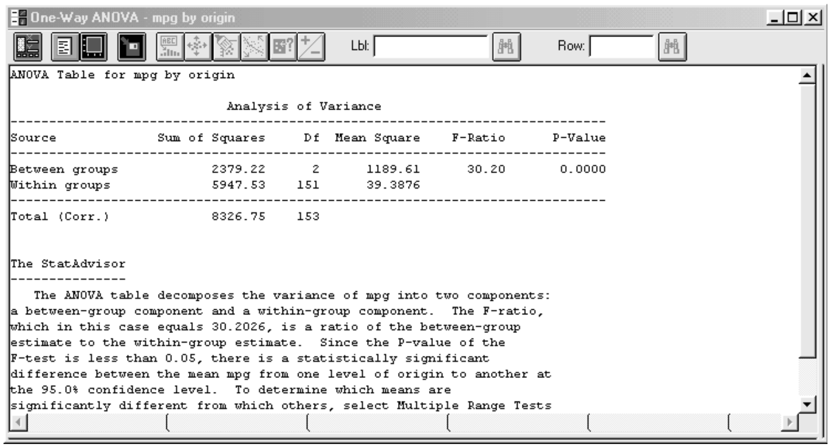
□ Skewness
☑ Std. Skewness
□ Kurtosis
☑ Std. Kurtosis
□ Coeff. of Var.
□ Sum

Figure 16-6. *Summary Statistics Options Dialog Box*

will be calculated (see Figure 16-6).

ANOVA Table

The ANOVA Table option creates a standard ANOVA table and decomposes the variance of the data into two components: Between Groups and Within Groups. The F-ratio is the mean square value for Between Groups divided by the mean square value for Within Groups. If the *p*-value is less than a given alpha, there is a statistically significant difference between the means of the variables at a given confidence level (see Figure 16-7).



```
ANOVA Table for mpg by origin

                            Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Between groups            2379.22      2      1189.61      30.20       0.0000
Within groups             5947.53    151      39.3876
-----------------------------------------------------------------------------
Total (Corr.)             8326.75    153


The StatAdvisor
---------------
   The ANOVA table decomposes the variance of mpg into two components:
a between-group component and a within-group component.  The F-ratio,
which in this case equals 30.2026, is a ratio of the between-group
estimate to the within-group estimate.  Since the P-value of the
F-test is less than 0.05, there is a statistically significant
difference between the mean mpg from one level of origin to another at
the 95.0% confidence level.  To determine which means are
significantly different from which others, select Multiple Range Tests
```

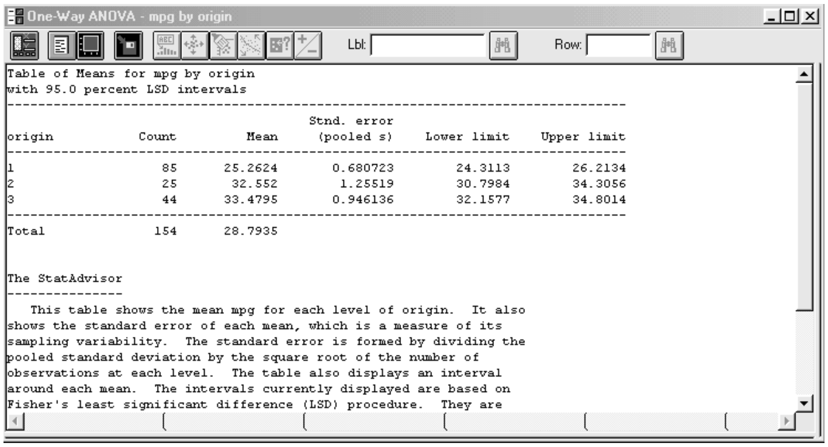
Figure 16-7. ANOVA Table

The Sum of Squares - Between Groups statistic is the measure of variability among the different samples. The Sum of Squares - Within Groups statistic is the measure of variability within each of the samples. The Total is the measure of variability for all the data around the grand mean. Each Mean Square is the Sum of Squares for the source of the variation divided by the degrees of freedom (df). The *p*-values indicate the significance level. Small significance levels (less than .05 for most practical applications) indicate that the averages of the samples differ significantly.

Table of Means

The Table of Means option shows the mean for each column of data; the pooled standard error for each mean; and the Lower and Upper limits for the confidence intervals for the means, which is a measure of sampling variability (see Figure 16-8). The pooled standard error is calculated by dividing the pooled standard deviation by the square root of the number of observations at each level.

The table also displays an interval around each mean, which is based on Fisher’s least significant difference (LSD). The intervals are constructed so that if two means are the same, their intervals will overlap a given percentage of time. You can use the intervals in the Multiple Range Tests option to determine which means are statistically different.



```
Table of Means for mpg by origin
with 95.0 percent LSD intervals
--------------------------------------------------------------------------------
                                       Stnd. error
origin           Count         Mean     (pooled s)    Lower limit    Upper limit
--------------------------------------------------------------------------------
1                   85      25.2624       0.680723        24.3113        26.2134
2                   25       32.552        1.25519        30.7984        34.3056
3                   44      33.4795       0.946136        32.1577        34.8014
--------------------------------------------------------------------------------
Total              154      28.7935


The StatAdvisor
---------------
   This table shows the mean mpg for each level of origin.  It also
shows the standard error of each mean, which is a measure of its
sampling variability.  The standard error is formed by dividing the
pooled standard deviation by the square root of the number of
observations at each level.  The table also displays an interval
around each mean.  The intervals currently displayed are based on
Fisher's least significant difference (LSD) procedure.  They are
```

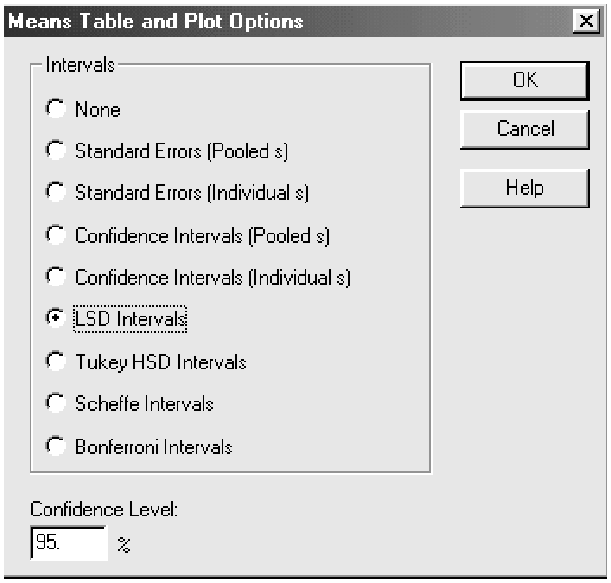
Figure 16-8. Table of Means

Use the *Means Table and Plot Options* dialog box to indicate the method and confidence level that will be used to calculate the intervals. LSD Intervals are the default for the method, and 95 percent is the default for the confidence level (see Figure 16-9).

Multiple Range Tests

The Multiple Range Tests option applies a multiple comparison analysis to the data to determine which means are significantly different (see Figure 16-10).

The top portion of the table identifies homogenous groups with columns of Xs that, within each column, indicate the group means for which there are no statistically significant differences. The bottom portion of the table shows the estimated difference between each pair of means. An asterisk indicates a statistically significant difference at a given confidence level.

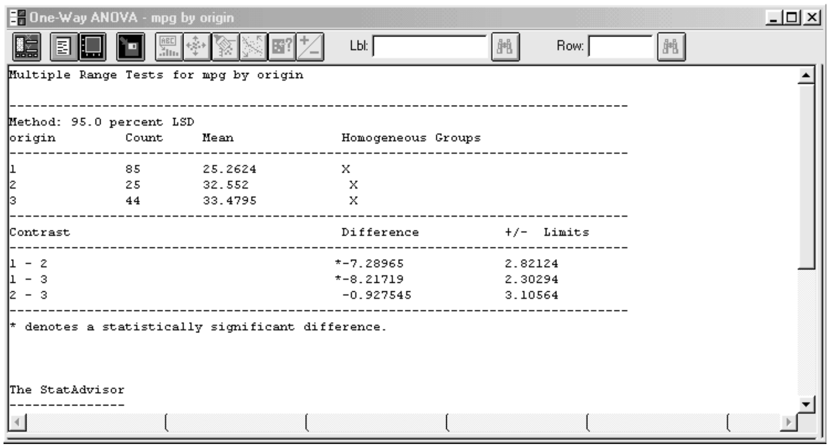


Intervals

○ None
○ Standard Errors (Pooled s)
○ Standard Errors (Individual s)
○ Confidence Intervals (Pooled s)
○ Confidence Intervals (Individual s)
● LSD Intervals
○ Tukey HSD Intervals
○ Scheffe Intervals
○ Bonferroni Intervals

Confidence Level: 95. %

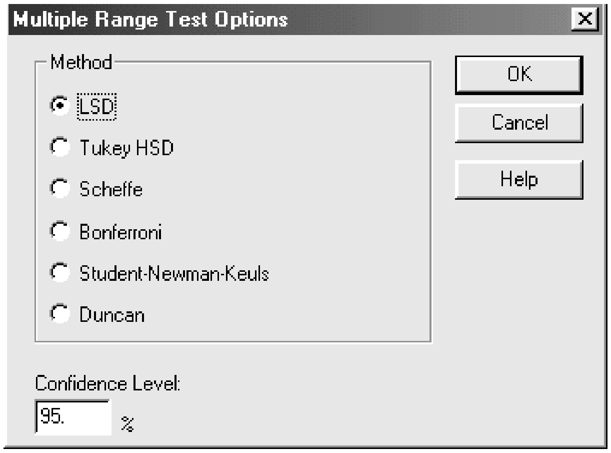
Figure 16-9. Means Table and Plot Options Dialog Box



```
Multiple Range Tests for mpg by origin

--------------------------------------------------------------------------------
Method: 95.0 percent LSD
origin         Count     Mean              Homogeneous Groups
--------------------------------------------------------------------------------
1              85        25.2624           X
2              25        32.552             X
3              44        33.4795            X
--------------------------------------------------------------------------------
Contrast                                   Difference          +/-  Limits
--------------------------------------------------------------------------------
1 - 2                                     *-7.28965            2.82124
1 - 3                                     *-8.21719            2.30294
2 - 3                                      -0.927545           3.10564
--------------------------------------------------------------------------------
* denotes a statistically significant difference.



The StatAdvisor
---------------
```

Figure 16-10. Multiple Range Tests

Use the *Multiple Range Tests Options* dialog box to choose the method and the confidence level that will be used to calculate the statistically significant differences among the means (see Figure 16-11).



Multiple Range Test Options

Method

● LSD

○ Tukey HSD

○ Scheffe

○ Bonferroni

○ Student-Newman-Keuls

○ Duncan

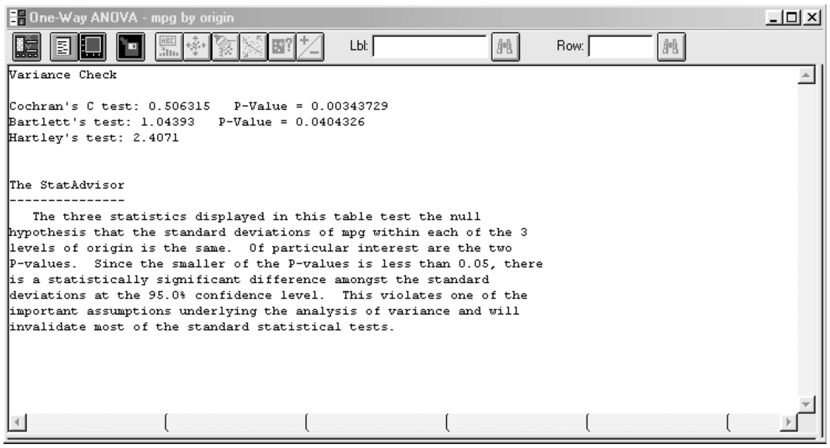
Confidence Level: 95. %

Figure 16-11. Multiple Range Tests Options Dialog Box

Variance Check

The Variance Check option tests the null hypothesis that the standard deviations within each of the samples are the same. It uses three tests to obtain the results: Cochran's C Test, Bartlett's Test, and Hartley's Test (see Figure 16-12). Each of these statistics examines the homogeneity of variances assumption that the variances are equal (homogeneous) across the cells of the between-groups design. If the *p*-value is less than a specified alpha, there is a statistically significant difference between the standard deviations at a given confidence level.

Levene's Test compares the sample variances by performing an analysis of variance on the absolute deviations of the data values from their respective sample menus. It is less sensitive than Bartlett's test to departures from normality of the underlying populations.

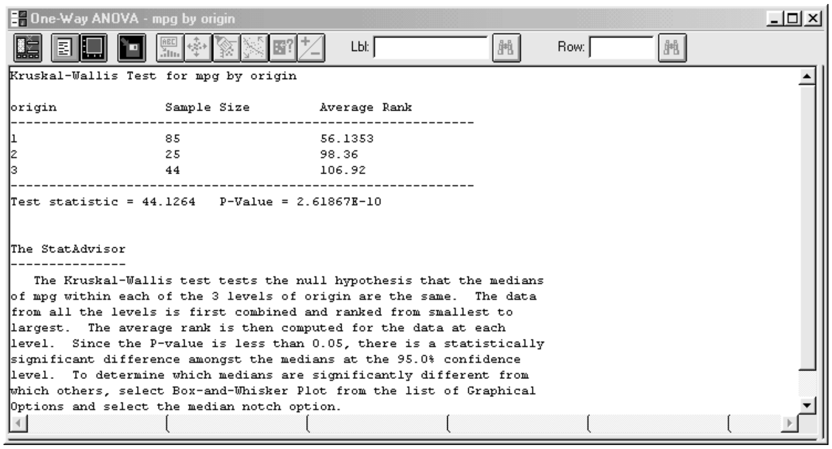


```
Variance Check

Cochran's C test: 0.506315   P-Value = 0.00343729
Bartlett's test: 1.04393   P-Value = 0.0404326
Hartley's test: 2.4071


The StatAdvisor
---------------
   The three statistics displayed in this table test the null
hypothesis that the standard deviations of mpg within each of the 3
levels of origin is the same.  Of particular interest are the two
P-values.  Since the smaller of the P-values is less than 0.05, there
is a statistically significant difference amongst the standard
deviations at the 95.0% confidence level.  This violates one of the
important assumptions underlying the analysis of variance and will
invalidate most of the standard statistical tests.
```

Figure 16-12. Variance Check

Kruskal-Wallis Test

The Kruskal-Wallis Test option is a nonparametric method that tests the assumption that the medians of the levels are equal (see Figure 16-13).



```
Kruskal-Wallis Test for mpg by origin

origin              Sample Size         Average Rank
------------------------------------------------------------
1                   85                  56.1353
2                   25                  98.36
3                   44                  106.92
------------------------------------------------------------
Test statistic = 44.1264   P-Value = 2.61867E-10


The StatAdvisor
---------------
   The Kruskal-Wallis test tests the null hypothesis that the medians
of mpg within each of the 3 levels of origin are the same.  The data
from all the levels is first combined and ranked from smallest to
largest.  The average rank is then computed for the data at each
level.  Since the P-value is less than 0.05, there is a statistically
significant difference amongst the medians at the 95.0% confidence
level.  To determine which medians are significantly different from
which others, select Box-and-Whisker Plot from the list of Graphical
Options and select the median notch option.
```

Figure 16-13. Kruskal-Wallis Test

The program first combines and ranks the values in the entire set of levels from the lowest to the highest, calculates the mean of the ranks for each level, and tests the hypothesis for the assumption that the levels are the same. If the *p*-value is less than .05, you can reject the null hypothesis at the 95 percent confidence level.

Graphical Options

Scatterplot

The Scatterplot option creates a plot of the dependent variable by levels of the factor. The points are plotted along a single axis as point symbols with no connecting lines (see Figure 16-14).

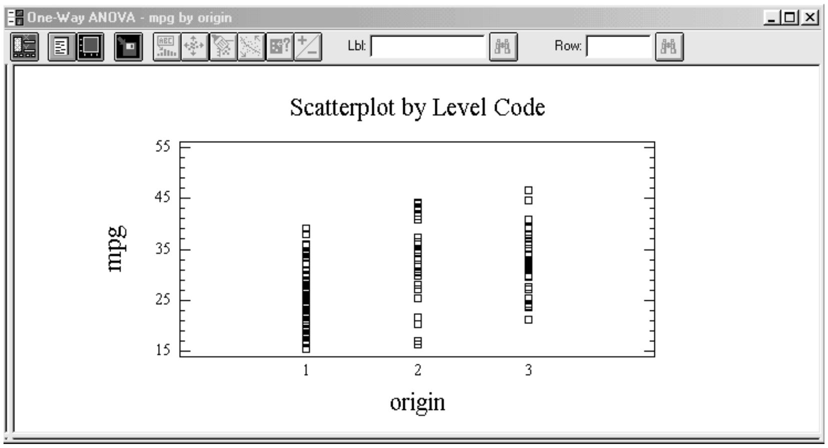


Figure 16-14. Scatterplot

Means Plot

The Means Plot option displays a plot of the means for each level and the intervals around each mean (see Figure 16-15). The intervals are constructed so that if two means are the same, their intervals will overlap a given percentage of the time. Any pair of intervals that do not overlap vertically correspond to a pair of means that have a statistically significant difference.

Use the *Means Table and Plot Options* dialog box to indicate the method and confidence level that will be used to calculate the intervals. LSD Intervals is

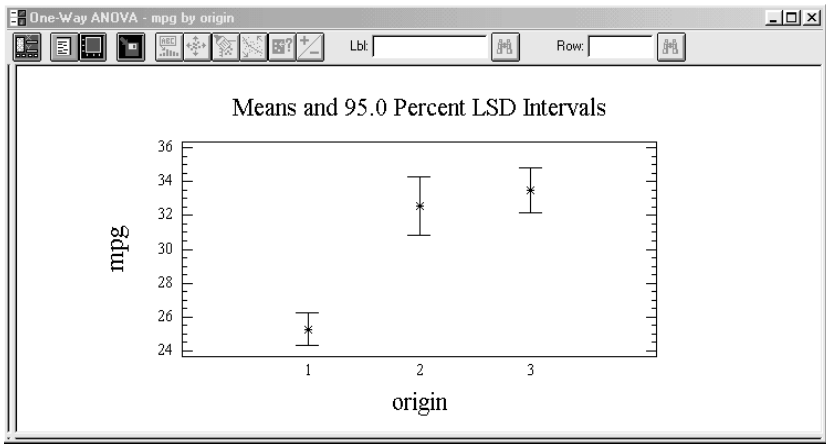

One-Way ANOVA - mpg by origin
Lbl:
Row:
Means and 95.0 Percent LSD Intervals
mpg
36
34
32
30
28
26
24
1
2
3
origin

Figure 16-15. Means Plot

the default for the method, and 95 percent is the default for the confidence level (see Figure 16-9 for an example of this dialog box).

Box-and-Whisker Plot

The Box-and-Whisker Plot option calculates a plot for each level (see Figure 16-16). The program divides the data into four areas of equal frequency (quartiles). The rectangular part of the plot extends from the lower quartile to the upper quartile, covering the center half of each sample. The center line in each box shows the location of the sample medians, while the plus (+) signs indicate the location of the sample means.

Horizontal lines, known as whiskers, extend from the box to the minimum and maximum values in each sample, except for any outside or far outside points, which are plotted separately. Outside points lie more than 1.5 times the interquartile range above or below the box and are shown as small squares. Far outside points lie more than 3.0 times the interquartile range above or below the box and are shown as small squares with plus (+) signs through them.

Use the *Box-and-Whisker Options* dialog box to indicate the direction of the plot and to choose the features that will appear on it (see Figure 16-17).

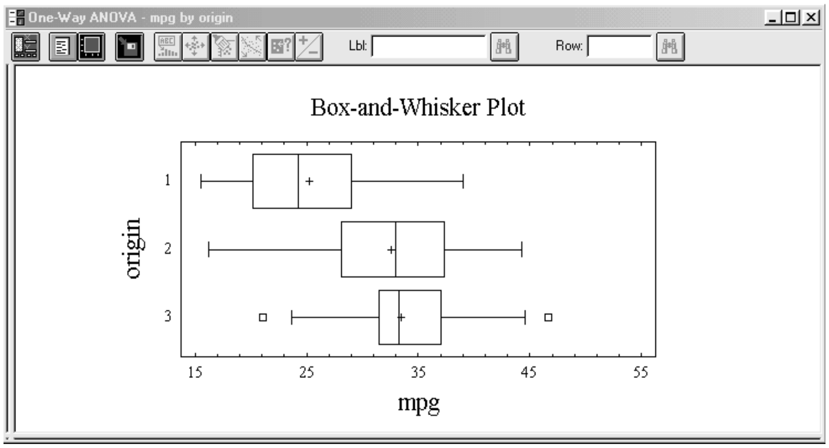
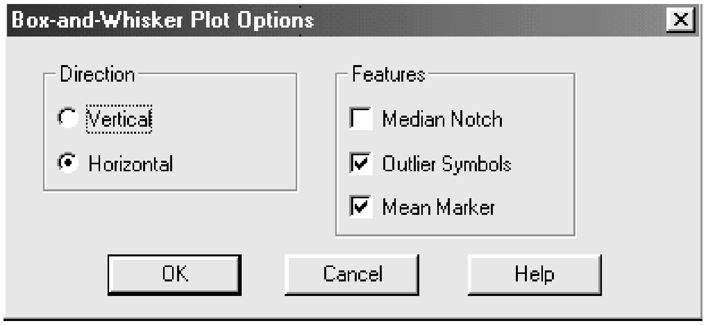


Figure 16-16. Box-and-Whisker Plot



Direction

○ Vertical

● Horizontal

Features

□ Median Notch

☑ Outlier Symbols

☑ Mean Marker

Figure 16-17. Box-and-Whisker Plot Options Dialog Box

Residuals versus Factor Levels

The Residuals versus Factor Level option creates a plot of the residuals versus the factor level (see Figure 16-18). The residuals are equal to the observed values of the dependent variable minus the mean of the dependent variable for the group from which it originates. Check the plot to ensure that the variability within each factor level is approximately the same.

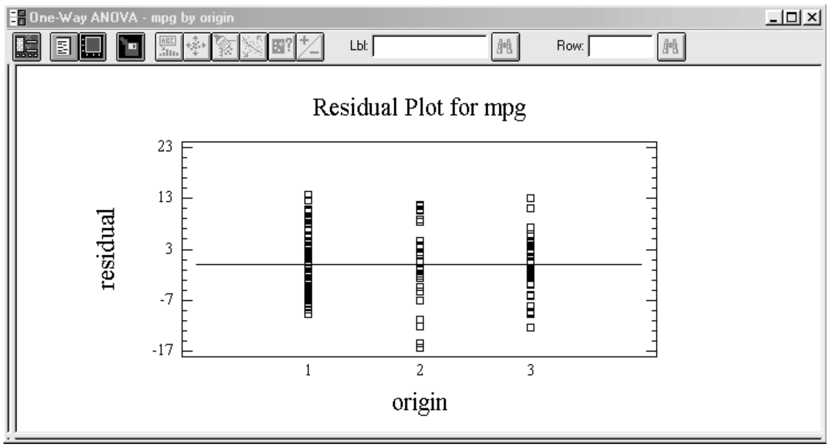


Figure 16-18. Residuals versus Factor Level Plot

Residuals versus Predicted

The Residuals versus Predicted option creates a plot of the residuals versus the predicted factor levels (see Figure 16-19). The residuals are equal to the observed values of the independent variable minus the mean of the dependent variable from the group from which it originates.

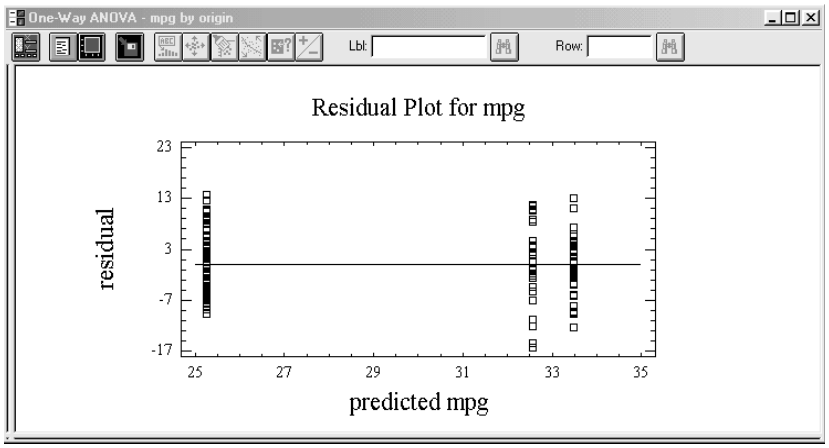


Figure 16-19. Residuals versus Predicted Plot

The plot is useful for detecting heteroscedasticity, in which the variance of the dependent variable changes together with the mean. If the points form a general funnel-shaped pattern, it indicates a type of heteroscedasticity that can often be corrected by transforming the dependent variable. You could try using the Analysis dialog box to enter LOG(*n*), SQRT(*n*), or 1/*n*, where *n* equals the name of the dependent variable.

Residuals versus Row Number

The Residuals versus Row Number option creates a plot of the residuals versus the row number (see Figure 16-20). The residuals appear on the plot in the order in which you entered them into the DataSheet. Any pattern other than a random scatter could indicate some serial correlation or time dependence in the data.

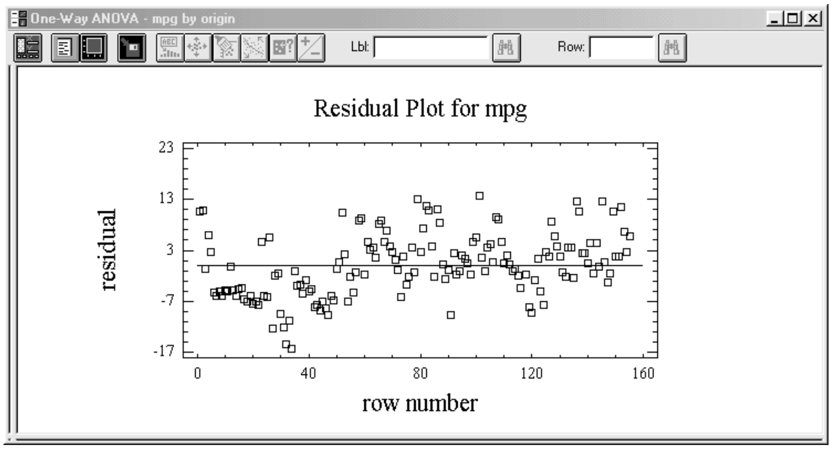
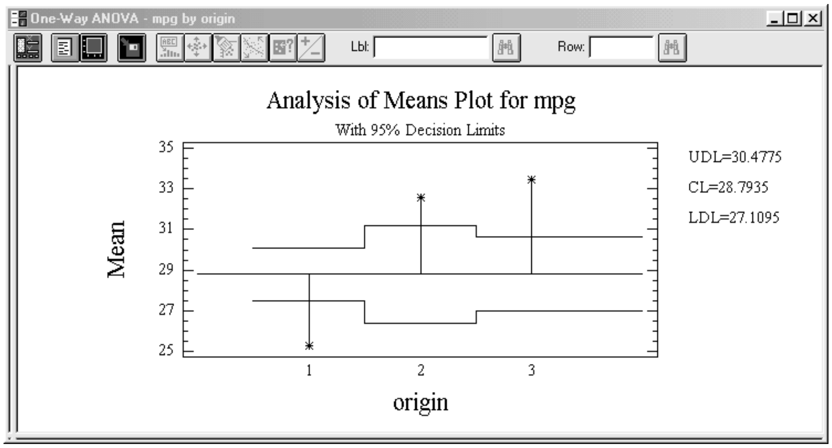


Figure 16-20. Residuals versus Row Number Plot

Analysis of Means (ANOM) Plot

The Analysis of Means (ANOM) Plot option creates a plot that shows the mean for each of the samples, the centerline at the grand mean, and the decision limits (see Figure 16-21). The decision limits determine the groups that differ significantly at the grand mean.

Use the *Analysis of Means Plot Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals and to



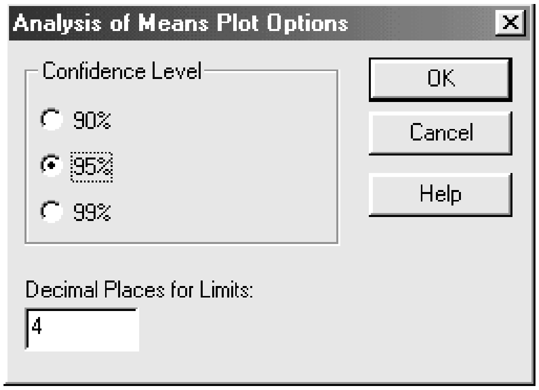
UDL=30.4775

CL=28.7935

LDL=27.1095

Figure 16-21. Analysis of Means (ANOM) Plot

enter the number of decimal places that will be used for the decision limits (see Figure 16-22).



Confidence Level

○ 90%

● 95%

○ 99%

Decimal Places for Limits: 4

Figure 16-22. Analysis of Means Plot Options Dialog Box

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are eight selections: Level Counts, Level Means, Level Sigmas, Level Standard Errors, Level Labels, Level Indicators, Residuals, and Ranges.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. New York: Wiley.

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: Wiley.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Hollander, M. and Wolfe, D. A. 1973. *Nonparametric Statistical Methods*. New York: Wiley.

Neter, J., Wasserman, W., and Kutner, M. 1985. *Applied Linear Statistical Models*. Homewood, Illinois: Richard E. Irwin, Inc.

Owen, D. B. 1962. *Handbook of Statistical Tables*. Redding, Massachusetts: Addison-Wesley.

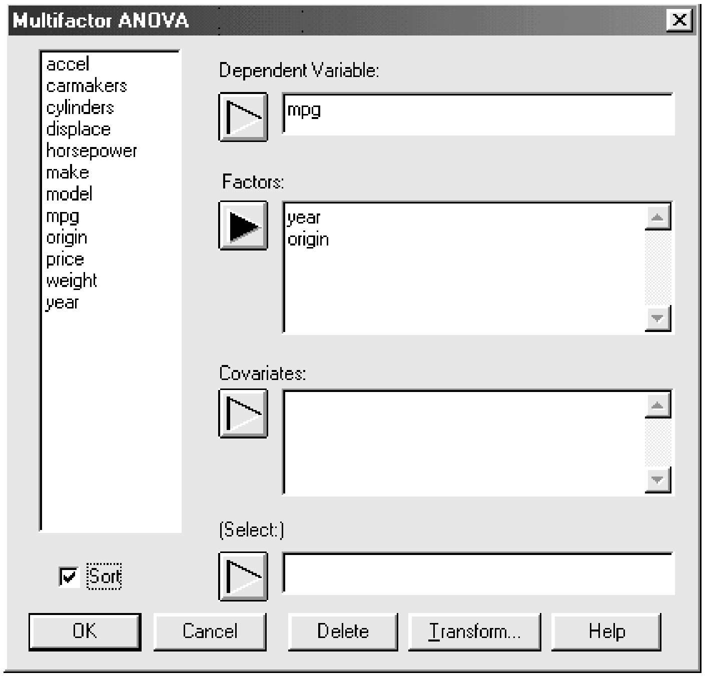
Using Multifactor ANOVA

Multifactor ANOVA is basically the same as One-Way ANOVA except that it examines the effects of two or more factors on one variable, and you typically use data you collect for a designed experiment. The design can be balanced or unbalanced and can also include covariates.

You can use the default values and perform a standard analysis, or you can customize the analysis in a variety of ways, such as using either Type I or Type III sums of squares, or selecting the order of interactions.

The Multifactor ANOVA Analysis is useful in a range of disciplines. For example, in marketing studies you might test how product sales are affected by alternative pricing, advertising techniques, and different packaging options. In medical studies, you might test how disease is combated by using different combinations of drugs, varying dosages of drugs, and alternative treatment schedules. In agricultural studies, you might measure how weight gain and the overall health of cattle is affected by brands of feed, amounts of medication, and type of breed.

To access the analysis, from the menus, choose: COMPARE... ANALYSIS OF VARIANCE... MULTIFACTOR ANOVA... (see Figure 16-23).



Multifactor ANOVA

Dependent Variable: mpg

Factors: year, origin

Covariates:

(Select:)

☑ Sort

Figure 16-23.   The Multifactor ANOVA Dialog Box

Tabular Options

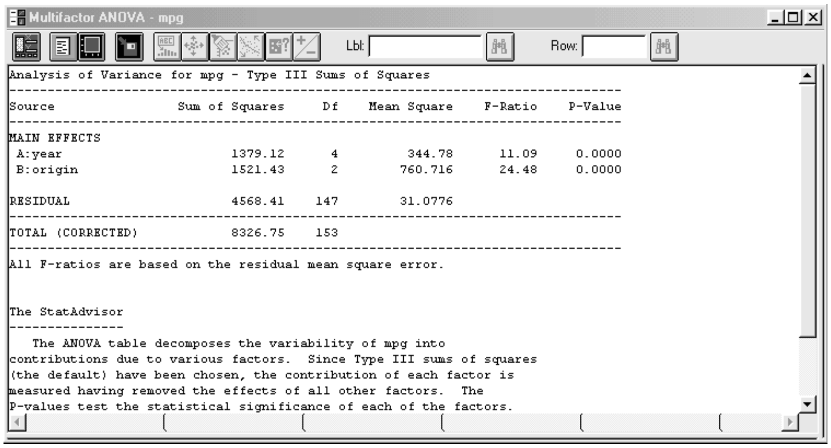
Analysis Summary

The Analysis Summary option creates a summary of the analysis that contains the name of the dependent variable and the factors. It also displays the number of complete cases in the analysis.

Use the *Multifactor ANOVA Options* dialog box to enter the number of highest order interactions that will be estimated.

ANOVA Table

The ANOVA Table option creates a standard analysis of variance table (see Figure 16-24). The table includes the residual sums of squares, the degrees of freedom (Df) for the residual, the residual mean square, an F-ratio that is calculated using the residual mean square (or the error term you choose), and the *p*-value for the F-ratio.

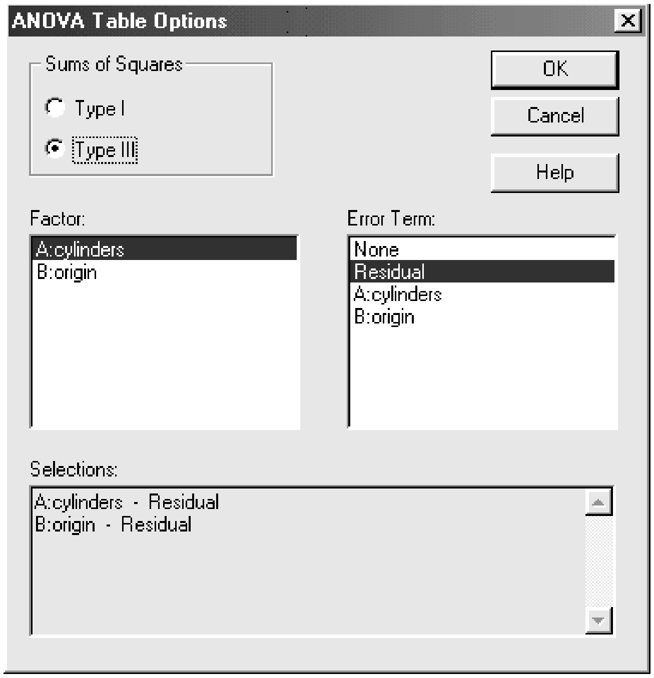


```
Analysis of Variance for mpg - Type III Sums of Squares
--------------------------------------------------------------------------------
Source             Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:year                   1379.12      4         344.78      11.09     0.0000
 B:origin                 1521.43      2        760.716      24.48     0.0000

RESIDUAL                  4568.41    147        31.0776
--------------------------------------------------------------------------------
TOTAL (CORRECTED)         8326.75    153
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.


The StatAdvisor
---------------
   The ANOVA table decomposes the variability of mpg into
contributions due to various factors.  Since Type III sums of squares
(the default) have been chosen, the contribution of each factor is
measured having removed the effects of all other factors.  The
P-values test the statistical significance of each of the factors.
```

Figure 16-24. The ANOVA Table

The table includes statistics for each main effect and interaction in the analysis. Notes at the bottom of the table identify the error term that was used as the denominator to calculate each F-ratio. Until you use the ANOVA Table Options dialog box to change the default, the table is initially calculated for Type III sums of squares.

Use the *ANOVA Table Options* dialog box to indicate which method will be used to calculate the sums of squares for each factor, to choose a factor for which you want to choose an Error term, to indicate the type of mean square that will be used as the denominator to calculate the F-ratio, and to view the error term (see Figure 16-25).



ANOVA Table Options

Sums of Squares
○ Type I
● Type III

Factor:
A:cylinders
B:origin

Error Term:
None
Residual
A:cylinders
B:origin

Selections:
A:cylinders - Residual
B:origin - Residual

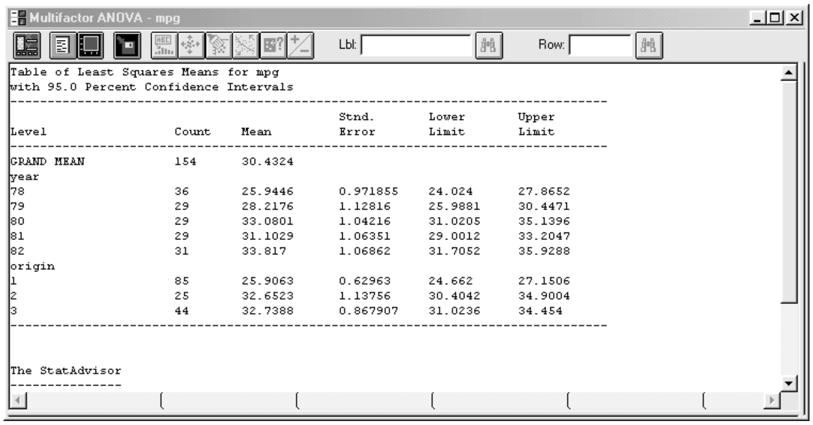
Figure 16-25. ANOVA Table Options Dialog Box

Table of Means

The Table of Means option creates a table that shows the mean of the variable for each level of the factors (see Figure 16-26). The table includes the number of observations (Count) at each level, the mean, the standard error, and the lower and upper limits for the confidence intervals.

In a balanced design, the least squares means are equal to the arithmetic average; in an unbalanced design, the least squares means differ. Using the arithmetic average in an unbalanced design instead of least squares means is not statistically valid because assumptions are made that are not true. See Hoaglin, et al. (1991) for more information.

Use the *Confidence Intervals Options* dialog box to enter a value for the confidence level that will be used to calculate the confidence intervals for the least squares means. The default is 95 percent; other common levels are 90 and 99.



```
Table of Least Squares Means for mpg
with 95.0 Percent Confidence Intervals
--------------------------------------------------------------------------------
                                           Stnd.        Lower        Upper
Level                 Count    Mean        Error        Limit        Limit
--------------------------------------------------------------------------------
GRAND MEAN            154      30.4324
year
78                    36       25.9446     0.971855     24.024       27.8652
79                    29       28.2176     1.12816      25.9881      30.4471
80                    29       33.0801     1.04216      31.0205      35.1396
81                    29       31.1029     1.06351      29.0012      33.2047
82                    31       33.817      1.06862      31.7052      35.9288
origin
1                     85       25.9063     0.62963      24.662       27.1506
2                     25       32.6523     1.13756      30.4042      34.9004
3                     44       32.7388     0.867907     31.0236      34.454
--------------------------------------------------------------------------------


The StatAdvisor
---------------
```

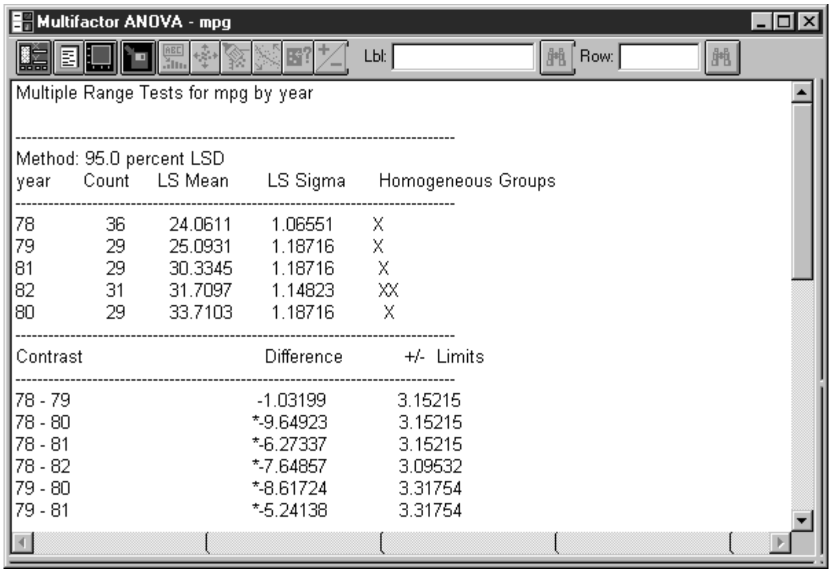
Figure 16-26.   Table of Means

Multiple Range Tests

The Multiple Range Tests option applies a multiple comparison analysis to the data to determine which means are significantly different (see Figure 16-27).

The top portion of the table identifies homogenous groups with columns of Xs that, within each column, indicate the group means for which there are no statistically significant differences. The bottom portion of the table shows the estimated difference between each pair of means. An asterisk indicates a statistically significant difference at a given confidence level; this statistic will vary depending on the method you are using.

Use the *Multiple Range Tests Options* dialog box to choose the method and the confidence level that will be used to calculate the statistically significant differences among the means (see Figure 16-11 for an example of this dialog box).



```
Multiple Range Tests for mpg by year

--------------------------------------------------------------------------------
Method: 95.0 percent LSD
year     Count    LS Mean      LS Sigma     Homogeneous Groups
--------------------------------------------------------------------------------
78        36      24.0611      1.06551      X
79        29      25.0931      1.18716      X
81        29      30.3345      1.18716       X
82        31      31.7097      1.14823       XX
80        29      33.7103      1.18716        X
--------------------------------------------------------------------------------
Contrast                        Difference        +/-  Limits
--------------------------------------------------------------------------------
78 - 79                         -1.03199          3.15215
78 - 80                         *-9.64923         3.15215
78 - 81                         *-6.27337         3.15215
78 - 82                         *-7.64857         3.09532
79 - 80                         *-8.61724         3.31754
79 - 81                         *-5.24138         3.31754
```

Figure 16-27.   Multiple Range Tests

Graphical Options

Scatterplot

The Scatterplot option creates a plot of the values for the response variable at each of the factor levels (see Figure 16-28). The values are plotted as point symbols with no connecting lines along a single horizontal axis.

Use the *Scatterplot Options* dialog box to choose the factor that will be used in the analysis.

Means Plot

The Means Plot option creates a plot of the means for each factor level and the intervals around each mean (see Figure 16-29). The intervals are constructed so that if two means are the same, their intervals will overlap a given percentage of the time. Any pair of intervals that do not overlap vertically correspond to a pair of means that have a statistically significant difference.

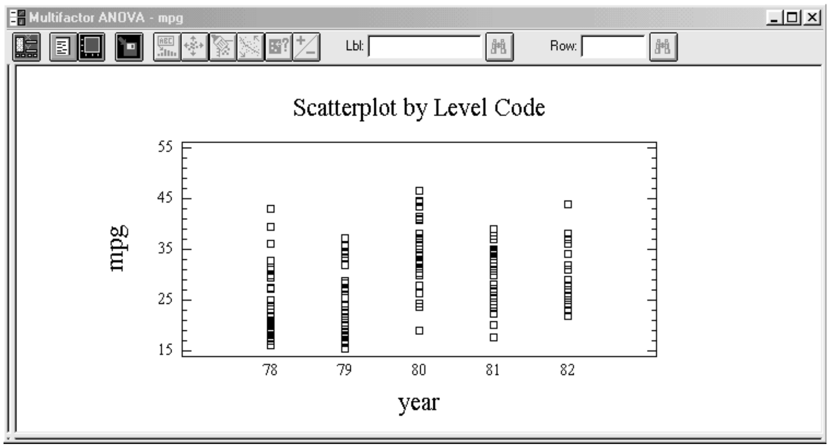


Figure 16-28. Scatterplot

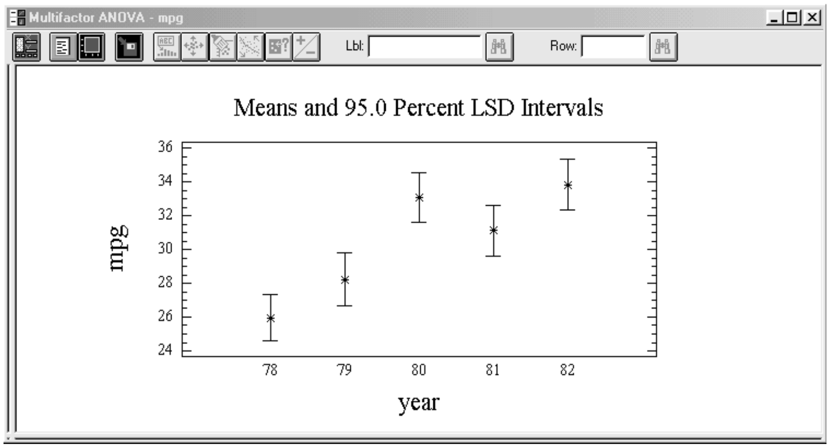
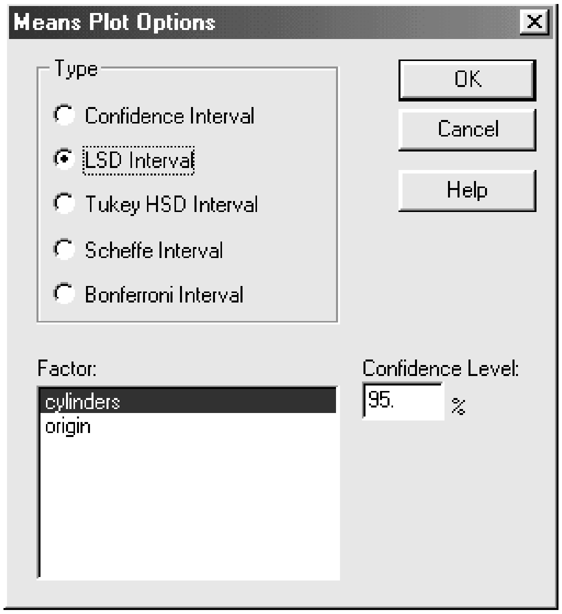


Figure 16-29. Means Plot

Use the *Means Plot Options* dialog box to indicate the interval, factor, and confidence level that will be used to calculate the intervals. LSD Intervals is the default for the method, and 95 percent is the default for the confidence level (see Figure 16-30).



Type

○ Confidence Interval
● LSD Interval
○ Tukey HSD Interval
○ Scheffe Interval
○ Bonferroni Interval

Factor:
cylinders
origin

Confidence Level: 95. %

Figure 16-30. Means Plot Options Dialog Box

Interaction Plot

The Interaction Plot option creates a plot that shows any two-factor interactions that were estimated in the analysis using the current model (see Figure 16-31).

Use the *Interaction Plot Options* dialog box to indicate which model will be used to determine the two-factor interactions; to choose an interaction; to enter a value for the confidence level that will be used to calculate the confidence interval; and to indicate the factor in the interaction that will be plotted against the response variable (see Figure 16-32).

Residuals versus Factor Level

The Residuals versus Factor Level option creates a plot of the residuals versus the factor level (see Figure 16-33). The residuals are equal to the observed values minus the mean for the group from which they originate. Check the plot to ensure that the variability within each level is approximately the same.

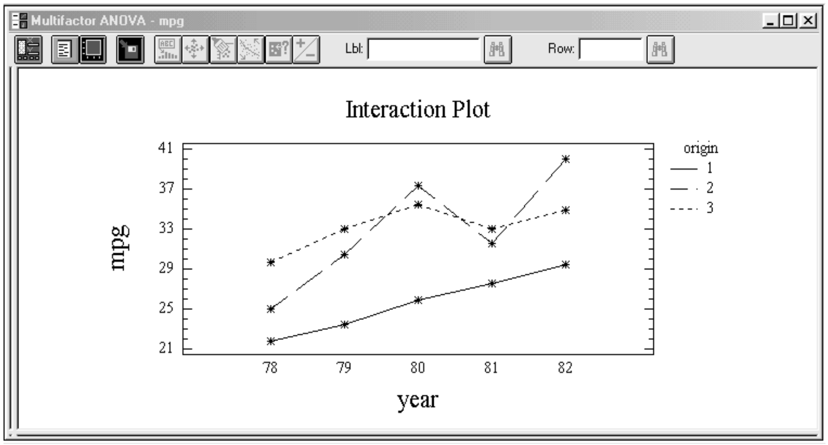
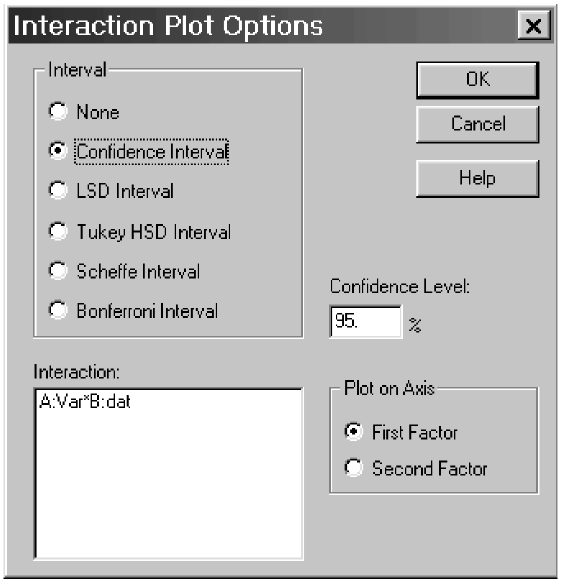


Figure 16-31. Interaction Plot



Interval

○ None
● Confidence Interval
○ LSD Interval
○ Tukey HSD Interval
○ Scheffe Interval
○ Bonferroni Interval

Confidence Level: 95. %

Interaction: A:Var*B:dat

Plot on Axis

● First Factor
○ Second Factor

Figure 16-32. Interaction Plot Options Dialog Box

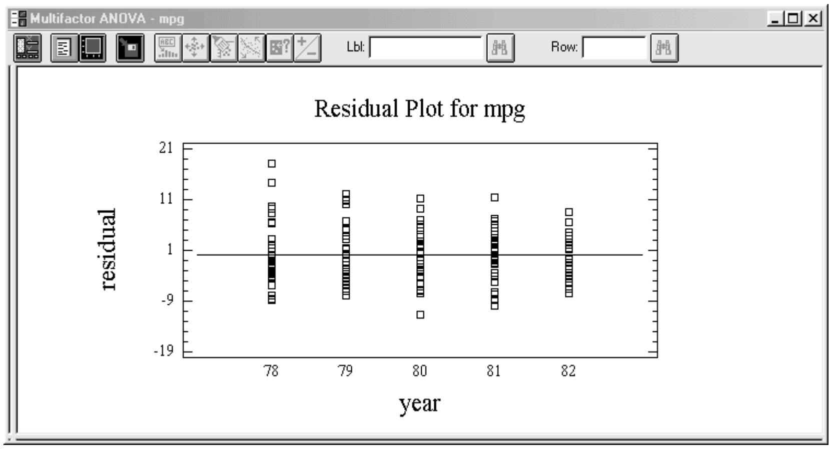


Figure 16-33.   Residuals versus Factor Level Plot

Use the *Residuals Plot Options* dialog box to choose the factor that will be plotted against the residuals.

Residuals versus Predicted

The Residuals versus Predicted option creates a plot of the residuals versus the predicted levels (see Figure 16-34). The plot is helpful for detecting heteroscedasticity, in which the variance changes together with the mean. If the points form a general funnel-shaped pattern, there is an indication of a type of heteroscedasticity that can often be corrected by transforming the dependent variable. Try using the Analysis dialog box to enter LOG(*n),* SQRT(*n*), or 1/*n* as the dependent variable.

Residuals versus Row Number

The Residuals versus Row Number option creates a plot of the residuals versus the row number (see Figure 16-35). The residuals appear on the plot in the order you entered the observations into the DataSheet. Any pattern other than a random scatter could indicate some serial correlation or time dependence in the data.

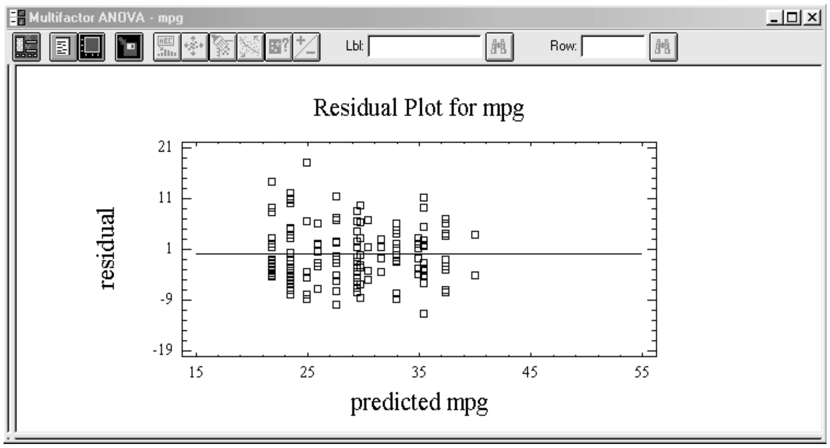


Figure 16-34. Residuals versus Predicted Plot

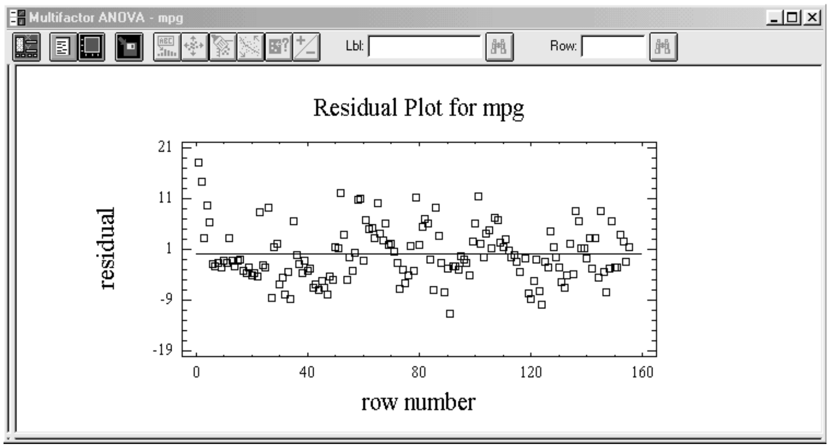


Figure 16-35. Residuals versus Row Number Plot

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are five selections: Level Counts, Level Means, Level Standard Errors, Least Squares Means, and Residuals.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. New York: Wiley.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., eds. 1991. *Fundamentals of Exploratory Analysis of Variance*. New York: Wiley.

Milliken, G. A. and Johnson, D. E. 1984. *Analysis of Messy Data, Vol. I: Designed Experiments*. New York: Van Nostrand Reinhold Company.

Montgomery, D. C. 1991. *Design and Analysis of Experiments*, third edition. New York: Wiley.

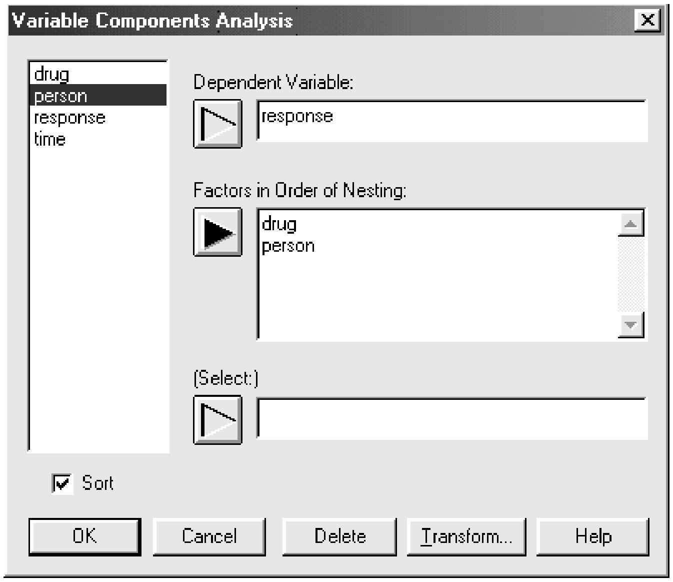
Neter, J., Wasserman, W., and Kutner, M. 1985. *Applied Linear Statistical Models*, second edition. Homewood, Illinois: Richard E. Irwin, Inc.

Using Variance Components Analysis

The Variance Components Analysis uses fully nested, random effects models, which are classification effects where the levels of the effect are assumed to be randomly selected from an infinite population of possible levels. The analysis estimates the contribution each of the random effects makes to the variance of the dependent variable, and analyzes the effect of one or more qualitative factors on one response variable. The number of observations do not need to be equal at all combinations of the factor levels.

An example of a nested design would be soil samples in an agricultural or geological study taken from five geographic regions, with subsamples taken from 12 areas within each region. Another example, a study of worker productivity, samples workers by selecting 10 factories at random, then randomly selects four departments within each factory, then again randomly selects five workers from each department.

To access the analysis, from the menus, choose: COMPARE... ANALYSIS OF VARIANCE... VARIANCE COMPONENTS... (see Figure 16-36).



Dependent Variable: response

Factors in Order of Nesting: drug, person

(Select:)

☑ Sort

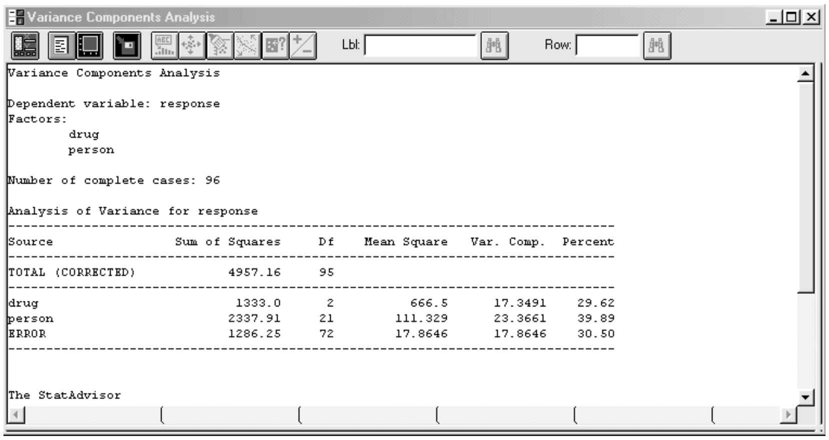
Figure 16-36. The Variance Components Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a variance table where the variance of the dependent variable is divided into one component for each factor (see Figure 16-37). The goal of the analysis is usually to estimate the amount of variability each of the factors contributes.

Each factor after the first one is nested in the factor above. The Percent column contains the values for the percentage of variance each factor accounts for. If you omit the last factor, a line on the table appears with the label “ERROR.”

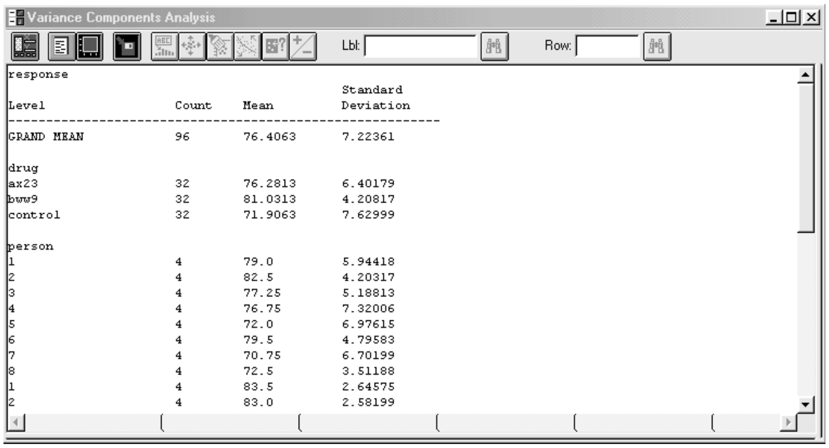


```
Variance Components Analysis

Dependent variable: response
Factors:
        drug
        person

Number of complete cases: 96

Analysis of Variance for response
--------------------------------------------------------------------------------
Source                Sum of Squares     Df    Mean Square   Var. Comp.  Percent
--------------------------------------------------------------------------------
TOTAL (CORRECTED)            4957.16     95
--------------------------------------------------------------------------------
drug                          1333.0      2          666.5      17.3491    29.62
person                       2337.91     21        111.329      23.3661    39.89
ERROR                        1286.25     72        17.8646      17.8646    30.50
--------------------------------------------------------------------------------


The StatAdvisor
```

Figure 16-37.   Analysis Summary

Summary Statistics

The Summary Statistics option creates summary statistics for the count, mean, and standard deviation of the data (see Figure 16-38). It also shows the grand mean, which is the mean of all the dependent variables.



```
response
                                             Standard
Level                 Count    Mean          Deviation
---------------------------------------------------------
GRAND MEAN            96       76.4063       7.22361

drug
ax23                  32       76.2813       6.40179
bww9                  32       81.0313       4.20817
control               32       71.9063       7.62999

person
1                     4        79.0          5.94418
2                     4        82.5          4.20317
3                     4        77.25         5.18813
4                     4        76.75         7.32006
5                     4        72.0          6.97615
6                     4        79.5          4.79583
7                     4        70.75         6.70199
8                     4        72.5          3.51188
1                     4        83.5          2.64575
2                     4        83.0          2.58199
```

Figure 16-38.   Summary Statistics

Graphical Options

Scatterplot

The Scatterplot option creates a plot of the variance in the dependent variable between the levels of the independent variables. The horizontal line represents the average values of the dependent variable. The points represent the average value for one independent variable nested within the other (see Figure 16-39).

Use the *Scatterplot Options* dialog box to choose a factor for the X-axis.

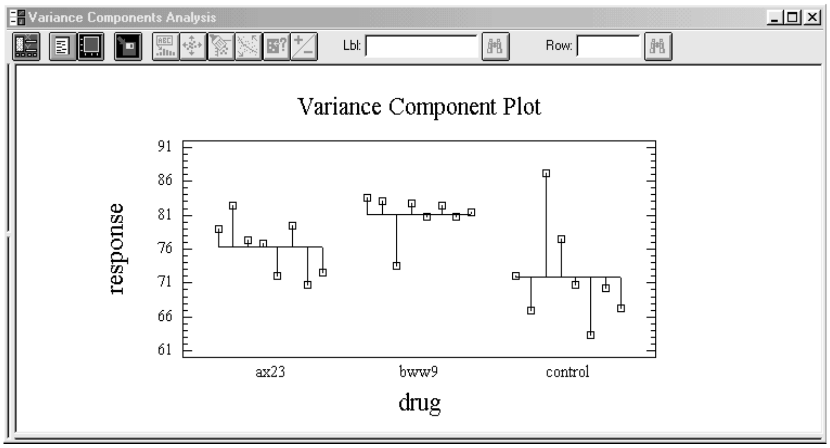


Figure 16-39. Scatterplot

Residuals versus Factor Level

The Residuals versus Factor Level option creates a plot that shows the residuals versus the levels of the independent variable (see Figure 16-40). The residuals are equal to the observed values minus the predicted values. Check this plot to ensure that the variability within each level is approximately the same.

Use the *Residuals Plot Options* dialog box to choose a factor for the X-axis.

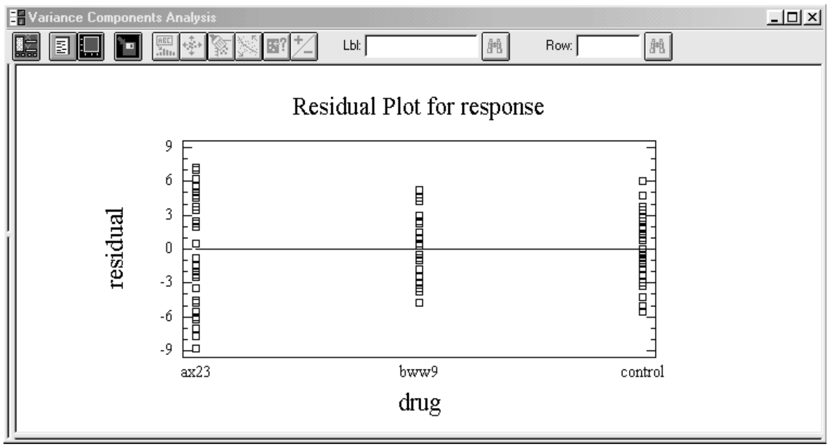


Figure 16-40. Residuals versus Factor Level Plot

Residuals versus Predicted

The Residuals versus Predicted option creates a plot of the residuals versus the predicted level of the dependent variable (see Figure 16-41). The plot is useful for detecting heteroscedasticity, which changes the variance of the dependent variable and the mean. If the points form a general funnel-shaped pattern, it is an indication of a type of heteroscedasticity that can often be corrected by transforming the data. When this happens, you can use the Analysis dialog box to enter LOG(*n*), SQRT(*n*), or 1/*n*, where *n* equals the name of the dependent variable.

Residuals versus Row Number

The Residuals versus Row Number option displays a plot of the residuals versus their row number (see Figure 16-42). Any pattern other than a random scatter could indicate some serial correlation or time dependence in the data.

Saving the Results

The Save Results Options dialog box allows you to choose the results you want to save. There are two selections: Variance Components and Residuals.

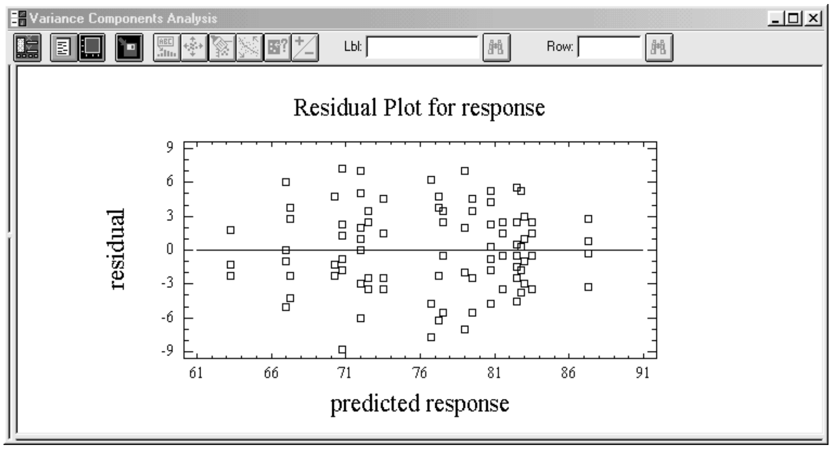


Figure 16-41. Residuals versus Predicted Plot

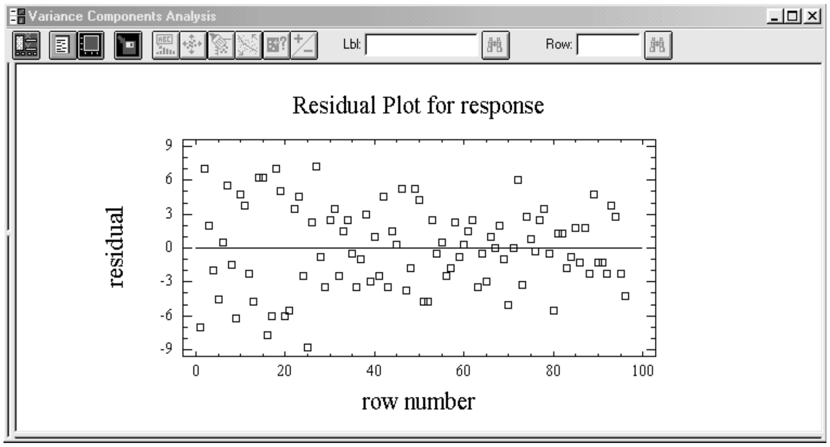


Figure 16-42. Residuals versus Row Number Plot

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. New York: Wiley.

Kempthorne, O. and Folks, L. 1971. *Probability, Statistics, and Data Analysis*. Ames, Iowa: Iowa University Press.

Searle, S. R., Casella, G., and McCulloch, C. E. 1992. *Variance Components*. In Wiley Series in *Probability and Mathematical Statistics: Applied Probability and Statistics Section*, ed. by

Barnett, V., Bradley, R. A., Fisher, N. I., Hunter, J. S., Kadane, J. B., Kendall, D. G., Smith, A. F. M., Stigler, S. M., Teugels, J., and Watson, G. S. New York: John Wiley & Sons, Inc.

17 Performing Regression Analysis

STATGRAPHICS *Plus* provides four techniques for modeling the relationship between dependent and independent variables: Simple Regression, Polynomial Regression, Box-Cox Transformations, and Multiple Regression. Modeling quantifies the linear relationship between the variables and measures the strength of the relationship.

- **Simple Regression**
  The Simple Regression Analysis fits a model that relates one dependent variable to one independent variable by minimizing the sum of the squares of the residuals for the fitted line.

- **Polynomial Regression**
  The Polynomial Regression Analysis fits a model between a single dependent variable Y and a single independent variable X. The analysis allows you to fit polynomial models up to Order 8.

- **Box-Cox Transformations**
  The Box-Cox Transformations Analysis finds the transformation parameter that minimizes the mean squared error of the fitted model.

- **Multiple Regression**
  The Multiple Regression Analysis analyzes the relationship among one dependent variable and one or more independent variables.

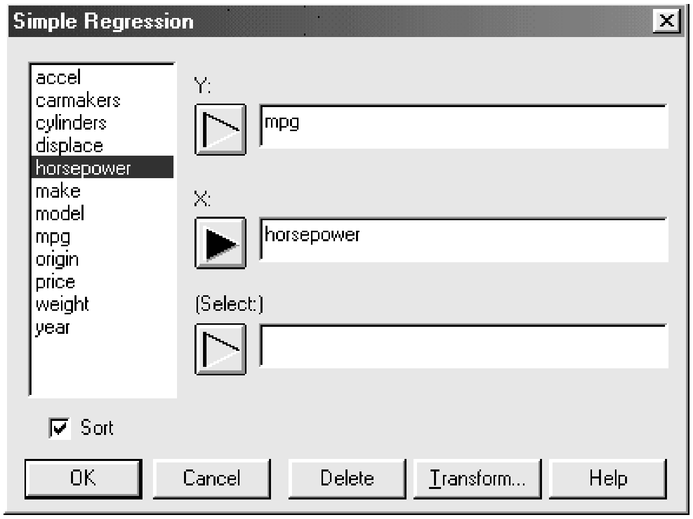
Using Simple Regression

The Simple Regression Analysis estimates a linear or nonlinear regression between two variables. The model relates one dependent variable to one independent variable by minimizing the sum of the squares of the residuals for the fitted line. In addition to a straight line, you can also use the analysis to estimate a number of standard nonlinear models. The analysis calculates and compares the goodness of fit for each model.

The analysis can automatically fit any of 12 models. The preferred model is the linear model unless another model type significantly increases the R-Squared value. To select the best model, you must fit the data to all the models and compare the statistics. You can also use the Comparison of Alternative Models tabular option to compare correlation coefficients and R-Squared values for several transformed model types.

You can plot the fitted line and the residuals for all the models. STATGRAPHICS *Plus* also creates and plots predictions for given values of X. You can also save the residuals and predictions for use in future analyses.

To access the analysis, from the menus, choose: RELATE... SIMPLE REGRESSION... (see Figure 17-1).



Y: mpg

X: horsepower

(Select:)

☑ Sort

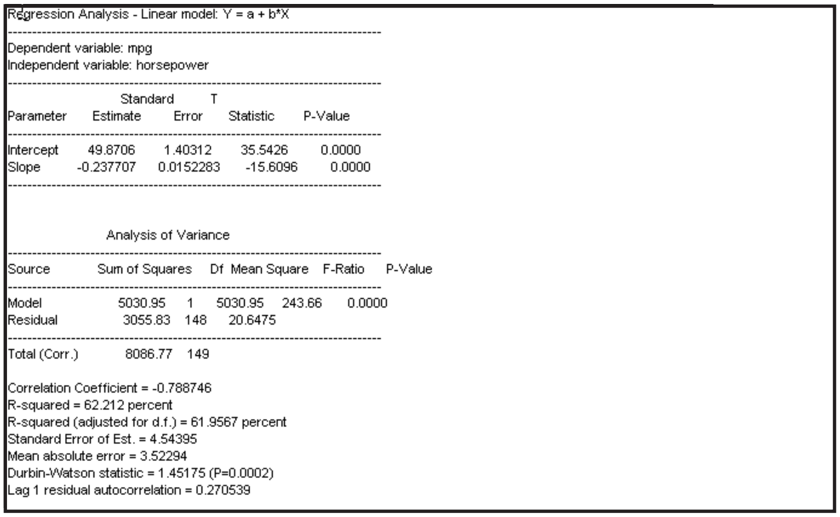
Figure 17-1. The Simple Regression Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which shows the results of fitting a linear model that describes the relationship between the two variables. The equation of the fitted linear model is

Y = a + b*X (see Figure 17-2). In the regression model, the estimate for the intercept is the value for “a”; the estimate for the slope is the value of “b”.



```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: mpg
Independent variable: horsepower
-----------------------------------------------------------------------------
                         Standard        T
Parameter     Estimate      Error     Statistic      P-Value
-----------------------------------------------------------------------------
Intercept      49.8706      1.40312      35.5426       0.0000
Slope       -0.237707    0.0152283     -15.6096        0.0000
-----------------------------------------------------------------------------


                     Analysis of Variance
-----------------------------------------------------------------------------
Source           Sum of Squares    Df  Mean Square   F-Ratio     P-Value
-----------------------------------------------------------------------------
Model                 5030.95     1     5030.95     243.66      0.0000
Residual               3055.83   148     20.6475
-----------------------------------------------------------------------------
Total (Corr.)          8086.77   149

Correlation Coefficient = -0.788746
R-squared = 62.212 percent
R-squared (adjusted for d.f.) = 61.9567 percent
Standard Error of Est. = 4.54395
Mean absolute error = 3.52294
Durbin-Watson statistic = 1.45175 (P=0.0002)
Lag 1 residual autocorrelation = 0.270539
```

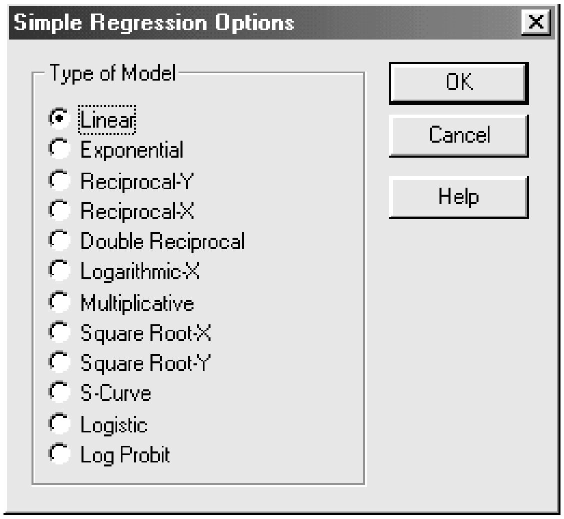
Figure 17-2. Analysis Summary

The summary includes the standard error, *t*-statistic, and *p*-value for each estimate. The *t*-statistic tests to see if the true value of the coefficient is equal to 0, which is equivalent to concluding that there is no linear relationship between the independent and dependent variables. Low probability levels (less than .05 for a 95 percent confidence level) for the intercept and slope suggest significant values.

The summary also includes an ANOVA Table for the model. The F-ratio in the summary indicates the significance of the results; a high F-ratio suggests a significant model. The summary also shows the correlation coefficient that measures the linear relationship between the independent and dependent variables, the R-Squared value (the square of the correlation coefficient, expressed as a percentage), and the standard error of estimation. The standard error of estimation is the square root of the residual mean square, which is the estimated standard deviation of the variable that is not explained by the estimated model. The R-Squared value shows the percentage of variability in the Y values explained by the X variable.

In addition, the summary now includes statistics for the following: Adjusted R-squared, MAE (Main Absolute Error), Durbin-Watson, and Lag 1 autocorrelation.

Use the *Simple Regression Options* dialog box to select the type of model that will be fit (see Figure 17-3). *See Online Help for a description of all the models as well as the equation.*



Type of Model

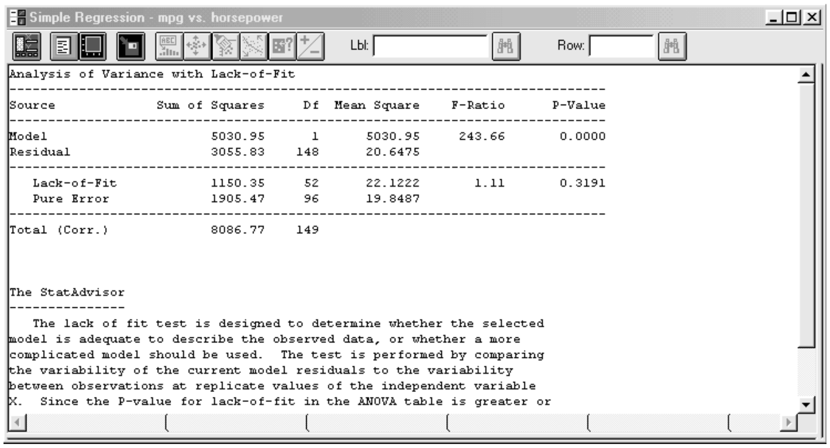
● Linear
○ Exponential
○ Reciprocal-Y
○ Reciprocal-X
○ Double Reciprocal
○ Logarithmic-X
○ Multiplicative
○ Square Root-X
○ Square Root-Y
○ S-Curve
○ Logistic
○ Log Probit

Figure 17-3. Simple Regression Options Dialog Box

Lack-of-Fit Test

The Lack-of-Fit Test option determines if the current regression model is adequate to fit the data or if you should select a more complicated model (see Figure 17-4). The test is performed by comparing the variability for the residuals of the current model with the variability between observations at replicate values of the independent variable X.

The test assumes that the observations Y for a given X are independent and normally distributed, and that the distributions of Y have the same variance. Significant lack-of-fit shows that the specified model does not adequately fit the response. The test requires replicated observations at one or more X values. A *p*-value less than .05 indicates that the model does not adequately fit the data at the 95 percent confidence level.

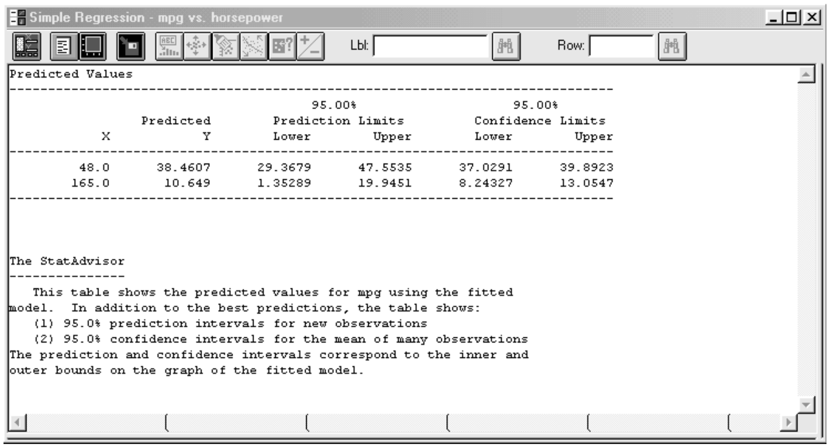


```
Analysis of Variance with Lack-of-Fit
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     5030.95      1      5030.95     243.66       0.0000
Residual                  3055.83    148      20.6475
-----------------------------------------------------------------------------
   Lack-of-Fit            1150.35     52      22.1222       1.11       0.3191
   Pure Error             1905.47     96      19.8487
-----------------------------------------------------------------------------
Total (Corr.)             8086.77    149



The StatAdvisor
---------------
   The lack of fit test is designed to determine whether the selected
model is adequate to describe the observed data, or whether a more
complicated model should be used.  The test is performed by comparing
the variability of the current model residuals to the variability
between observations at replicate values of the independent variable
X.  Since the P-value for lack-of-fit in the ANOVA table is greater or
```

Figure 17-4.   Lack-of-Fit Test

Forecasts

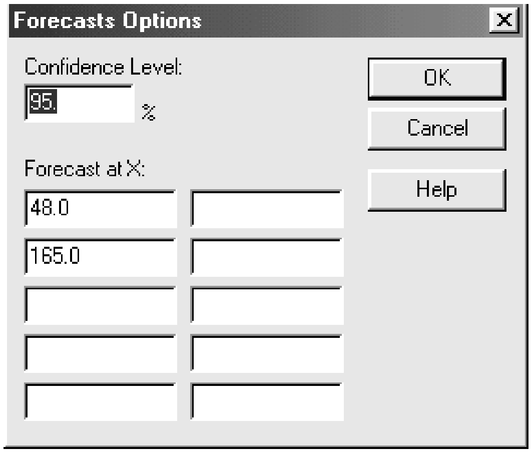
The Forecasts option creates a table of the predicted values for Y using the fitted model. The table displays the values for the prediction limits and confidence limits. The prediction and confidence intervals correspond to the Inner and Outer bounds on the plot of the fitted model (see Figure 17-5).



```
Predicted Values
------------------------------------------------------------------------------
                                      95.00%                    95.00%
                  Predicted        Prediction Limits         Confidence Limits
            X             Y        Lower        Upper        Lower        Upper
------------------------------------------------------------------------------
         48.0       38.4607      29.3679      47.5535      37.0291      39.8923
        165.0        10.649      1.35289      19.9451      8.24327      13.0547
------------------------------------------------------------------------------



The StatAdvisor
---------------
   This table shows the predicted values for mpg using the fitted
model.  In addition to the best predictions, the table shows:
   (1) 95.0% prediction intervals for new observations
   (2) 95.0% confidence intervals for the mean of many observations
The prediction and confidence intervals correspond to the inner and
outer bounds on the graph of the fitted model.
```

Figure 17-5.   Forecasts

Use the *Forecasts Options* dialog box to enter a percentage for the confidence level that will be used to calculate the confidence limits, and to enter values for X that will be used to calculate a value for Y (see Figure 17-6).



Confidence Level: 95. %

Forecast at X:

48.0

165.0

Figure 17-6. Forecasts Options Dialog Box

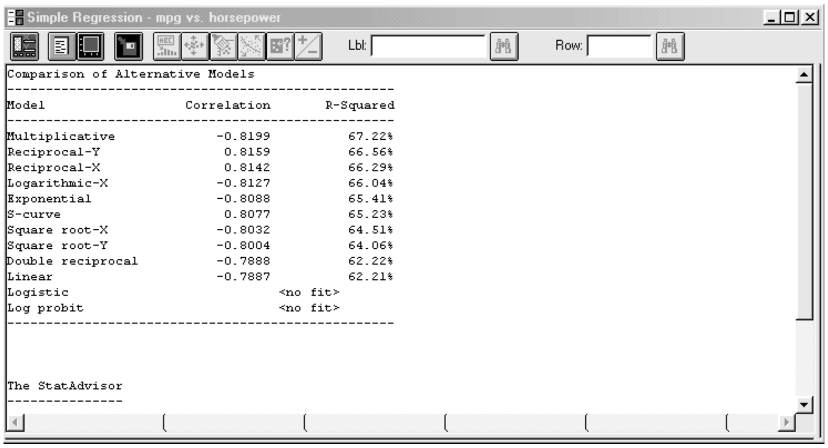
Comparison of Alternative Models

The Comparison of Alternative Models option creates a table of the results of fitting several curvilinear models to the data (see Figure 17-7).

It displays the correlation coefficients and R-Squared values for several transformed model types. Compare the R-Squared values for each transformed model with the values for the linear model. The preferred model is linear, unless another type of model significantly increases the R-Squared value.

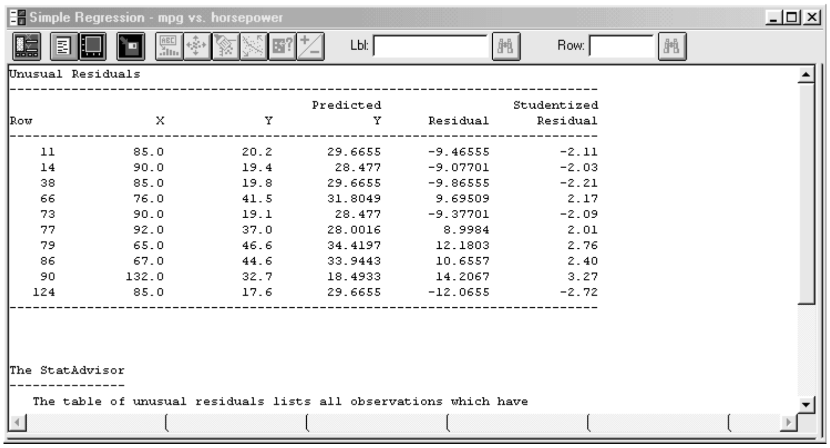
Unusual Residuals

The Unusual Residuals option creates a table of all the observations that have studentized residuals less than -2 or greater than 2 (see Figure 17-8). The table also shows the corresponding row, observed independent and dependent values, fitted value, and the residual. Take a close look at any values greater than 3 to determine if they are outliers, which you should remove from the analysis.



```
Comparison of Alternative Models
--------------------------------------------------
Model                     Correlation       R-Squared
--------------------------------------------------
Multiplicative                -0.8199          67.22%
Reciprocal-Y                   0.8159          66.56%
Reciprocal-X                   0.8142          66.29%
Logarithmic-X                 -0.8127          66.04%
Exponential                   -0.8088          65.41%
S-curve                        0.8077          65.23%
Square root-X                 -0.8032          64.51%
Square root-Y                 -0.8004          64.06%
Double reciprocal             -0.7888          62.22%
Linear                        -0.7887          62.21%
Logistic                            <no fit>
Log probit                          <no fit>
--------------------------------------------------



The StatAdvisor
---------------
```

Figure 17-7.   Comparison of Alternative Models



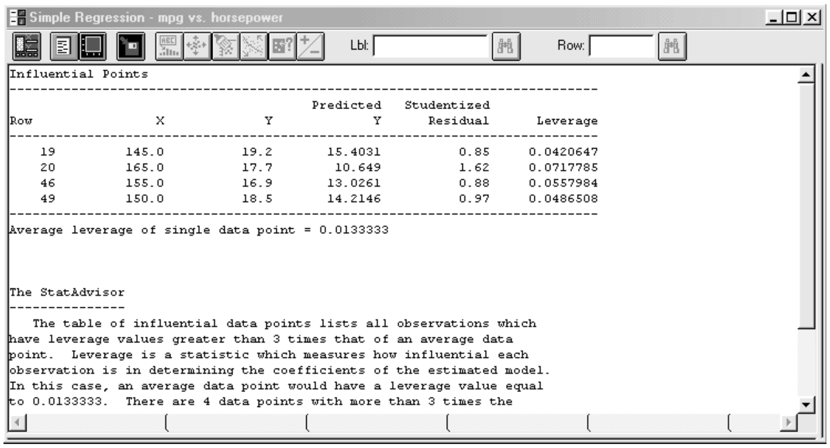
```
Unusual Residuals
----------------------------------------------------------------------------
                                       Predicted                Studentized
Row              X             Y             Y      Residual       Residual
----------------------------------------------------------------------------
    11         85.0          20.2       29.6655      -9.46555          -2.11
    14         90.0          19.4        28.477      -9.07701          -2.03
    38         85.0          19.8       29.6655      -9.86555          -2.21
    66         76.0          41.5       31.8049       9.69509           2.17
    73         90.0          19.1        28.477      -9.37701          -2.09
    77         92.0          37.0       28.0016        8.9984           2.01
    79         65.0          46.6       34.4197       12.1803           2.76
    86         67.0          44.6       33.9443       10.6557           2.40
    90        132.0          32.7       18.4933       14.2067           3.27
   124         85.0          17.6       29.6655      -12.0655          -2.72
----------------------------------------------------------------------------



The StatAdvisor
---------------
   The table of unusual residuals lists all observations which have
```

Figure 17-8.   Unusual Residuals

Influential Points

The Influential Points option creates a table that lists all the observations that have leverage values above three times that of an average point (see

Figure 17-9). The average leverage value is 2 divided by the number of complete observations.



```
Influential Points
---------------------------------------------------------------------------
                                       Predicted   Studentized
Row                X             Y             Y      Residual      Leverage
---------------------------------------------------------------------------
    19         145.0          19.2       15.4031          0.85     0.0420647
    20         165.0          17.7        10.649          1.62     0.0717785
    46         155.0          16.9       13.0261          0.88     0.0557984
    49         150.0          18.5       14.2146          0.97     0.0486508
---------------------------------------------------------------------------
Average leverage of single data point = 0.0133333



The StatAdvisor
---------------
   The table of influential data points lists all observations which
have leverage values greater than 3 times that of an average data
point.  Leverage is a statistic which measures how influential each
observation is in determining the coefficients of the estimated model.
In this case, an average data point would have a leverage value equal
to 0.0133333.  There are 4 data points with more than 3 times the
```

Figure 17-9.   Influential Points

Leverage determines how influential each observation is in determining the coefficients for the estimated model. Look carefully at points more than five times the average leverage to determine how much the model would change if you removed them from the data. *For more information see Belsley, Kuh, and Welsch (1980).*

Graphical Options

Plot of Fitted Model

The Plot of Fitted Model creates a plot of the results of fitting a linear model that describes the relationship between two selected variables. The equation of the fitted model is shown as a solid line (see Figure 17-10). The plot includes both confidence limits for the means (the middle lines) and prediction limits (the outer lines).

Use the *Plot of Fitted Model Options* dialog box to indicate the type of limits that will be included, to enter a value for the confidence level, and to enter a

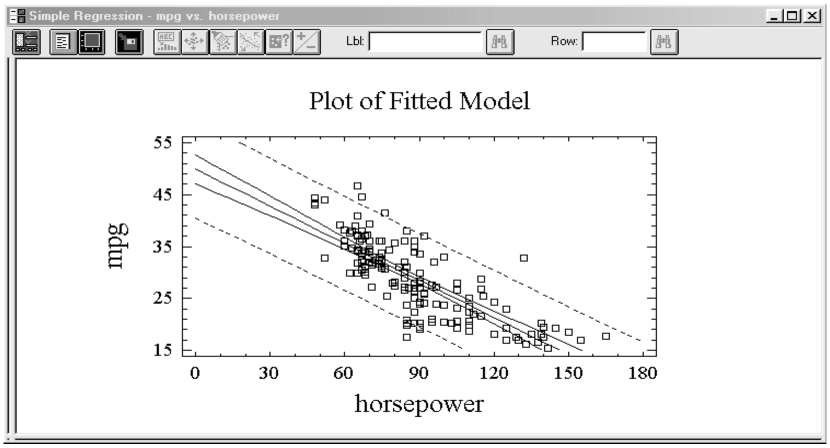
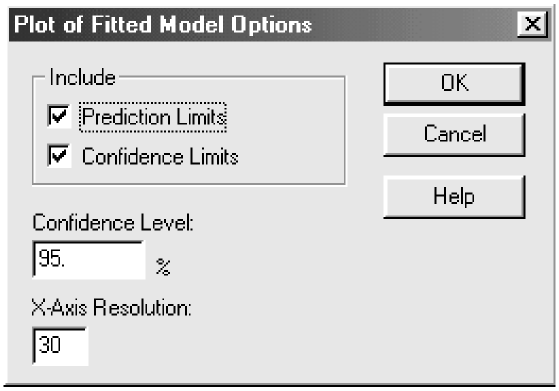


Figure 17-10. Plot of Fitted Model

number that will be used to change the resolution of the lines (see Figure 17-11).


Plot of Fitted Model Options

Include

☑ Prediction Limits

☑ Confidence Limits

Confidence Level: 95. %

X-Axis Resolution: 30

Figure 17-11. Plot of Fitted Model Options Dialog Box

Observed versus Predicted

The Observed versus Predicted option creates a plot of the observed values versus the predicted values for the dependent variable (the fitted model; see

Figure 17-12). The closer the points lie to the diagonal line, the better the model (at predicting the observed data).

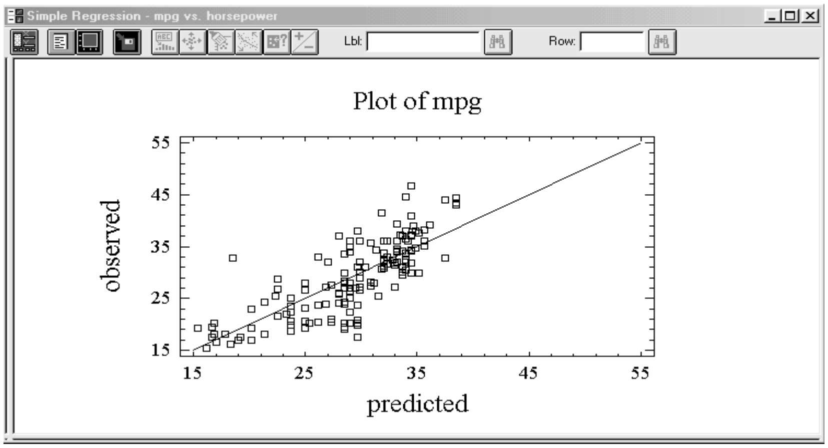


Figure 17-12. Observed versus Predicted Plot

Look for various anomalies, such as increases in variability around the line, as the value of the variable (the observed data) increases. This is known as heteroscedasticity. Also look for individual points that lie far away from the line, which are known as outliers. You can use the plot to detect cases in which the variance is not constant, which indicates you may need to transform the dependent variable.

Residuals versus X

The Residuals versus X option creates a plot of the studentized residuals versus the values for the independent variable (see Figure 17-13).

Nonrandom patterns could indicate that the model you selected does not adequately describe the observed data, and any values outside the range of -3 to +3 are outliers. Use the plot to detect the nonlinear relationship between Y and X. You can also use the plot to determine if the variance of the residuals is constant. If the model is correct, and if all the assumptions are satisfied, the residuals should appear structureless; that is, no pattern will be apparent.

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot (see Figure 17-14).

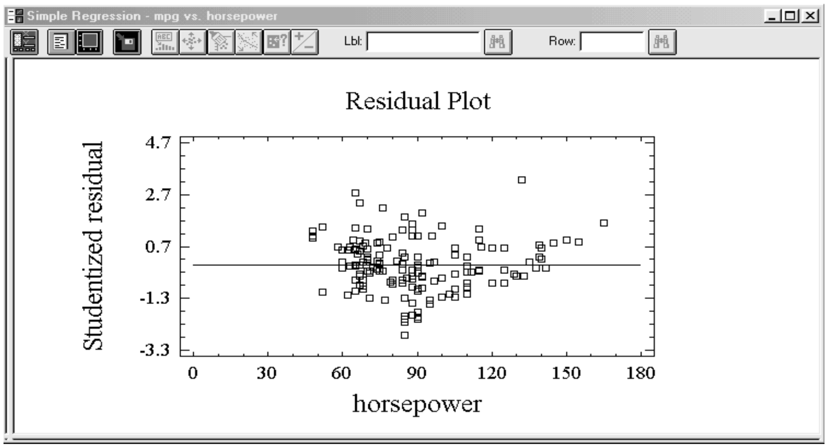
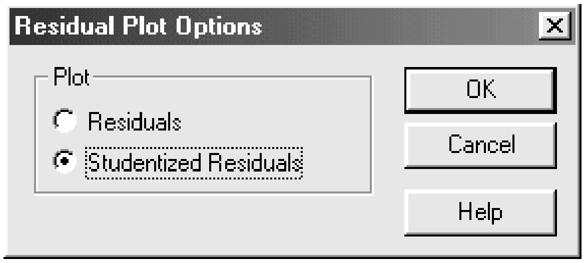


Figure 17-13. Residuals versus X Plot



Plot
○ Residuals
● Studentized Residuals

Figure 17-14. Residuals Plot Options Dialog Box

Residuals versus Predicted

The Residuals versus Predicted option creates a plot of the studentized residuals versus the predicted values of the variable you selected (see Figure 17-15). Nonrandom patterns indicate that the model you selected does not adequately describe the observed data.

This plot is especially helpful in showing heteroscedasticity, in which the variability of the residuals change as the values of the dependent variables change. If the model is correct and if all the assumptions are satisfied, the residuals should be structureless.

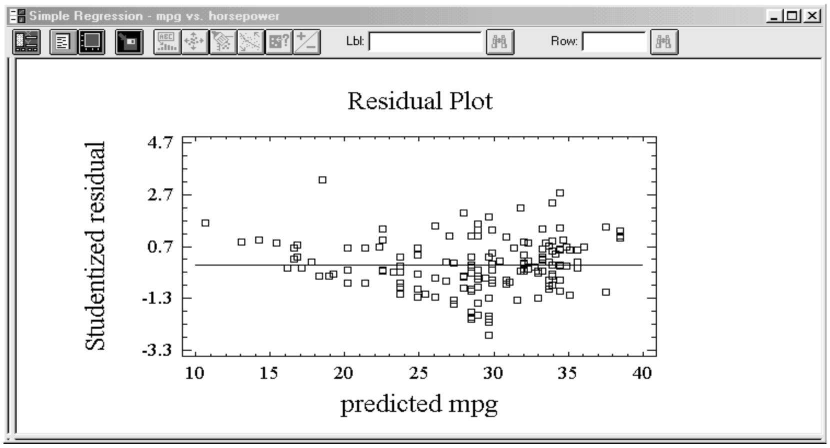


Figure 17-15. Residuals versus Predicted Plot

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot (see Figure 17-14 for an example of this dialog box).

Residuals versus Row Number

The Residuals versus Row Number option creates a plot of the studentized residuals versus their row number (see Figure 17-16). If the row order corresponds to the order in which the data were collected, any nonrandom pattern could indicate serial correlation in the data. Ideally, the points should appear structureless.

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot (see Figure 17-14 for an example of this dialog box).

Saving the Results

The Save Results Options dialog box allows you to select the results you want to save. There are eight selections: Predicted Values, Lower Limits for Predictions, Upper Limits for Predictions, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Studentized Residuals, and Leverages.

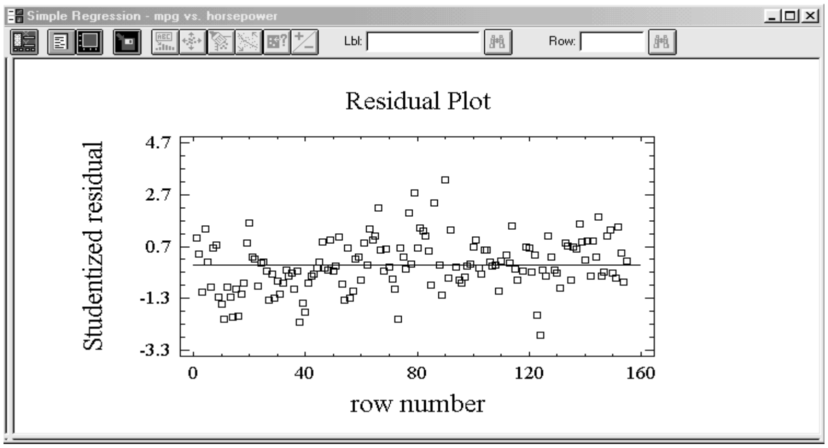


Figure 17-16. Residuals versus Row Number Plot

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results button on the Analysis toolbar (the fourth button from the left).

References

Belsley, D. A., Kuh, E., and Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity*. New York: Wiley.

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: Wiley.

Neter, J., Wasserman, W., and Kutner, M. 1985. *Applied Linear Statistical Models*. Homewood, Illinois: Richard E. Irwin, Inc.

Using Polynomial Regression Analysis

Polynomial regression analysis is regression analysis for relationships that are or are suspected to be nonlinear. It is known as polynomial regression because the regression equation for a curvilinear (nonlinear) relationship is a polynomial equation. The more turns in the regression line, the higher the power the terms in the equation must be.

The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on for the higher order effects (see Table 17-1 for the degree of freedom through the fifth).

Table 17-1. Polynomial Regression Degrees of Freedom

Turns of Regression Line	Order of Degree	Number of Terms	Type of Analysis
0	First	1	Linear
1	Second	2	Quadratic
2	Third	3	Cubic
3	Fourth	4	Quartic
4	Fifth	5	Quintic

A polynomial equation is an equation in which one of the terms is raised to a power greater than 1. For example, the regression equation $Y = a + b_1X + b_2X^2 + b_3X^3$ is a polynomial equation because X is raised to the second and third powers. The highest power of a term gives the equation its “degree” or “order.” The equation shown above is a third-order (or degree) polynomial equation.

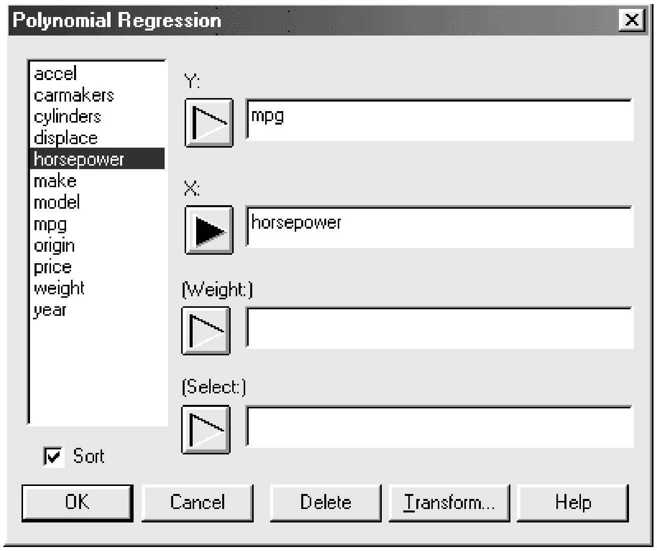
The Polynomial Regression Analysis calculates a polynomial regression model between a single dependent variable Y and a single independent variable X. Use this analysis if you think your model fits a polynomial instead of a linear model. You can fit polynomial models up to Order 8.

To access the analysis, from the menus, choose: RELATE... POLYNOMIAL REGRESSION... (see Figure 17-17).

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis, which fits a second-order polynomial model that describes the relationship between two variables. The model-fitting results in the top portion of the pane include

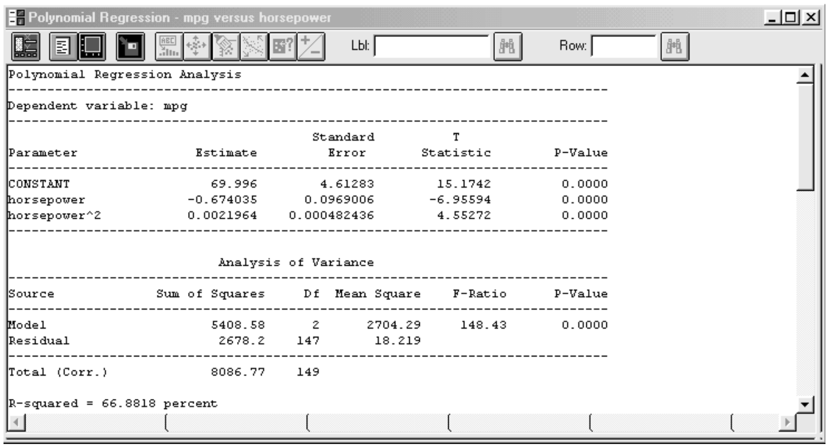


Y: mpg

X: horsepower

Figure 17-17. The Polynomial Regression Analysis Dialog Box

estimates for the constant (intercept term) and the model coefficients for each power of the independent variable (see Figure 17-18).



```
Polynomial Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: mpg
-----------------------------------------------------------------------------
                                       Standard          T
Parameter                Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                   69.996       4.61283        15.1742         0.0000
horsepower              -0.674035     0.0969006       -6.95594         0.0000
horsepower^2            0.0021964   0.000482436        4.55272         0.0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     5408.58      2      2704.29     148.43       0.0000
Residual                   2678.2    147       18.219
-----------------------------------------------------------------------------
Total (Corr.)             8086.77    149

R-squared = 66.8818 percent
```

Figure 17-18. Analysis Summary

The model is fit to an Order 2 polynomial by default. Use the Order option on the Polynomial Regression Options dialog box to change the order of the model.

The table also displays several key measures you can use to judge the adequacy of the model: the standard errors of the coefficients; the *t* statistics (which are calculated by dividing the estimate by its standard error), and the *p*-value for each *t* statistic. The *p*-value corresponds to tests of hypotheses that the coefficients are equal to zero. Small *p*-values indicate statistically significant regression coefficients. Large *p*-values indicate regression coefficients that have little or no effect on the dependent variable at the 95 percent confidence level. The estimates that have a high probability level (greater than .05) have little or no effect on the dependent variable at the 95 percent confidence level. The F-ratio tests the overall model significance; in other words, it tests to see if the coefficients are all equal to zero.

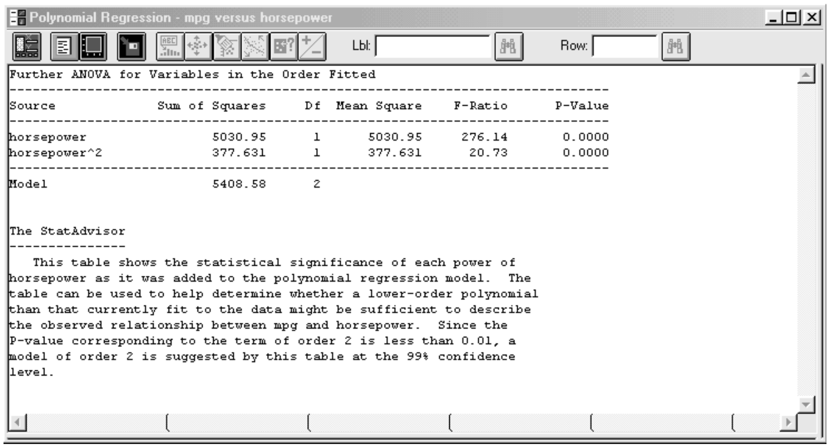
The bottom portion of the summary displays the analysis of variance statistics for the full regression, and provides five additional values to help you judge the utility of the model.

- The R-Squared statistic measures the proportion of variability explained in the model for the dependent variable.
- The Adjusted R-Squared value is the amount of variation explained in the model for the dependent variable. It adjusts the R-Squared value based on the number of coefficients in the model. Unlike the R-Square statistic, the Adjusted R-Squared value may decrease if higher orders do not significantly add to the fit.
- The standard error of the estimate is the square root of the residual mean square (the mean squared error). It measures the unexplained variability in the dependent variable.
- The mean absolute error is the average of the absolute values of the residuals. It is the average error size you can expect in a prediction.
- The Durbin-Watson statistic is a measure of the serial correlation of the residuals. In general, a Durbin-Watson statistic of less than 1.3 indicates serial correlation in the residuals. You can find tables that interpret the statistic in Durbin and Watson (1951) or in standard text books such as Draper and Smith (1981).

Use the *Polynomial Regression Options* dialog box to enter the number of the highest power you want to include in the model.

Conditional Sums of Squares

The Conditional Sums of Squares option creates additional analysis of variance statistics (see Figure 17-19). The table shows the statistical significance of each power as it was added to the polynomial regression model, and includes the contribution each independent variable made to the total regression sums of squares when it entered the regression (also called Type I sums of squares).

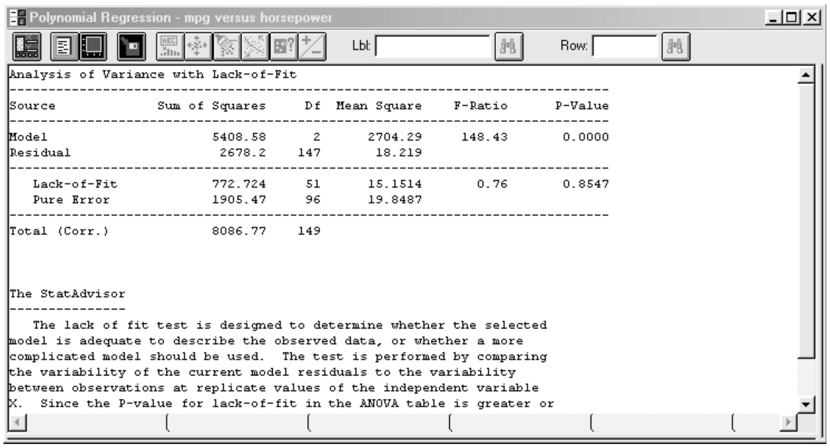


```
Further ANOVA for Variables in the Order Fitted
------------------------------------------------------------------------------
Source            Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
horsepower               5030.95      1      5030.95     276.14       0.0000
horsepower^2             377.631      1      377.631      20.73       0.0000
------------------------------------------------------------------------------
Model                    5408.58      2


The StatAdvisor
---------------
   This table shows the statistical significance of each power of
horsepower as it was added to the polynomial regression model.  The
table can be used to help determine whether a lower-order polynomial
than that currently fit to the data might be sufficient to describe
the observed relationship between mpg and horsepower.  Since the
P-value corresponding to the term of order 2 is less than 0.01, a
model of order 2 is suggested by this table at the 99% confidence
level.
```

Figure 17-19. Conditional Sums of Squares

Lack-of-Fit Test

The Lack-of-Fit Test option creates a modified ANOVA table with a line for lack-of-fit (see Figure 17-20). The test is designed to determine if the regression model adequately fits the data. The Lack-of-Fit test is performed by comparing the variability of the residuals for the current model with the variability between observations at replicate values of the independent variable. The test assumes that the observations, Y for a given X, are independent and normally distributed, and that the distributions of Y have the same variance. Significant lack-of-fit shows that the specified model does not adequately fit the response.

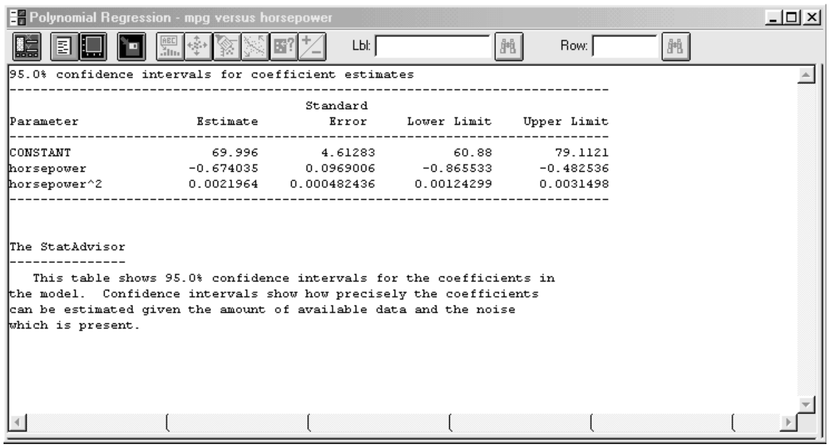


```
Analysis of Variance with Lack-of-Fit
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     5408.58      2      2704.29     148.43       0.0000
Residual                   2678.2    147       18.219
-----------------------------------------------------------------------------
   Lack-of-Fit            772.724     51      15.1514       0.76       0.8547
   Pure Error             1905.47     96      19.8487
-----------------------------------------------------------------------------
Total (Corr.)             8086.77    149



The StatAdvisor
---------------
   The lack of fit test is designed to determine whether the selected
model is adequate to describe the observed data, or whether a more
complicated model should be used.  The test is performed by comparing
the variability of the current model residuals to the variability
between observations at replicate values of the independent variable
X.  Since the P-value for lack-of-fit in the ANOVA table is greater or
```

Figure 17-20.   Lack-of-Fit Test

Confidence Intervals

The Confidence Intervals option creates 95 percent confidence intervals for the coefficient estimates (see Figure 17-21). The estimates and the standard errors are identical to those shown in the initial table of model-fitting results.



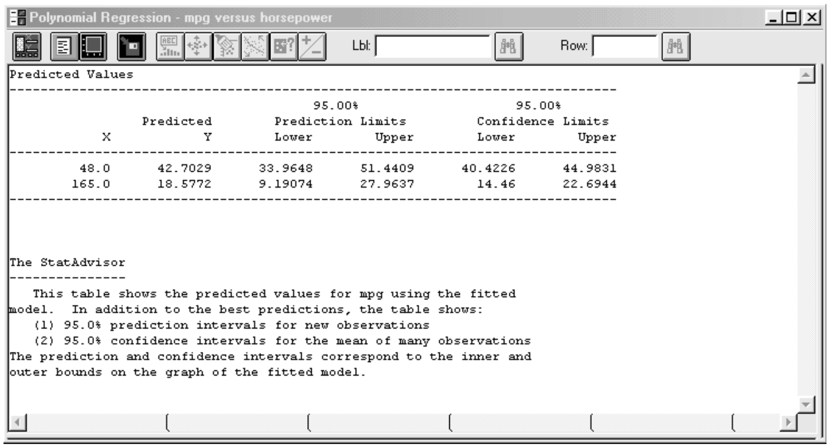
```
95.0% confidence intervals for coefficient estimates
-----------------------------------------------------------------------------
                                      Standard
Parameter               Estimate         Error    Lower Limit    Upper Limit
-----------------------------------------------------------------------------
CONSTANT                  69.996       4.61283          60.88        79.1121
horsepower             -0.674035     0.0969006      -0.865533      -0.482536
horsepower^2           0.0021964   0.000482436     0.00124299      0.0031498
-----------------------------------------------------------------------------


The StatAdvisor
---------------
   This table shows 95.0% confidence intervals for the coefficients in
the model.  Confidence intervals show how precisely the coefficients
can be estimated given the amount of available data and the noise
which is present.
```

Figure 17-21.   Confidence Inervals

Use the *Confidence Intervals Options* dialog box to change the number used to calculate the confidence intervals.

Forecasts

The Forecasts option creates predicted values for the Y variable using the fitted model (see Figure 17-22). In addition to the best predictions, the table displays the prediction intervals for new observations and confidence intervals for the mean of many observations (at a given percentage).



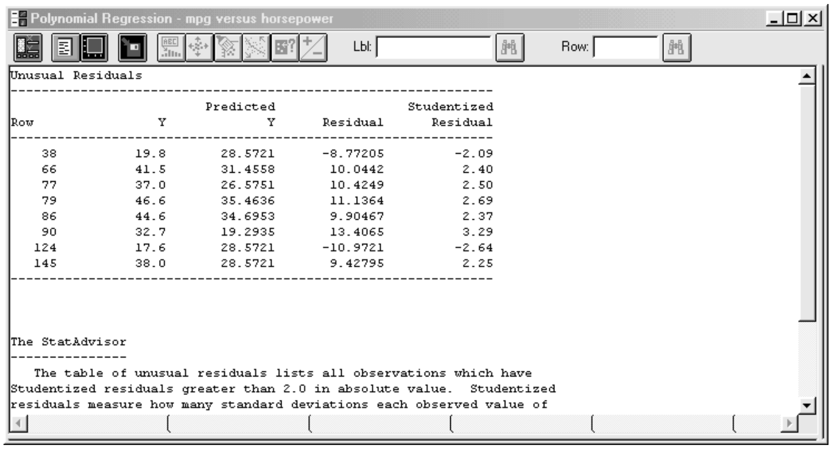
```
Predicted Values
------------------------------------------------------------------------------
                                       95.00%                    95.00%
                 Predicted       Prediction Limits         Confidence Limits
            X            Y        Lower        Upper        Lower        Upper
------------------------------------------------------------------------------
         48.0      42.7029      33.9648      51.4409      40.4226      44.9831
        165.0      18.5772      9.19074      27.9637        14.46      22.6944
------------------------------------------------------------------------------



The StatAdvisor
---------------
   This table shows the predicted values for mpg using the fitted
model.  In addition to the best predictions, the table shows:
   (1) 95.0% prediction intervals for new observations
   (2) 95.0% confidence intervals for the mean of many observations
The prediction and confidence intervals correspond to the inner and
outer bounds on the graph of the fitted model.
```

Figure 17-22. Forecasts

Use the *Forecasts Options* dialog box to enter values to change the confidence level that will be used to calculate the confidence limits, and to enter values for X for which you want to predict a Y value (see Figure 17-6 for an example of this dialog box).

Unusual Residuals

The Unusual Residuals option creates a table that displays the values for all the observations that have studentized residuals less than -2 or greater than 2 in absolute value (see Figure 17-23). Studentized residuals measure the number of standard deviations each observed value of the independent variable deviates from a model fitted using all the data except that observation. Take a close look at any values greater than 3 in absolute value to determine if they are outliers, which you should remove from the analysis.



```
Unusual Residuals
--------------------------------------------------------------
                        Predicted                  Studentized
Row               Y             Y      Residual      Residual
--------------------------------------------------------------
    38         19.8       28.5721      -8.77205         -2.09
    66         41.5       31.4558       10.0442          2.40
    77         37.0       26.5751       10.4249          2.50
    79         46.6       35.4636       11.1364          2.69
    86         44.6       34.6953       9.90467          2.37
    90         32.7       19.2935       13.4065          3.29
   124         17.6       28.5721      -10.9721         -2.64
   145         38.0       28.5721       9.42795          2.25
--------------------------------------------------------------



The StatAdvisor
---------------
   The table of unusual residuals lists all observations which have
Studentized residuals greater than 2.0 in absolute value.  Studentized
residuals measure how many standard deviations each observed value of
```

Figure 17-23.   Unusual Residuals

Influential Points

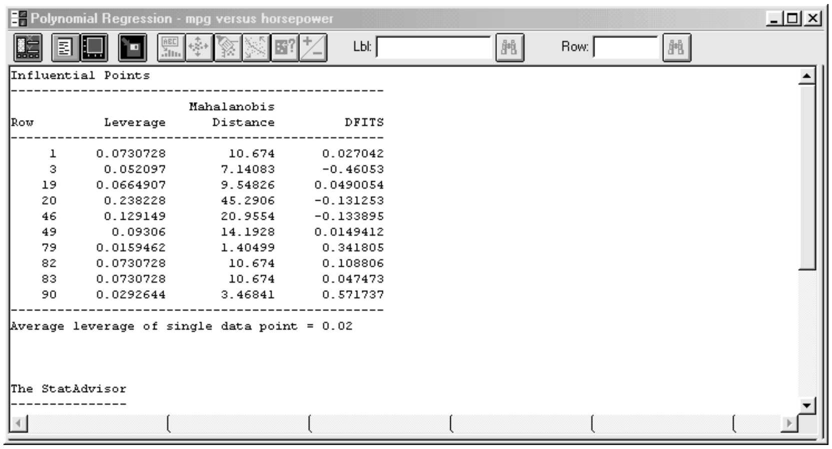
The Influential Points option creates a table that lists all the observations that have leverage values greater than three times the average (see Figure 17-24). Leverage is a statistic that measures the influence of each observation in determining the coefficients of the estimated model. DFITS is a statistic that measures how much the estimated coefficients would change if each observation was removed from the data.

The table also displays values for the Mahalanobis distance, which is the square of the standardized value of X. *For more information, see Belsley, Kuh, and Welsch (1980)*.

Graphical Options

Plot of Fitted Model

The Plot of Fitted Model option creates a plot of the results of fitting a second-order polynomial model to describe the relationship between the



```
Influential Points
------------------------------------------------
                       Mahalanobis
Row         Leverage      Distance         DFITS
------------------------------------------------
     1     0.0730728        10.674      0.027042
     3      0.052097       7.14083      -0.46053
    19     0.0664907       9.54826     0.0490054
    20      0.238228       45.2906     -0.131253
    46      0.129149       20.9554     -0.133895
    49       0.09306       14.1928     0.0149412
    79     0.0159462       1.40499      0.341805
    82     0.0730728        10.674      0.108806
    83     0.0730728        10.674      0.047473
    90     0.0292644       3.46841      0.571737
------------------------------------------------
Average leverage of single data point = 0.02



The StatAdvisor
---------------
```

Figure 17-24. Influential Points

dependent and independent variables (see Figure 17-25). The plot shows the curve or the fitted line, and includes both confidence limits for the means (the inner pair of lines) and prediction limits (the outer pair of lines).

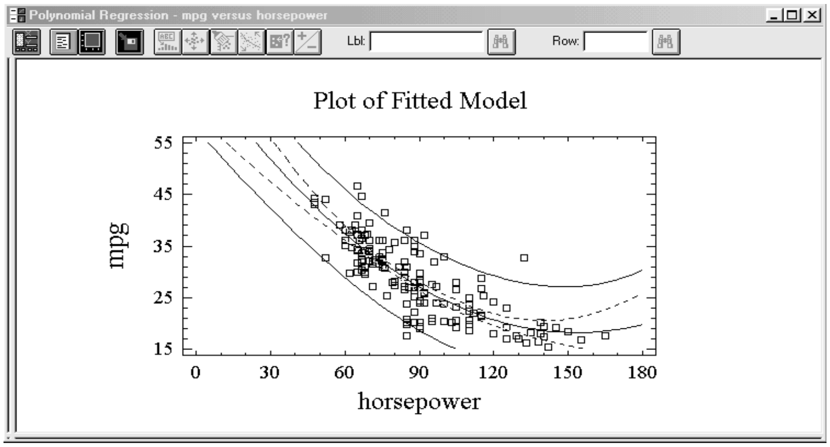


Figure 17-25. Plot of Fitted Model

Use the *Plot of Fitted Model Options* dialog box to indicate which limits will be shown on the plot, to enter a value for the confidence limits that will be used to calculate the confidence intervals, and to enter a number that will be used to change the resolution of the lines on the plot (see Figure 17-11 for an example of this dialog box).

Observed versus Predicted

The Observed versus Predicted option creates a plot of the observed values of the independent variable versus the values predicted by the fitted model (see Figure 17-26). The closer the points are to the diagonal line, the better the model.

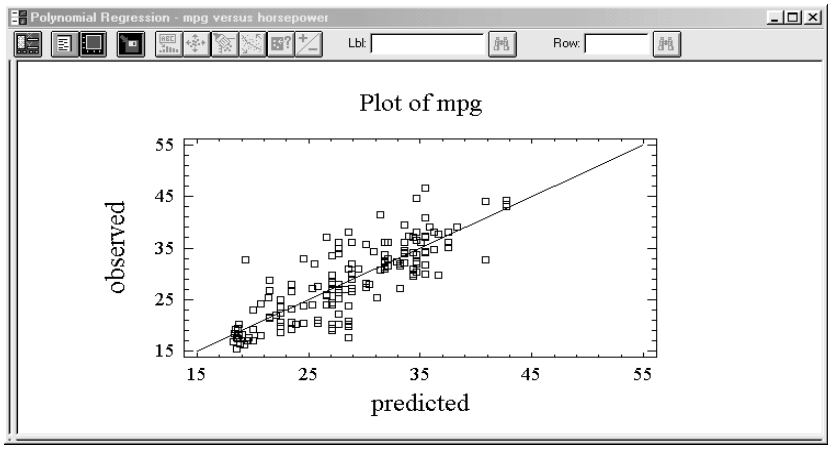


Figure 17-26. Observed versus Predicted Plot

Residuals versus X

The Residuals versus X option creates a plot of the residuals versus the dependent variable (see Figure 17-27). A nonrandom pattern indicates that the selected model does not adequately describe the observed data. Any values outside the range of -3 to +3 are probably outliers.

Use the *Residual Plot Options* dialog box to indicate if residuals or studentized residuals will appear on the plot.

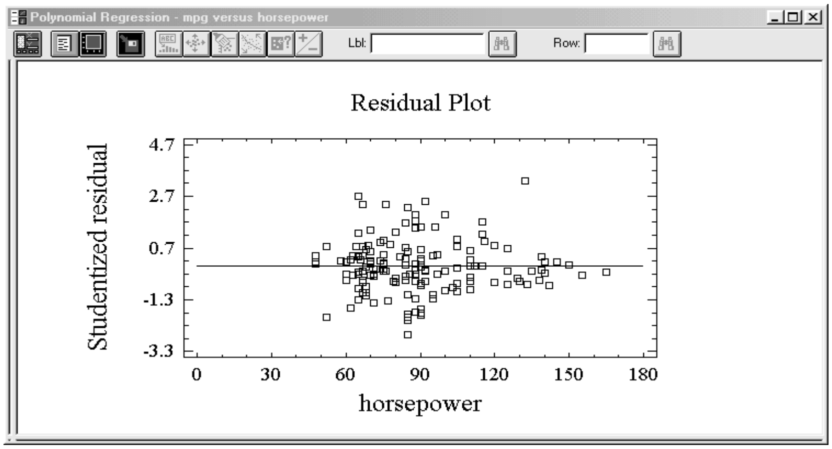


Figure 17-27. Residuals versus X Plot

Residuals versus Predicted

The Residuals versus Predicted option creates a plot of the residuals versus the predicted values for the independent variable (see Figure 17-28). A nonrandom pattern could indicate that the selected model does not adequately describe the observed data. The plot is especially helpful in detecting heteroscedasticity, in which the variability of the residuals changes as the values of the dependent variable change. If the model is correct and if all the assumptions are satisfied, the residuals should be structureless.

Use the *Residual Plot Options* dialog box to indicate if residuals or studentized residuals will appear on the plot.

Residuals versus Row Number

The Residuals versus Row Number option creates a plot of the residuals versus their row number (see Figure 17-29). The residuals are plotted in the order that the observations appear in the dependent variable. Any nonrandom pattern could indicate serial correlation in the data.

Use the *Residual Plot Options* dialog box to indicate if residuals or studentized residuals will appear on the plot.

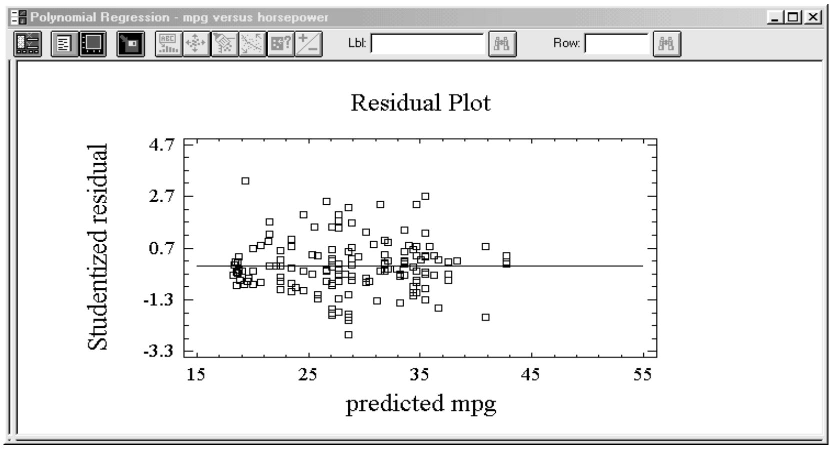


Figure 17-28. Residuals versus Predicted Plot

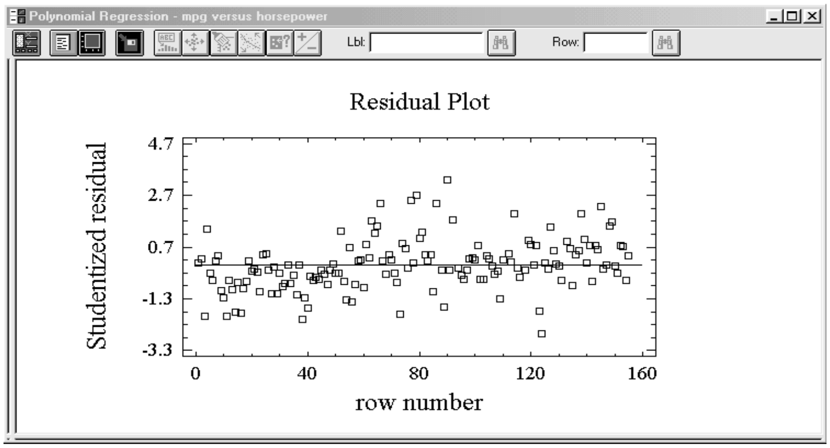


Figure 17-29. Residuals versus Row Number Plot

Saving the Results

The Save Results Options dialog box allows you to select the results you want to save. There are 12 selections: Predicted Values, Standard Errors of Predictions, Lower Limits for Predictions, Upper Limits for Predictions, Standard Error of Means, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Studentized Residuals, Leverages, DFITS Statistics, and Mahalanobis Distances.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Belsley, D. A., Kuh, E., and Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity.* New York: Wiley.

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: Wiley.

Durbin, J. and Watson, G. S. 1951. "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, **38**.

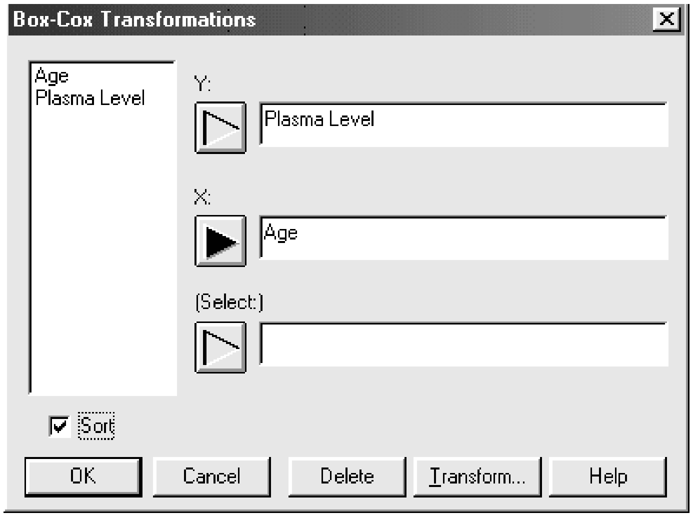
Montgomery, D. C. 1991. *Design and Analysis of Experiments,* third edition. New York: Wiley.

Using Box-Cox Transformations Analysis

In the context of fitting a simple regression model, it is assumed that the error variances in the different groups are homogeneous and uncorrelated with the means. It is also assumed that the error terms are normally distributed. If there are departures from the simple regression model, it is helpful to transform the data to stabilize the error variances and normalize the error terms. Transforming the data can also help linearize a curvilinear regression relation.

The Box-Cox Transformations Analysis automatically identifies a transformation from a family of power transformations. The analysis determines the appropriate transformation parameter, Lambda1, which minimizes the mean squared error (MSE) of the fitted model.

To access the analysis, from the menus, choose: RELATE... BOX-COX TRANSFORMATIONS... (see Figure 17-30).



Box-Cox Transformations

Age
Plasma Level

Y: Plasma Level

X: Age

(Select:)

☑ Sort

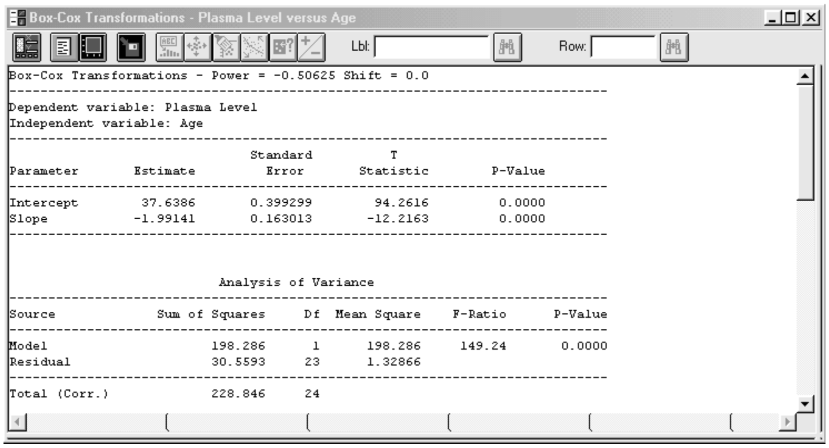
Figure 17-30. The Box-Cox Transformations Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the value of Lambda1, which minimizes the mean squared error (MSE). The analysis compares the effect of various power transformations (of the dependent variable on the linear regression) with the independent variable (see Figure 17-31). The summary includes the equation for the fitted model, which is a Box-Cox transformation with power determined to minimize the MSE.

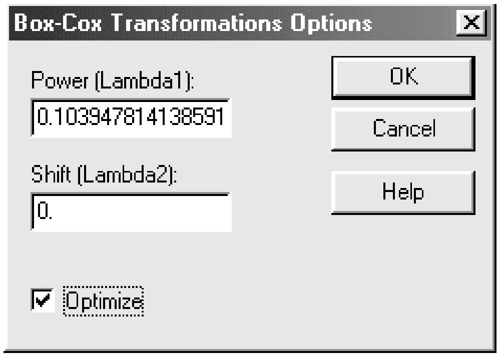
If the *p*-value in the ANOVA table is less than .05, there is a statistically significant relationship between the transformed values of the two variables at the 95 percent confidence level. The R-Squared statistic explains the percentage of variability that is explained by the model. Also shown is the correlation coefficient, which indicates the relationship between the two variables. The standard error of the estimate shows the standard deviation of the residuals. Also, the Durbin-Watson statistic is now included.



```
Box-Cox Transformations - Power = -0.50625 Shift = 0.0
-----------------------------------------------------------------------------
Dependent variable: Plasma Level
Independent variable: Age
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        37.6386       0.399299        94.2616         0.0000
Slope           -1.99141       0.163013       -12.2163         0.0000
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     198.286      1      198.286     149.24       0.0000
Residual                  30.5593     23      1.32866
-----------------------------------------------------------------------------
Total (Corr.)             228.846     24
```

Figure 17-31.   Analysis Summary

Use the *Box-Cox Transformations Options* dialog box to enter values for power (Lambda1), and shift (lambda2), and to indicate if Lambda1 should be optimized (see Figure 17-32).



Power (Lambda1): 0.103947814138591

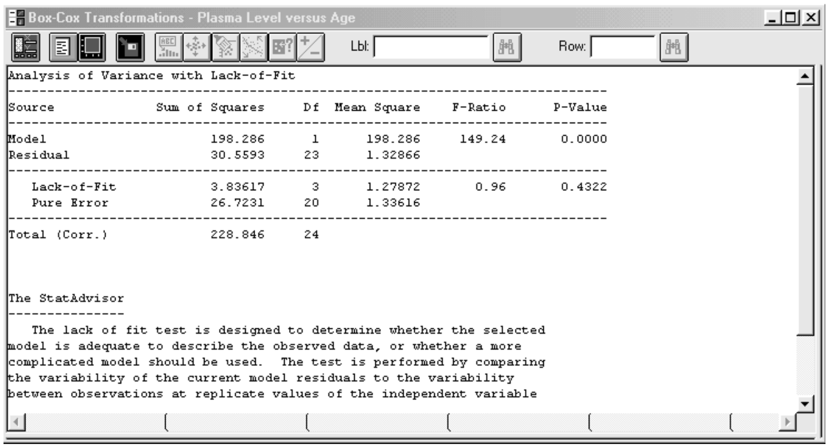
Shift (Lambda2): 0.

☑ Optimize

Figure 17-32.   Box-Cox Transformations Options Dialog Box

Lack-of-Fit Test

The Lack-of-Fit Test option creates a table that displays the results of a test designed to determine whether the selected model adequately describes the observed data, or whether a more complicated model should be used instead (see Figure 17-33). The analysis compares the variability of the residuals for the current model with the variability between observations at replicate values of the independent variable X. If the *p*-value is greater than or equal to .05, the current model should be adequate for the observed data at the 95 percent confidence level.



```
Analysis of Variance with Lack-of-Fit
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     198.286      1      198.286     149.24       0.0000
Residual                  30.5593     23      1.32866
-----------------------------------------------------------------------------
   Lack-of-Fit            3.83617      3      1.27872       0.96       0.4322
   Pure Error             26.7231     20      1.33616
-----------------------------------------------------------------------------
Total (Corr.)             228.846     24



The StatAdvisor
---------------
   The lack of fit test is designed to determine whether the selected
model is adequate to describe the observed data, or whether a more
complicated model should be used.  The test is performed by comparing
the variability of the current model residuals to the variability
between observations at replicate values of the independent variable
```

Figure 17-33. Lack-of-Fit Test

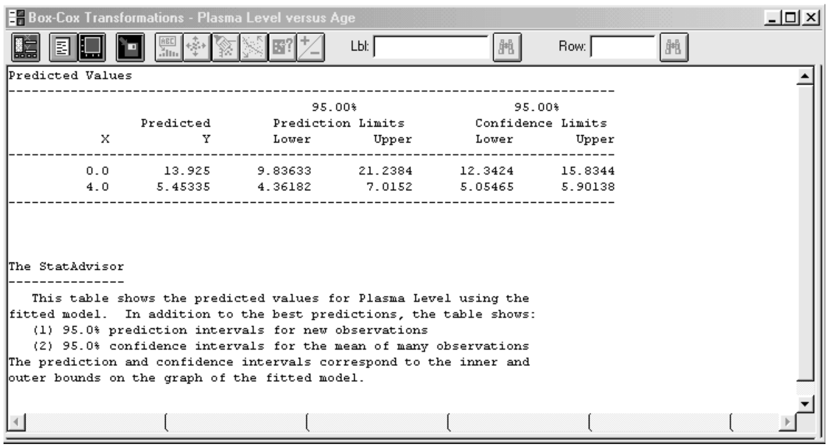
Forecasts

The Forecasts option creates a table of the predicted values for the Y (dependent) variable using the fitted model (see Figure 17-34). In addition to the best predictions, the table displays the prediction intervals for the new observations and the confidence intervals for the mean of many observations (at a given percentage).

Use the *Forecasts Options* dialog box to enter values to change the confidence level that will be used to calculate the confidence limits, and to enter values for X for which you want to predict a Y value (see Figure 17-6 for an example of this dialog box).

MSE Comparison Table

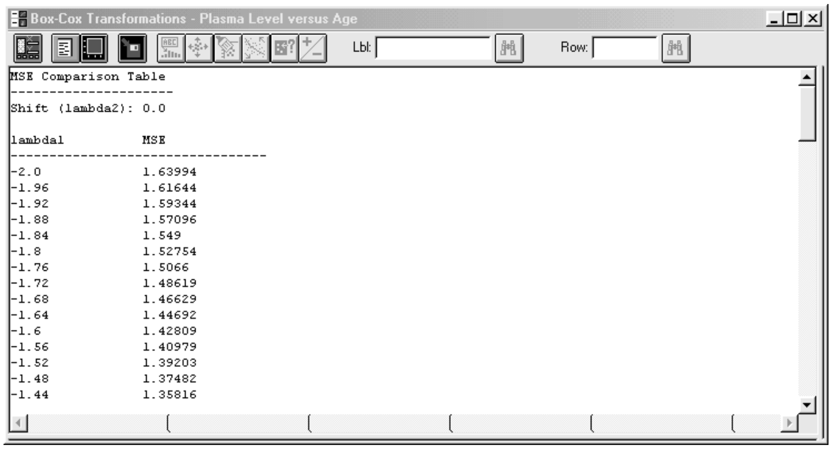
The MSE Comparison Table option creates a table that displays the mean squared error (MSE) for various values of the power transformation parameter (Lambda1) between -2 and +2 (see Figure 17-35). The MSE Table



```
Predicted Values
------------------------------------------------------------------------------
                                       95.00%                    95.00%
                 Predicted       Prediction Limits         Confidence Limits
            X            Y        Lower        Upper        Lower        Upper
------------------------------------------------------------------------------
          0.0       13.925      9.83633      21.2384      12.3424      15.8344
          4.0      5.45335      4.36182       7.0152      5.05465      5.90138
------------------------------------------------------------------------------



The StatAdvisor
---------------
   This table shows the predicted values for Plasma Level using the
fitted model.  In addition to the best predictions, the table shows:
   (1) 95.0% prediction intervals for new observations
   (2) 95.0% confidence intervals for the mean of many observations
The prediction and confidence intervals correspond to the inner and
outer bounds on the graph of the fitted model.
```

Figure 17-34.   Forecasts

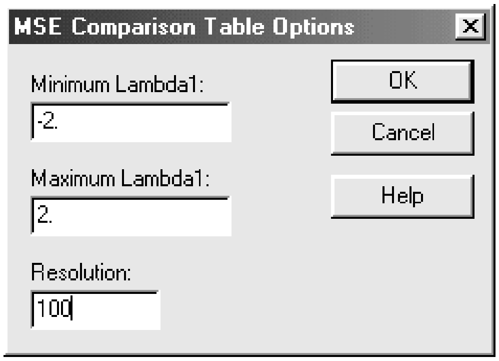
shows the estimated model MSE for a range of lambda values. By default, this analysis will always choose the optimal Lambda1. Using the table you can determine the value of Lambda1, which minimizes the model MSE. You can also see the MSE for other values of lambda, other than the optimal.



```
MSE Comparison Table
---------------------
Shift (lambda2): 0.0

lambda1           MSE
---------------------------------
-2.0              1.63994
-1.96             1.61644
-1.92             1.59344
-1.88             1.57096
-1.84             1.549
-1.8              1.52754
-1.76             1.5066
-1.72             1.48619
-1.68             1.46629
-1.64             1.44692
-1.6              1.42809
-1.56             1.40979
-1.52             1.39203
-1.48             1.37482
-1.44             1.35816
```

Figure 17-35.   MSE Comparison Table

Use the *MSE Comparison Table Options* dialog box to enter values for the minimum and maximum Lambda1, and for the resolution (see Figure 17-36).


MSE Comparison Table Options

Minimum Lambda1: -2.

Maximum Lambda1: 2.

Resolution: 100

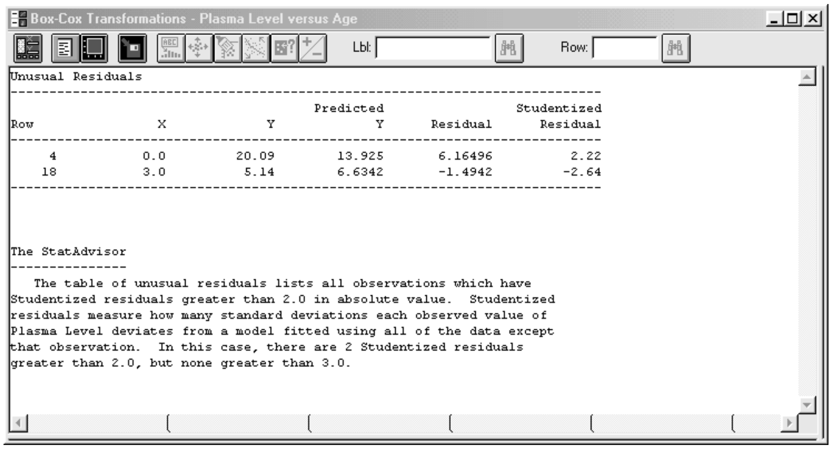
Figure 17-36. MSE Comparison Table Options Dialog Box

Unusual Residuals

The Unusual Residuals option creates a table that lists all the observations that have studentized residuals greater than 2.0 in absolute value (see Figure 17-37). Studentized residuals measure the number of standard deviations each observed value of the dependent variable deviates from a model fitted using all the data except that observation. Take a close look at any values greater than 3 in absolute value to determine if they are outliers, which you should remove from the analysis.

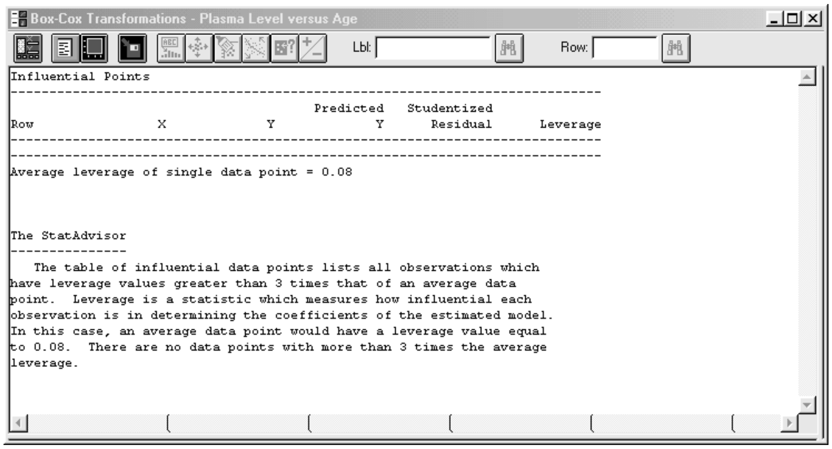
Influential Points

The Influential Points option creates a table that lists all the observations that have leverage values greater than three times the average (see Figure 17-38). Leverage is a statistic that measures the amount of influence of each observation in determining the coefficients for the estimated model.



```
Unusual Residuals
----------------------------------------------------------------------------
                                       Predicted                 Studentized
Row               X             Y             Y      Residual       Residual
----------------------------------------------------------------------------
     4          0.0         20.09        13.925       6.16496           2.22
    18          3.0          5.14        6.6342       -1.4942          -2.64
----------------------------------------------------------------------------



The StatAdvisor
---------------
   The table of unusual residuals lists all observations which have
Studentized residuals greater than 2.0 in absolute value.  Studentized
residuals measure how many standard deviations each observed value of
Plasma Level deviates from a model fitted using all of the data except
that observation.  In this case, there are 2 Studentized residuals
greater than 2.0, but none greater than 3.0.
```

Figure 17-37. Unusual Residuals Plot



```
Influential Points
----------------------------------------------------------------------------
                                       Predicted   Studentized
Row               X             Y             Y      Residual       Leverage
----------------------------------------------------------------------------
----------------------------------------------------------------------------
Average leverage of single data point = 0.08



The StatAdvisor
---------------
   The table of influential data points lists all observations which
have leverage values greater than 3 times that of an average data
point.  Leverage is a statistic which measures how influential each
observation is in determining the coefficients of the estimated model.
In this case, an average data point would have a leverage value equal
to 0.08.  There are no data points with more than 3 times the average
leverage.
```

Figure 17-38. Influential Points

Graphical Options

Plot of Fitted Model

The Plot of Fitted Model creates a plot of the results of fitting a linear model that describes the relationship between two selected variables. The equation of the fitted model is shown as a solid line (see Figure 17-39). The plot includes both confidence limits for the means (the middle lines) and prediction limits (the outer lines).

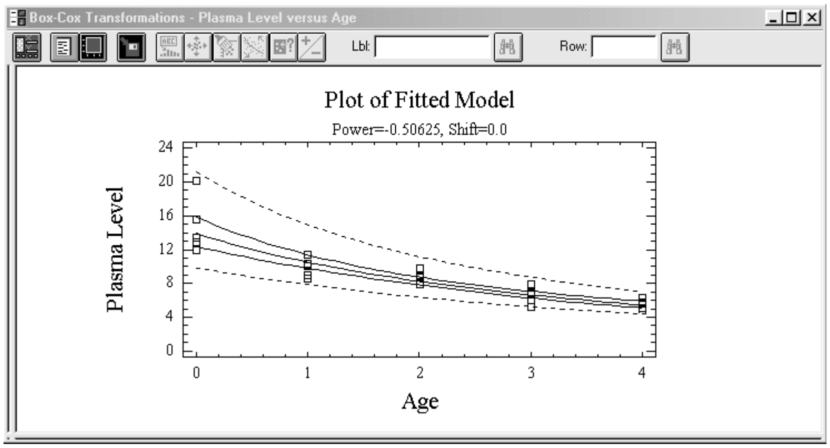

Box-Cox Transformations - Plasma Level versus Age
Lbl:
Row:
Plot of Fitted Model
Power=-0.50625, Shift=0.0
Plasma Level
24
20
16
12
8
4
0
0
1
2
3
4
Age

Figure 17-39. Plot of Fitted Model

Use the *Plot of Fitted Model Options* dialog box to indicate the type of limits that will be included, to enter a value for the confidence level, and to enter a number that will be used to change the resolution of the lines (see Figure 17-11 for an example of this dialog box).

Observed versus Predicted

The Observed versus Predicted option creates a plot of the observed values versus the predicted values for the dependent variable (the fitted model; see Figure 17-40). The closer the points lie to the diagonal line, the better the model (at predicting the observed data).

Look for various anomalies, such as increases in variability around the line, as the value of the variable (the observed data) increases. This is known as heteroscedasticity. Also look for individual points that lie far away from the line, which are known as outliers. You can use the plot to detect cases in which the variance is not constant, which indicates you may need to transform the dependent variable.

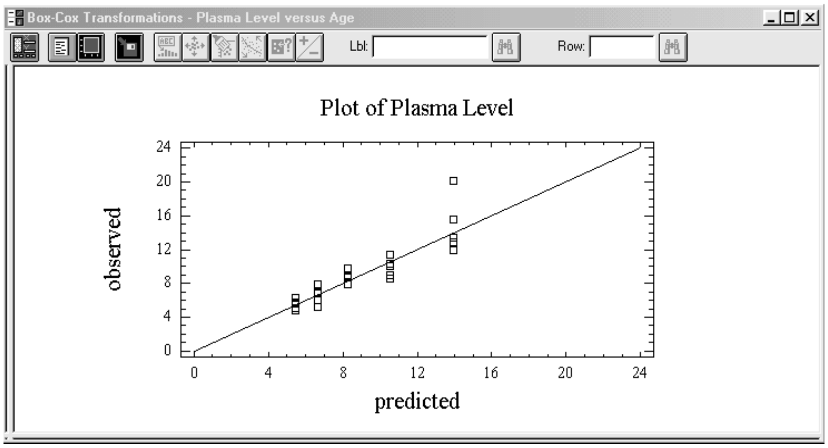


Figure 17-40. Observed versus Predicted Plot

MSE Comparison Plot

The MSE Comparison Plot option creates a plot of the mean squared error (MSE) for various values of the power transformation parameter Lambda1 between -2 and +2 (see Figure 17-41). The analysis, by default, will always choose Lambda1.

Use the *MSE Comparison Plot Options* dialog box to enter values for the minimum and maximum Lambda1, and for the resolution (see Figure 17-42).

Residuals versus X

The Residuals versus X option creates a plot of the studentized residuals versus the values for the independent variable (see Figure 17-43).

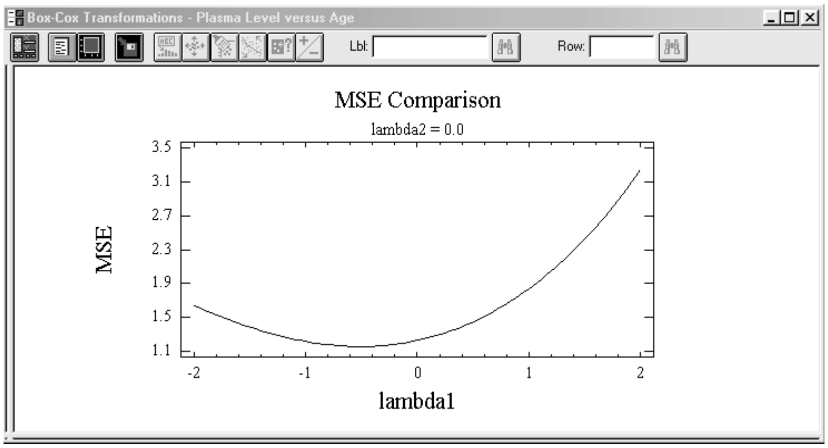
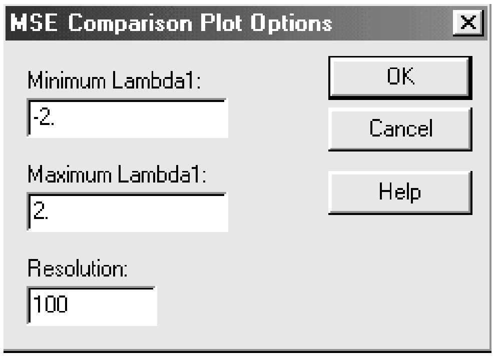


Figure 17-41. MSE Comparison Plot



Minimum Lambda1: -2.

Maximum Lambda1: 2.

Resolution: 100

Figure 17-42. MSE Comparison Plot Options Dialog Box

Nonrandom patterns could indicate that the model you selected does not adequately describe the observed data, and any values outside the range of -3 to +3 are outliers. Use the plot to detect the nonlinear relationship between Y and X. You can also use the plot to determine if the variance of the residuals is constant. If the model is correct, and if all the assumptions are satisfied, the residuals should appear structureless; that is, no pattern will be apparent.

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot (see Figure 17-14 for an example of this dialog box).

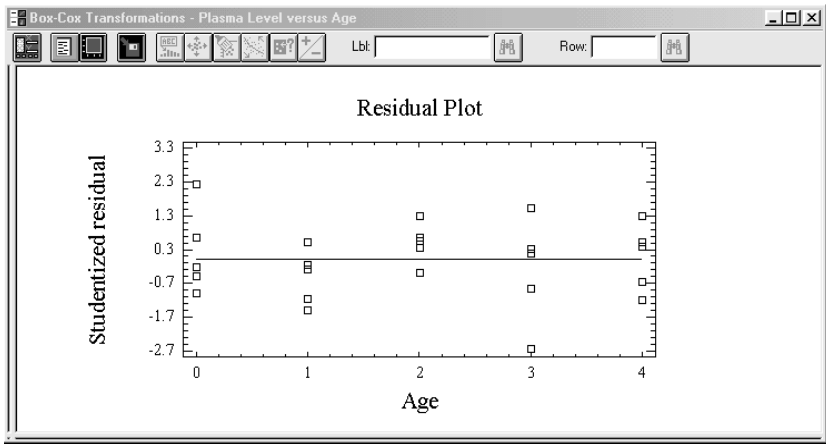


Figure 17-43. Residuals versus X Plot

Residuals versus Predicted

The Residuals versus Predicted option creates a plot of the studentized residuals versus the predicted values of the variable you selected (see Figure 17-44). Nonrandom patterns indicate that the model you selected does not adequately describe the observed data.

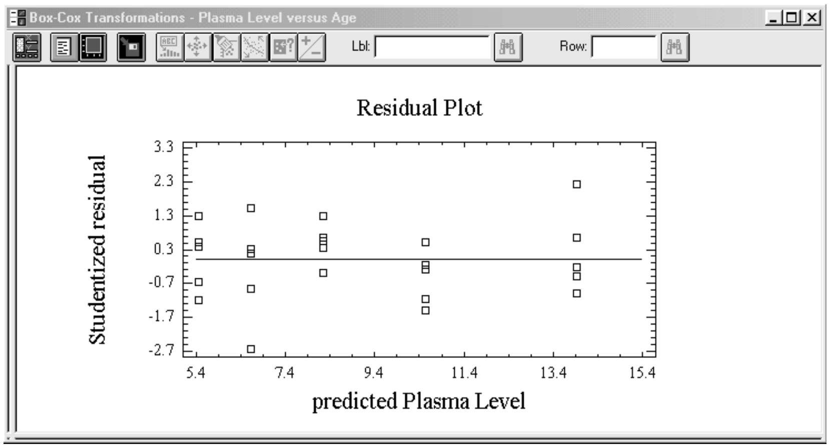


Figure 17-44. Residuals versus Predicted Plot

This plot is especially helpful in showing heteroscedasticity, in which the variability of the residuals change as the values of the dependent variables change. If the model is correct and if all the assumptions are satisfied, the residuals should be structureless.

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot (see Figure 17-14 for an example of this dialog box).

Residuals versus Row Number

The Residuals versus Row Number option creates a plot of the studentized residuals versus their row number (see Figure 17-45). If the row order corresponds to the order in which the data were collected, any nonrandom pattern could indicate serial correlation in the data. Ideally, the points should appear structureless.

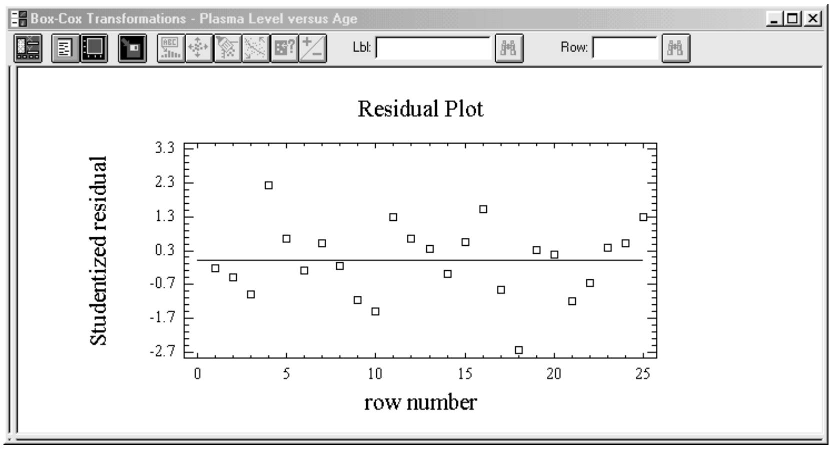


Figure 17-45. Residuals versus Row Number Plot

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot (see Figure 17-14 for an example of this dialog box).

Skewness and Kurtosis Plot

The Skewness and Kurtosis Plot option creates a plot that shows the values for the standardized skewness and standardized kurtosis at various values of Lambda1 (see Figure 17-46). A vertical line is drawn at the optimal value; horizontal lines are drawn at 0 and +/-2.

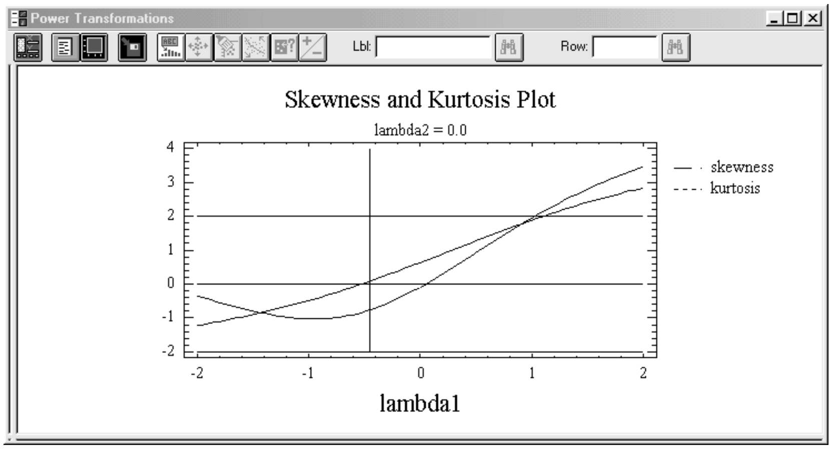
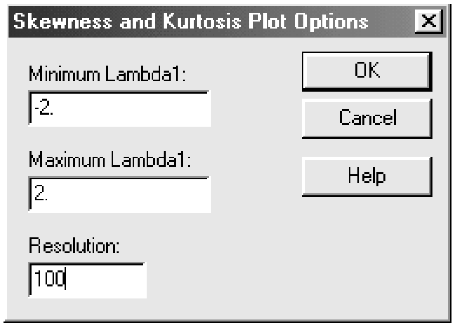


Figure 17-46 Skewness and Kurtosis Plot

Use the Skewness and Kurtosis Plot Options dialog box to set the values for which the MSE will be determined, and to enter a value for the resolution (see Figure 17-47).



Minimum Lambda1: -2.

Maximum Lambda1: 2.

Resolution: 100

Figure 17-47. Skewness and Kurtosis Plot Options Dialog Box

Saving the Results

The Save Results Options dialog box allows you to select the results you want to save. There are nine selections: Predicted Values, Lower Limits for Predictions, Upper Limits for Predictions, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Studentized Residuals, Leverages, and Transformed Data.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Myers, R. H. 1990. *Classical and Modern Regression with Applications*, second edition. Belmont, California: Duxbury Press.

Neter, J., Wasserman, W., and Kutner, M. H. 1989. *Applied Linear Regression Models*, second edition. Homewood, Illinois: Irwin.

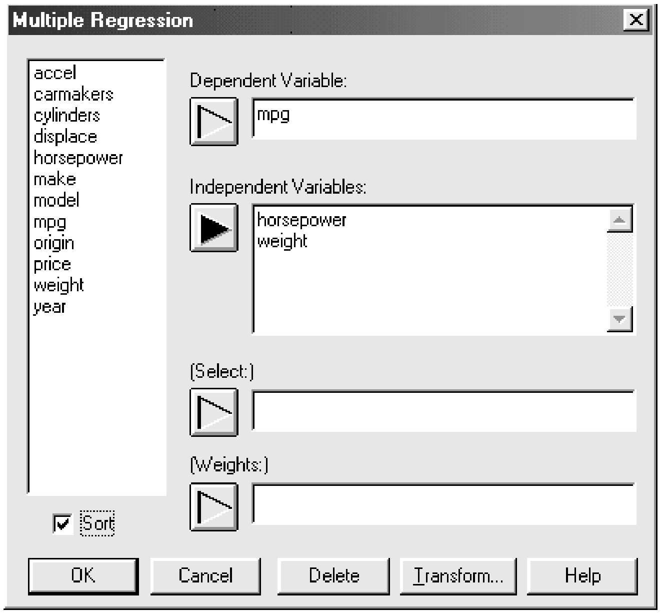
Using Multiple Regression Analysis

There are times in regression analysis when you want to draw conclusions about the relationship of variables as well as have a description of the observed data. Multiple regression is appropriate to use when the data consist of two or more numeric independent variables and a numeric dependent variable. The ability to include several independent variables also provides a method to use when there is a need to avoid the risk of oversimplifying the effect of one independent variable on one dependent variable.

Real explanatory power comes from the way interpretations are made about the regression coefficients. Multiple regression lets you estimate the effect of an independent variable on a dependent variable and, at the same time, control for the effects of all the other independent variables in the model. In other words, because there are multiple independent variables in the model, you can compare their effects and, in doing so, estimate which independent variable has the most explanatory power.

Basically, multiple regression includes specifying the model, interpreting the regression statistics, comparing the independent variables, and interpreting the model. Multiple regression analysis also allows you to use weights to force the intercept to 0 (regression through the origin), or to perform a stepwise regression. Like simple regression, multiple regression uses least squares to estimate the regression model.

To access the analysis, from the menus, choose: RELATE... MULTIPLE REGRESSION... (see Figure 17-48).



Multiple Regression

accel
carmakers
cylinders
displace
horsepower
make
model
mpg
origin
price
weight
year

☑ Sort

Dependent Variable: mpg

Independent Variables:
horsepower
weight

(Select:)

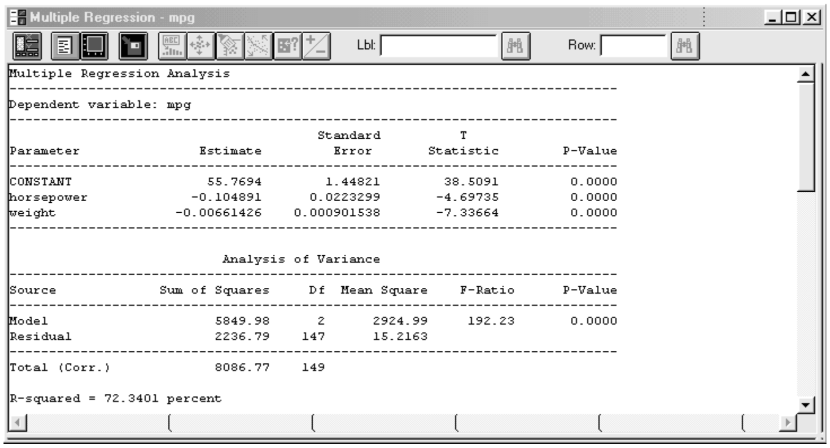
(Weights:)

Figure 17-48.   The Multiple Regression Analysis Dialog Box

Tabular Options

Analysis Summary

The Analysis Summary option creates a summary of the analysis that shows the results of fitting a multiple linear regression model to describe the relationship between the dependent variable and two or more independent variables (see Figure 17-49). The summary also shows the estimated coefficients, standard errors, *t*-values, and *p*-values.



```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: mpg
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 55.7694        1.44821        38.5091         0.0000
horsepower             -0.104891      0.0223299       -4.69735         0.0000
weight               -0.00661426    0.000901538       -7.33664         0.0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     5849.98      2      2924.99     192.23       0.0000
Residual                  2236.79    147      15.2163
-----------------------------------------------------------------------------
Total (Corr.)             8086.77    149

R-squared = 72.3401 percent
```

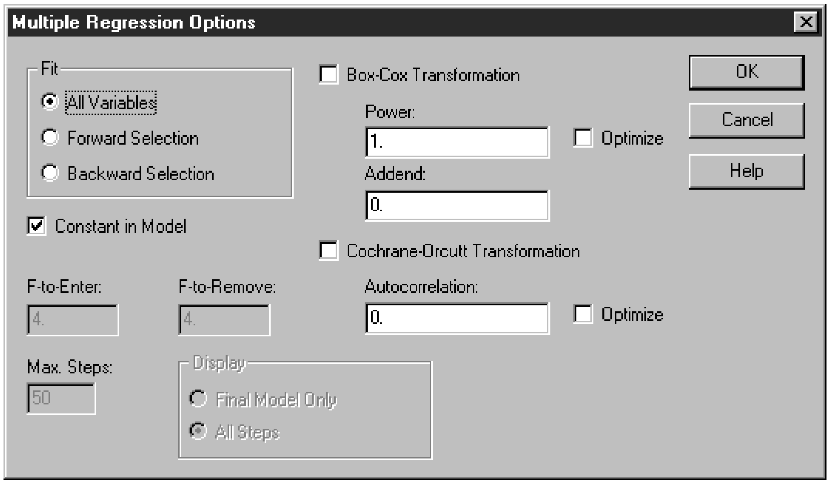
Figure 17-49. Analysis Smart

The *p*-values correspond to tests of the hypotheses that the coefficients are equal to 0 which, in turn, is equivalent to concluding that there is no linear relationship between the independent and dependent values. Small *p*-values (less than an alpha level such as .05) indicate statistically significant nonzero coefficients.

The R-Squared statistic explains the percentage of variability that is explained by the regression model. The Adjusted R-Squared statistic at the bottom of the ANOVA table adjusts the standard R-Squared value based on the number of coefficients in the model. This statistic is useful for comparing regression models that have different numbers of independent variables. The standard error of estimation is the estimated standard deviation of the residuals around the fitted line.

The mean absolute error is the mean of the absolute values of the residuals. The Durbin-Watson statistic tests the residuals to determine if there is any significant correlation based on the order in which they were entered into the file.

Use the *Multiple Regression Options* dialog box to indicate the independent variables that will be fit in the model and how they will be entered in the data file (see Figure 17-50). In addition, you can indicate if you want the constant included the model; enter values for the F-ratio at or above and below which you want to enter the variables into the model; enter values for the maximum number of steps that will be performed before the selection process stops; and indicate the method that will display the model for a stepwise regression.



Multiple Regression Options

Fit

● All Variables
○ Forward Selection
○ Backward Selection

☑ Constant in Model

F-to-Enter: 4.
F-to-Remove: 4.
Max. Steps: 50

Display

○ Final Model Only
● All Steps

□ Box-Cox Transformation

Power: 1. □ Optimize
Addend: 0.

□ Cochrane-Orcutt Transformation

Autocorrelation: 0. □ Optimize

Figure 17-50. Multiple Regression Options Dialog Box

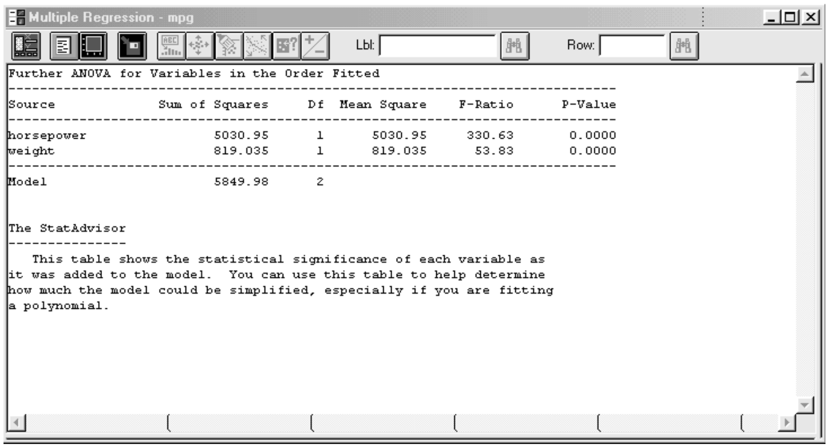
Additionally, you can enter values for the transformation parameters or request that the program find optimal values.

The Box-Cox Transformations handle only one X variable.

The Cochrane-Orcutt Transformation procedure handles sigutations in which the residuals are autocorrelated, as frequently occurs when the data is collected over time. It is an iterative procedure in which an ordinary regression model is fit and the residuals are saved. Then the first-order autocorrelation coefficient of the residuals is calculated and the regression model is refit on the transformed variables. If you select the Optimize option, recalculate the first-order autocorrelation coefficient of the residuals and refit the regression model on the transformed variables until the autocorrelation of the residuals from the transformed model is small.

Conditional Sums of Squares

The Conditional Sums of Squares option creates a table that displays the statistical significance of each variable as it was added to the model, which helps determine how much the model could be simplified (see Figure 17-51).



```
Further ANOVA for Variables in the Order Fitted
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
horsepower                 5030.95      1      5030.95     330.63       0.0000
weight                     819.035      1      819.035      53.83       0.0000
-----------------------------------------------------------------------------
Model                      5849.98      2


The StatAdvisor
---------------
   This table shows the statistical significance of each variable as
it was added to the model.  You can use this table to help determine
how much the model could be simplified, especially if you are fitting
a polynomial.
```

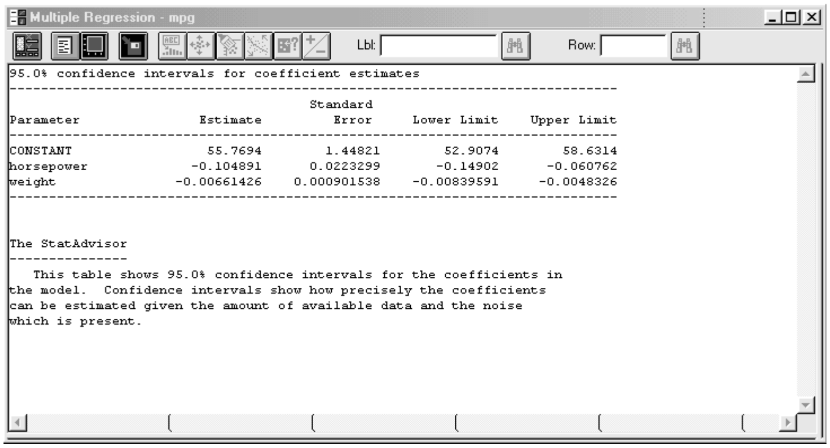
Figure 17-51.   Conditional Sums of Squares

If you select the All Variables option on the Multiple Regression Options dialog box, the results depend on the order of the variables as you entered them into the Analysis dialog box. For a stepwise regression, the results reflect the order in which the program entered the variables into the final model. The program calculates F-ratios by dividing the mean square terms by the mean square error from the primary analysis of variance table. The total sum of squares for the variables is the model sum of squares in the primary analysis of variance table.

Confidence Intervals

The Confidence Intervals option creates confidence intervals for the coefficient estimates (see Figure 17-52). The estimates and the standard errors are identical to those shown in the initial table of model-fitting results.

Use the *Confidence Intervals Options* dialog box to change the number used to calculate the confidence intervals.

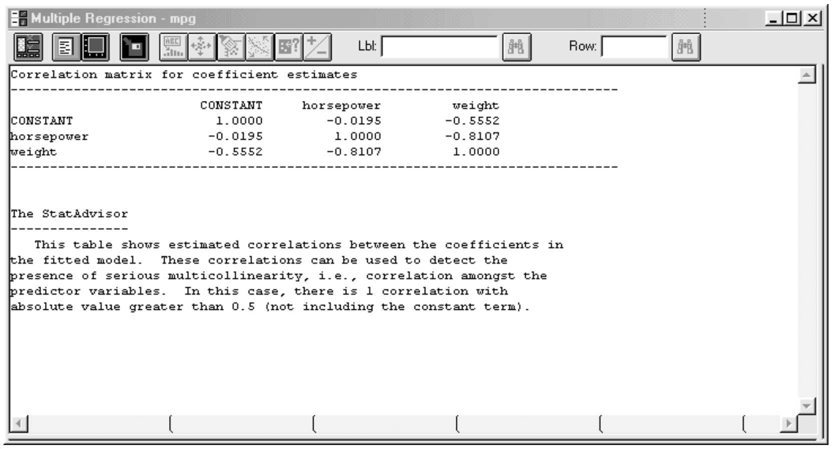


```
95.0% confidence intervals for coefficient estimates
-----------------------------------------------------------------------------
                                        Standard
Parameter               Estimate         Error     Lower Limit    Upper Limit
-----------------------------------------------------------------------------
CONSTANT                 55.7694        1.44821        52.9074        58.6314
horsepower             -0.104891      0.0223299       -0.14902      -0.060762
weight               -0.00661426    0.000901538    -0.00839591     -0.0048326
-----------------------------------------------------------------------------


The StatAdvisor
---------------
   This table shows 95.0% confidence intervals for the coefficients in
the model.  Confidence intervals show how precisely the coefficients
can be estimated given the amount of available data and the noise
which is present.
```

Figure 17-52.   Confidence Intervals

Correlation Matrix

The Correlation Matrix option creates a table that shows the estimated correlations between the coefficients in the fitted model (see Figure 17-53).



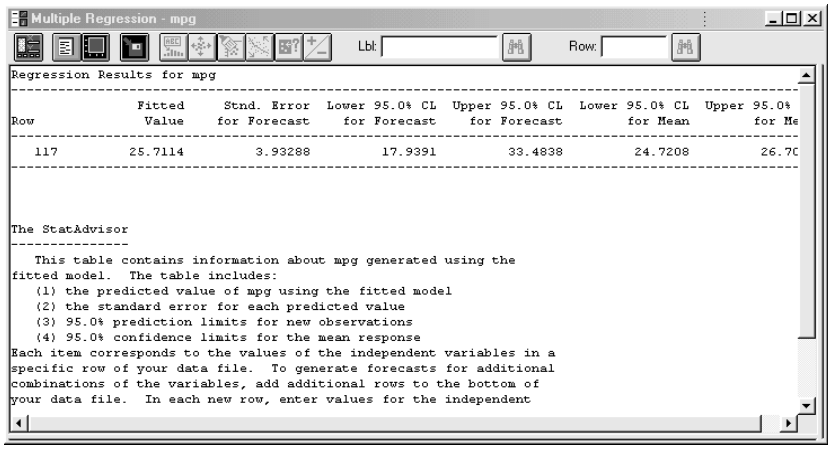
```
Correlation matrix for coefficient estimates
-----------------------------------------------------------------------------
                        CONSTANT     horsepower         weight
CONSTANT                  1.0000        -0.0195        -0.5552
horsepower               -0.0195         1.0000        -0.8107
weight                   -0.5552        -0.8107         1.0000
-----------------------------------------------------------------------------


The StatAdvisor
---------------
   This table shows estimated correlations between the coefficients in
the fitted model.  These correlations can be used to detect the
presence of serious multicollinearity, i.e., correlation amongst the
predictor variables.  In this case, there is 1 correlation with
absolute value greater than 0.5 (not including the constant term).
```

Figure 17-53.   Correlation Matrix

The correlations are helpful in detecting the presence of serious multicollinearity; that is, the correlation among the independent variables.

Reports

The Reports option creates a table that displays information about the dependent variable (see Figure 17-54). The table also includes values for the predicted value of the dependent variable using the fitted model; values for the standard error for each predicted value; the standard error for each predicted value; and the prediction and confidence limits at a given percentage for new observations and the mean response, respectively. Each item in the table corresponds to the values for the independent variables in a specific row of the data file.

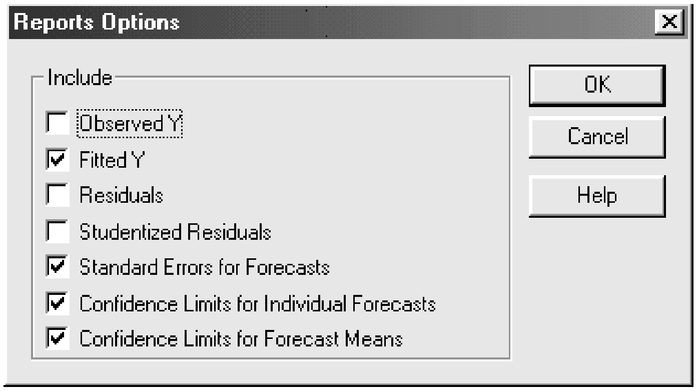


```
Regression Results for mpg
--------------------------------------------------------------------------------------------------------
                 Fitted     Stnd. Error  Lower 95.0% CL  Upper 95.0% CL  Lower 95.0% CL  Upper 95.0%
Row               Value    for Forecast    for Forecast    for Forecast        for Mean        for Me
--------------------------------------------------------------------------------------------------------
   117          25.7114         3.93288         17.9391         33.4838         24.7208         26.7C
--------------------------------------------------------------------------------------------------------



The StatAdvisor
---------------
   This table contains information about mpg generated using the
fitted model.  The table includes:
   (1) the predicted value of mpg using the fitted model
   (2) the standard error for each predicted value
   (3) 95.0% prediction limits for new observations
   (4) 95.0% confidence limits for the mean response
Each item corresponds to the values of the independent variables in a
specific row of your data file.  To generate forecasts for additional
combinations of the variables, add additional rows to the bottom of
your data file.  In each new row, enter values for the independent
```

Figure 17-54. Reports

If you want to create forecasts for additional combinations of the variables, add additional rows to the bottom of the data file. In each new row, enter values for the independent variables, but leave the cell for the dependent variable empty. When you return to this pane, the program will add the forecasts to this table for the new rows, but the model will remain unaffected.

Use the *Reports Options* dialog box to indicate the reports you want to include in the Reports table (see Figure 17-55).



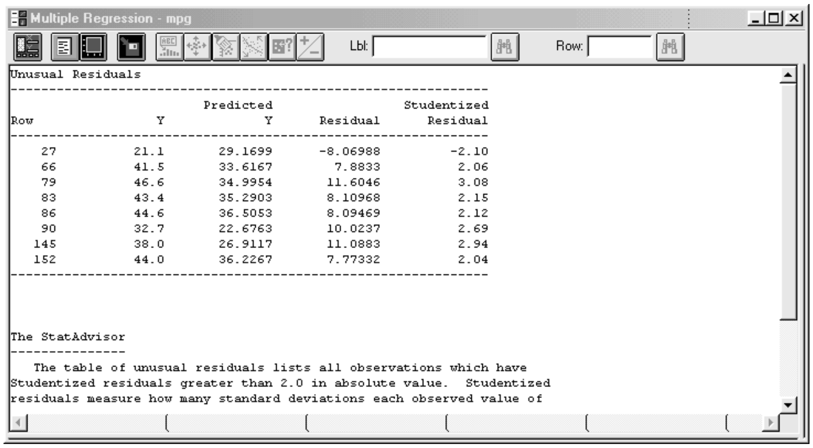
Include

- ☐ Observed Y
- ☑ Fitted Y
- ☐ Residuals
- ☐ Studentized Residuals
- ☑ Standard Errors for Forecasts
- ☑ Confidence Limits for Individual Forecasts
- ☑ Confidence Limits for Forecast Means

Figure 17-55. Reports Options Dialog Box

Unusual Residuals

The Unusual Residuals option creates a report that lists all the observations that have Studentized residuals greater than 2.0 in absolute value (see Figure 17-56). Studentized residuals measure the standard deviations each observed value of the dependent deviates from a model fitted using all the data except that observation. Look carefully at observations greater than 3.0 in absolute value to determine if they are outliers that should be removed from the model and handled separately.

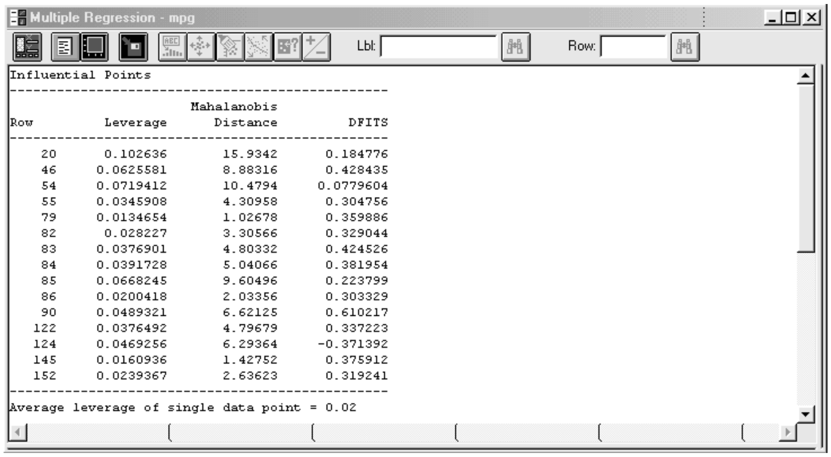


```
Unusual Residuals
--------------------------------------------------------------
                         Predicted                 Studentized
Row                 Y            Y      Residual      Residual
--------------------------------------------------------------
    27           21.1      29.1699      -8.06988         -2.10
    66           41.5      33.6167        7.8833          2.06
    79           46.6      34.9954       11.6046          3.08
    83           43.4      35.2903       8.10968          2.15
    86           44.6      36.5053       8.09469          2.12
    90           32.7      22.6763       10.0237          2.69
   145           38.0      26.9117       11.0883          2.94
   152           44.0      36.2267       7.77332          2.04
--------------------------------------------------------------



The StatAdvisor
---------------
   The table of unusual residuals lists all observations which have
Studentized residuals greater than 2.0 in absolute value.  Studentized
residuals measure how many standard deviations each observed value of
```

Figure 17-56. Unusual Residuals

Influential Points

The Influential Points option creates a table that displays all the observations that have leverage values greater than three times that of an average data point, or that have an unusually large DFITS value (see Figure 17-57). Leverage is a statistic that measures how influential each observation is in determining the coefficients of the estimated model. DFITS is a statistic that measures how much the estimated coefficients would change if each observation was removed from the dataset.



```
Influential Points
------------------------------------------------
                        Mahalanobis
Row         Leverage      Distance         DFITS
------------------------------------------------
    20      0.102636       15.9342      0.184776
    46     0.0625581       8.88316      0.428435
    54     0.0719412       10.4794     0.0779604
    55     0.0345908       4.30958      0.304756
    79     0.0134654       1.02678      0.359886
    82      0.028227       3.30566      0.329044
    83     0.0376901       4.80332      0.424526
    84     0.0391728       5.04066      0.381954
    85     0.0668245       9.60496      0.223799
    86     0.0200418       2.03356      0.303329
    90     0.0489321       6.62125      0.610217
   122     0.0376492       4.79679      0.337223
   124     0.0469256       6.29364     -0.371392
   145     0.0160936       1.42752      0.375912
   152     0.0239367       2.63623      0.319241
------------------------------------------------
Average leverage of single data point = 0.02
```

Figure 17-57.   Influential Points

Graphical Options

Component Effects

The Component Effects option creates a plot that shows the portion of the fitted model relating variables (see Figure 17-58). It also displays the equation of the line, which shows the relative change in the predicted values of the dependent variable that occur when changing the independent variable over its observed range.

Each point is plotted by adding its residual to the line. Examine the size of the residuals relative to the change in the predicted values of the response, so you can judge the importance of the selected independent variable.

Use the *Component Effects Plot Options* dialog box to indicate the independent variable that will be used for the plot.

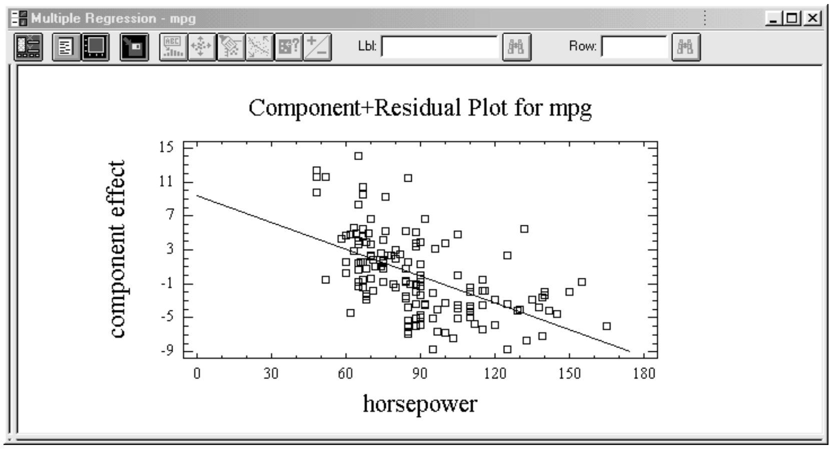


Figure 17-58. Component Effects Plot

Observed versus Predicted

The Observed versus Predicted option creates a plot of the observed values of the dependent variable versus the values predicted by the fitted model (see Figure 17-59). The closer the points lie to the diagonal line, the better the model is at predicting the observed data.

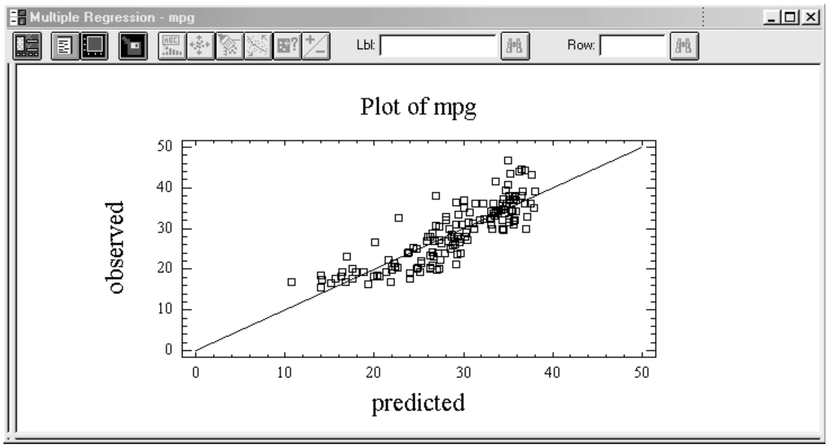


Figure 17-59. Observed versus Predicted Plot

Residuals versus X

The Residuals versus X option creates a plot of the residuals versus the predicted value for the dependent variable (see Figure 17-60). Any nonrandom pattern could indicate that the selected model does not adequately describe the observed data. In addition, any points outside the range of -3 to +3 are considered to be outliers.

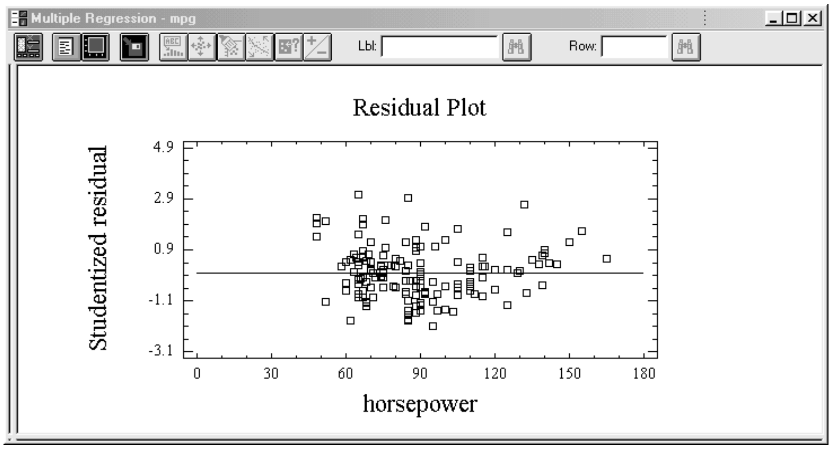


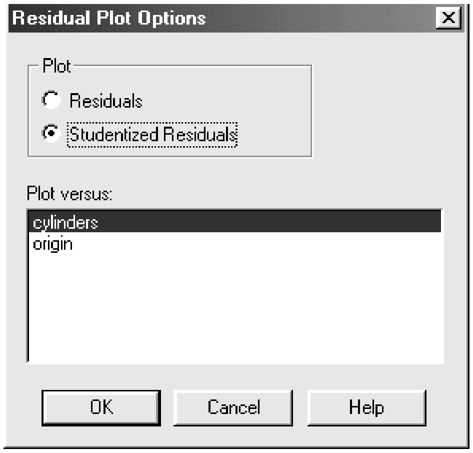
Figure 17-60 Residuals versus X Plot

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot and to select the independent variable that will be used on the plot (see Figure 17-61).

Residuals versus Predicted

The Residuals versus Predicted option creates a plot of the residuals versus the predicted values for the dependent variable (see Figure 17-62). Any nonrandom pattern could indicate that the selected model does not adequately describe the observed data. The plot is helpful in determining heteroscedasticity, in which the variability of the residuals changes as the values for the dependent variable change.

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot (see Figure 17-14 for an example of this dialog box).



Residual Plot Options

Plot

○ Residuals

● Studentized Residuals

Plot versus:

cylinders
origin

Figure 17-61. Residual Plot Options Dialog Box

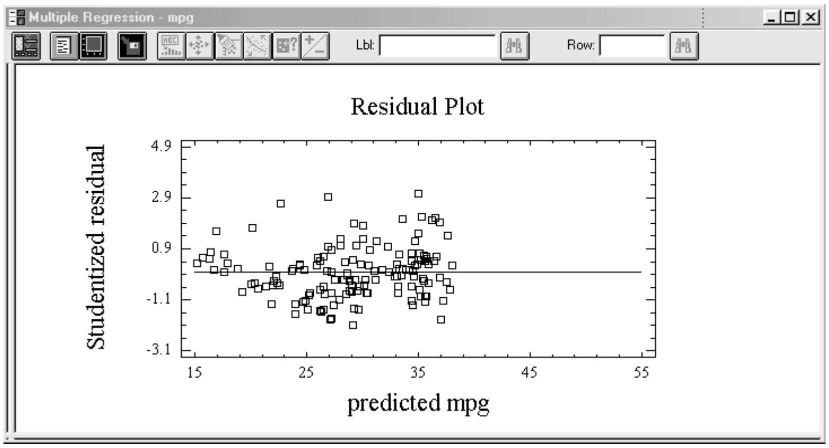


Figure 17-62. Residuals versus Predicted Plot

Residuals versus Row Number

The Residuals versus Row Number option creates a plot of the residuals versus the row number (see Figure 17-63). Any nonrandom pattern could indicate serial correlation in the data, if the row order corresponds to the order in which the data were collected.

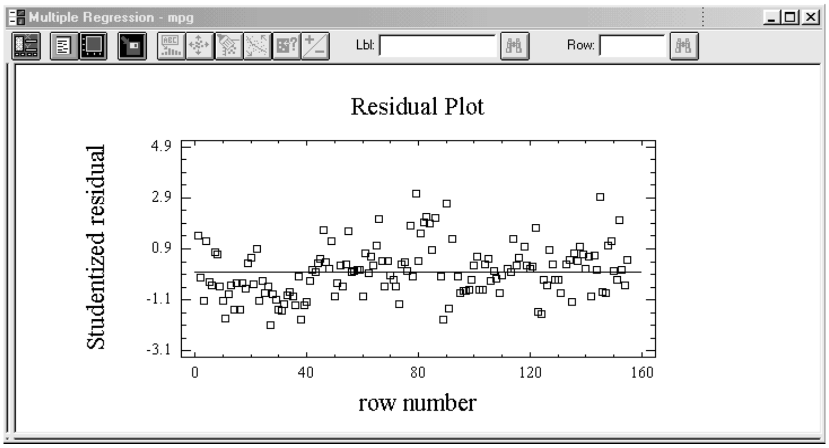


Figure 17-63. Residuals versus Row Number Plot

Use the *Residual Plot Options* dialog box to indicate the type of residuals that will appear on the plot (see Figure 17-14 for an example of this dialog box).

Interval Plots

The Interval Plots option creates a plot of the observed and predicted values of the dependent variable versus the values for the selected independent variable (see Figure 17-64).

Use the *Interval Plots Options* dialog box to select the type of confidence limits you want to plot, to select the independent variable that will be used on the plot, and to enter a value for the confidence level that will be used to calculate the confidence intervals (see Figure 17-65).

Saving the Results

The Save Results Options dialog box allows you to select the results you want to save. There are 12 selections: Predicted Values, Standard Errors of Predictions, Lower Limits for Predictions, Upper Limits for Predictions,

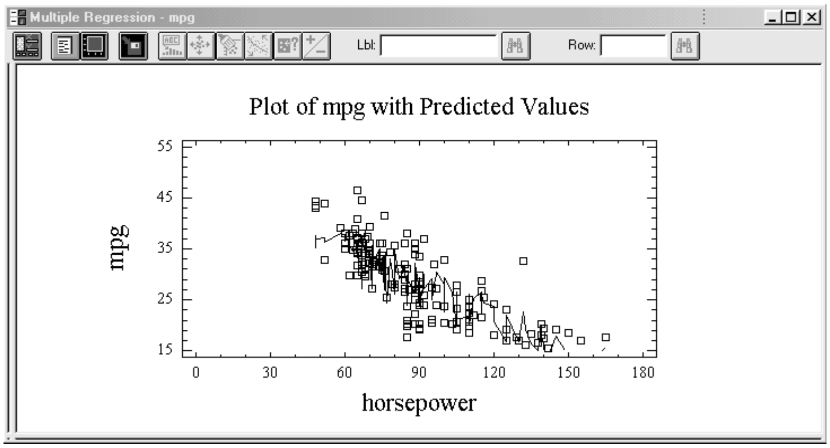
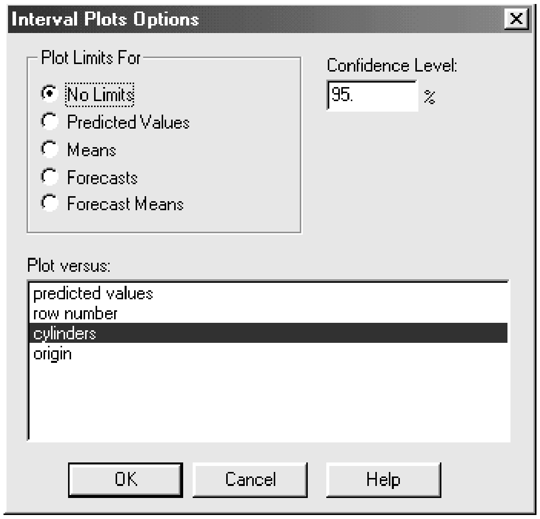


Figure 17-64.   Interval Plot


Interval Plots Options

Plot Limits For

- ● No Limits
- ○ Predicted Values
- ○ Means
- ○ Forecasts
- ○ Forecast Means

Confidence Level: 95. %

Plot versus:

- predicted values
- row number
- cylinders
- origin

Figure 17-65.   Interval Plot Options Dialog Box

Standard Errors of Means, Lower Limits for Forecast Means, Upper Limits for Forecast Means, Residuals, Studentized Residuals, Leverages, DFITS Statistics, and Mahalanobis Distances.

You can also use the Target Variables text boxes to enter the names of the variables in which you want to save the values generated during the analysis. You can enter new names or accept the defaults.

Note: To access the Save Results Options dialog box, click the Save Results Options button on the Analysis toolbar (the fourth button from the left).

References

Belsley, D. A., Kuh, E., and Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity*. New York: Wiley.

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: Wiley.

Durbin, J. and Watson, G. S. 1951. "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, **38**.

Kvalseth, T. O. 1985. "Cautionary Note about R^2," *The American Statistician*, **39**.

Montgomery, D. C. and Peck, E. A. 1992. *Introduction to Linear Regression Analysis*, second edition. New York: Wiley & Sons.

Neter, J., Wasserman, W., and Kutner, M. 1985. *Applied Linear Statistical Methods*. Homewood, Illinois: Richard E. Erwin, Inc.

18 Using SnapStats

This chapter discusses SnapStats. SnapStats are analyses designed to provide one-page summaries for commonly encountered data analysis problems.

You can choose from among nine SnapStats:

- One Sample Analysis (see *Using the One-Variable Analysis* in Chapter 9 for more information).
- Two Sample Comparison (see *Using the Two-Sample Comparison Analysis* in Chapter 14 for more information).
- Paired Sample Comparison (see *Using the Paired-Sample Comparison Analysis* in Chapter 14 for more information).
- Multiple Sample Comparison (see *Using the Multiple-Sample Comparison Analysis* Chapter 15 for more information).
- Curve Fitting (see *Using Simple Regression* in Chapter 17 for more information).
- Capability Assessment (Individuals) (see the advanced Quality and Design Manual for more information).
- Capability Assessment (Grouped Data) (see the advanced Quality and Design Manual for more information).
- Gage R&R (see the advanced Quality and Design Manual for more information).
- Automatic Forecasting (see the advanced Time Series Manual for more information).

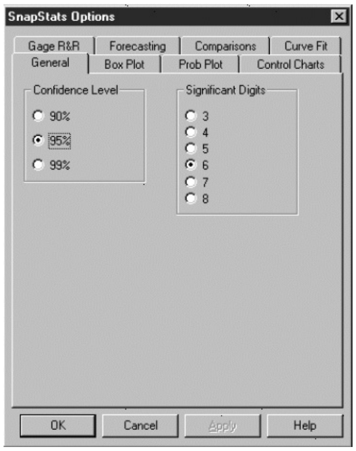
Each is a streamlined version of an existing analysis. However, they create a window with a single pane, a one page summary of the complete analysis.

Tables and graphs are laid out in a standard manner, which when printed yield a single printed page. See Figure 8-3 for an example.

There are only two options, available through either:

- **Graphics Options** - using the right-mouse button, you can set various features such as colors and scaling. The choices apply to all graphs in the window.

- **SnapStat Options** - the second button on the analysis toolbar allows you to set global preferences for your SnapStats. The first tab sets *General* options that apply to all SnapStats (see Figure 18-1).



Confidence Level
○ 90%
● 95%
○ 99%

Significant Digits
○ 3
○ 4
○ 5
● 6
○ 7
○ 8

Figure 18-1. SnapStats Options Dialog Box

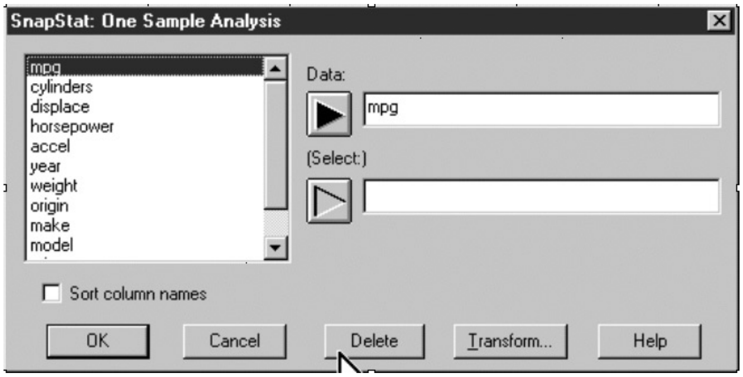
The fields include

- Confidence Level - applies to all confidence intervals, hypothesis tests, and interpretation by the StatAdvisor.
- Significant Digits - applies to the display of all numerical results.

Using SnapStat 1: One Sample Analysis

To Access the SnapStat One Sample Analysis, from the menus, choose: SNAPSTATS!!…ONE SAMPLE ANALYSIS…(see Figure 18-2).

The One Sample SnapStat provides a one-page summary for a single column of numeric data. It calculates statistics, performs hypothesis tests and constructs four graphs. The summary statistics include the variable name and sample size, the sample average, median, and standard deviation, minimum and maximum observations, and the standardized skewness and kurtosis values. Also provided is a confidence interval for the mean and standard deviation of the variable, the Shapiro-Wilkes test, and lag 1 autocorrelation.

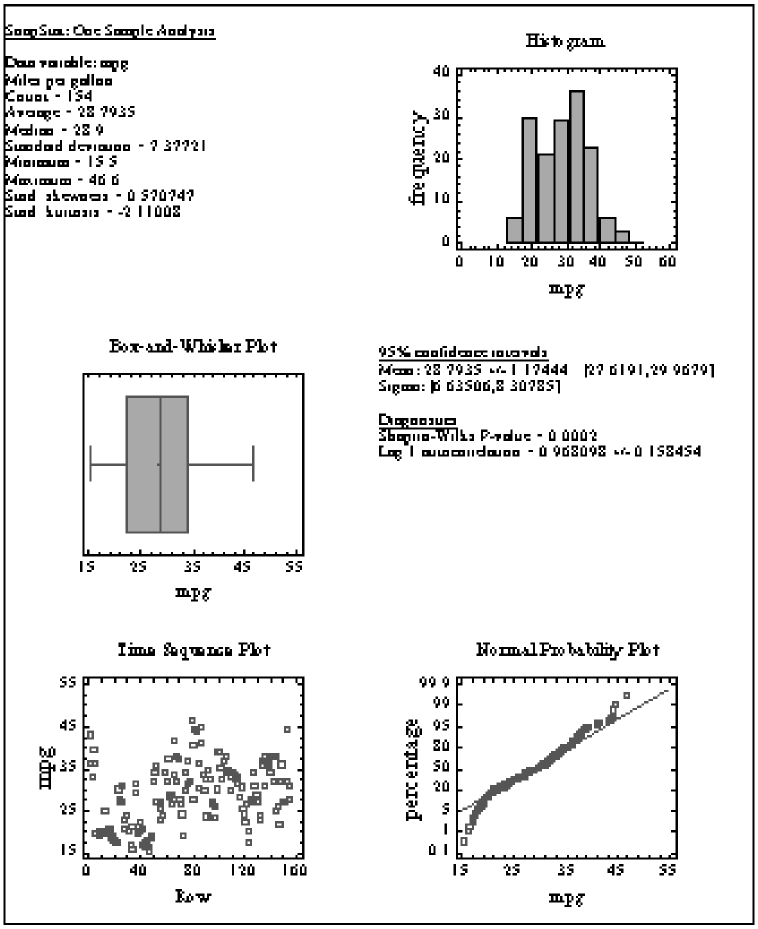


Data: mpg

☐ Sort column names

Figure 18-2. SnapStat: One Sample Analysis Input Dialog Box

Four graphs are provided to summarize the data: Frequency Histogram, a Box-and-Whisker Plot, a Time Sequence Plot and a Normal Probability Plot (see Figure 18-3).



SnapStat: One Sample Analysis

Data variable: mpg
Miles per gallon
Count = 154
Average = 28.7935
Median = 28.9
Standard deviation = 7.37721
Minimum = 15.5
Maximum = 46.6
Stnd. skewness = 0.570747
Stnd. kurtosis = -2.11008

95% confidence intervals
Mean: 28.7935 +/- 1.17444   [27.6191,29.9679]
Sigma: [6.63506,8.30785]

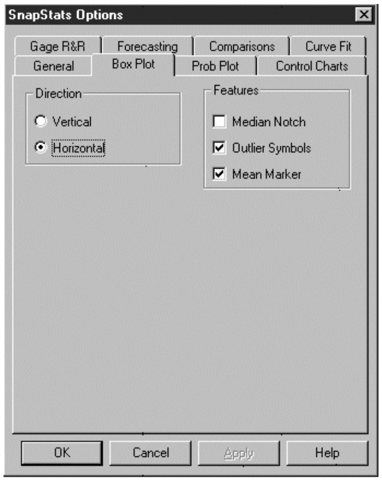
Diagnostics
Shapiro-Wilks P-value = 0.0002
Lag 1 autocorrelation = 0.968098 +/- 0.158454

Figure 18-3.   SnapStat: One Sample Analysis

Options include the Box Plot Tab (see Figure 18-4) and the Prob Plot Tab (see Figure 18-5).

Fields include

- Direction - determines the direction of the plot.
- Features - selects features to be added to the basic plot.



Direction

○ Vertical

● Horizontal

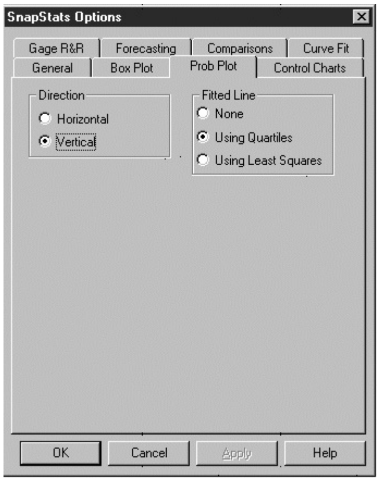
Features

□ Median Notch

☑ Outlier Symbols

☑ Mean Marker

Figure 18-4. Box Plot Tab



Direction

○ Horizontal

● Vertical

Fitted Line

○ None

● Using Quartiles

○ Using Least Squares

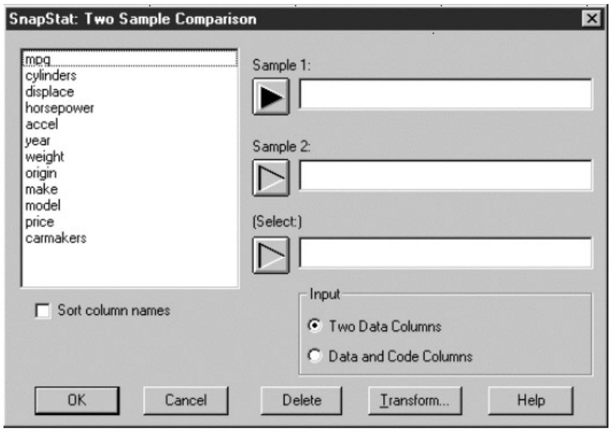
Figure 18-5. Prob Plot Tab

Fields include:

- Direction - determines the direction of the plot.
- Fitted line - selects the method for determining the reference line.

Using SnapStat 2: Two Sample Comparison

To access the SnapStat Two Sample Comparison, from the menus, choose: SNAPSTATS!!…TWO SAMPLE COMPARISON…(see Figure 18-6).



□ Sort column names

Input

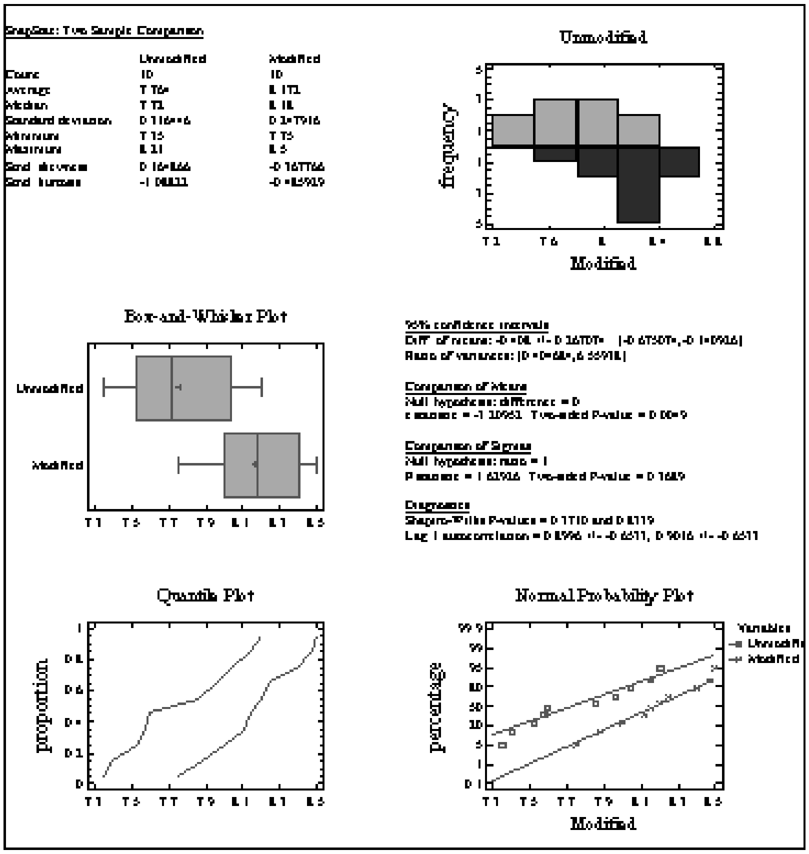
● Two Data Columns

○ Data and Code Columns

Figure 18-6. SnapStat: Two Sample Comparision Input Dialog Box

The output for the *Two Sample Comparison SnapStat* creates a summary of the analysis that includes the names of the variables for each of the two samples; the number of values in each sample; the average, variance, and standard deviation; and the minimum and maximum values. It also includes values for the standardized skewness and standardized kurtosis (see Figure 18-7). The standard skewness and standardized kurtosis values are of particular interest because they help to determine if the samples are from normal distributions. Values outside the range of -2 to +2 indicate a significant departure from normality, which tends to invalidate the tests that compare the standard deviations.

The primary purpose of the Two-Sample Comparison Analysis is to calculate confidence intervals for the difference between the population means and the ratio of the population variances, and to compare hypothesis tests of the means and variances. The analysis runs tests to determine if there are statistically significant differences between the two samples, then creates various reports and graphs that contain the results for each data sample.



```
SnapStat: Two Sample Comparison

                     Unmodified     Modified
Count                10             10
Average              [illegible]    [illegible]
Median               [illegible]    [illegible]
Standard deviation   [illegible]    [illegible]
Minimum              [illegible]    [illegible]
Maximum              [illegible]    [illegible]
Stnd. skewness       [illegible]    [illegible]
Stnd. kurtosis       [illegible]    [illegible]
```

```
95% confidence intervals
Diff. of means: [illegible]
Ratio of variances: [illegible]

Comparison of Means
Null hypothesis: difference = 0
t statistic = [illegible]  Two-sided P-value = [illegible]

Comparison of Sigmas
Null hypothesis: ratio = 1
F statistic = [illegible]  Two-sided P-value = [illegible]

Diagnostics
Shapiro-Wilks P-values = [illegible] and [illegible]
Lag 1 autocorrelation = [illegible]
```

Figure 18-7. SnapStat: Two Sample Comparison

Using SnapStat 3: Paired Sample Comparison

To Access the SnapStat Paired Sample Comparison, from the menus, choose: SNAPSTATS!!…PAIRED SAMPLE COMPARISON…(see Figure 18-8).

The Paired-Sample Comparison SnapStat parallels the Two-Sample Comparison SnapStat, except it compares data collected in pairs; that is, you want to compare two sets of data because the observations in the samples are paired. For example, suppose you are analyzing data that contain information about the miles per gallon ratings achieved by the same automobile during highway driving and city driving. The data represent two different sets of values for each automobile: miles per gallon ratings for highway driving and miles per gallon ratings for city driving. When data are from paired samples, you analyze it by calculating the difference between each value in each pair of observations. That is, you treat the paired differences as though they were a single sample from a population of differences.

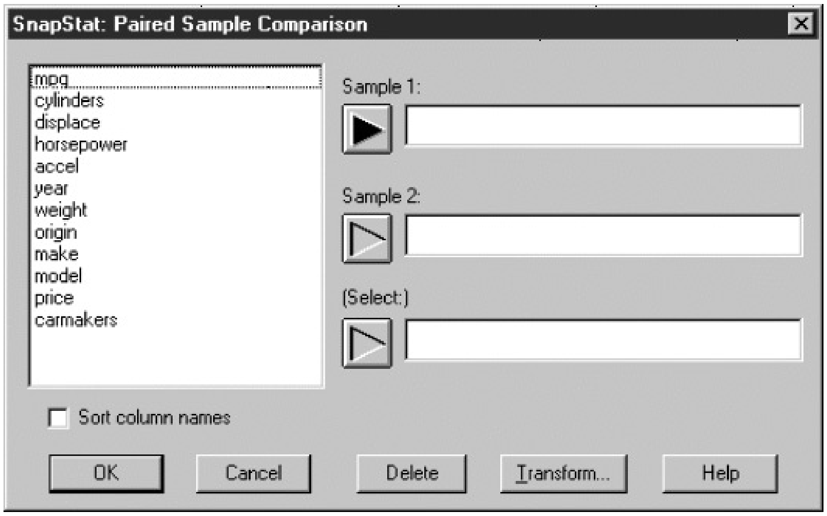
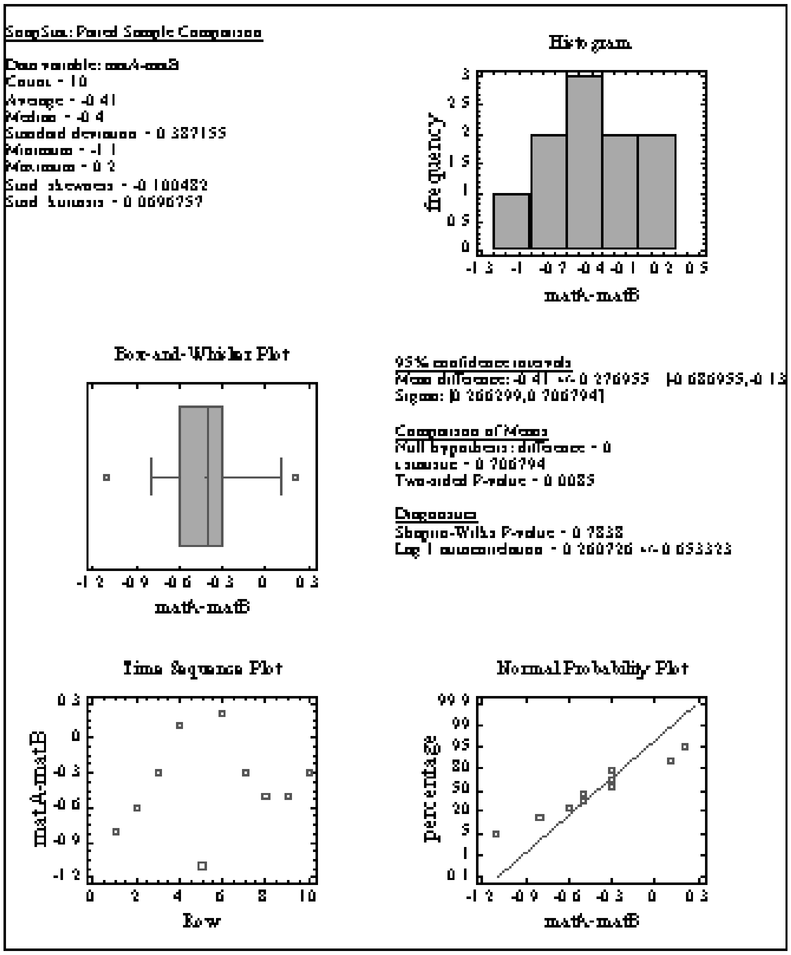


Figure 18-8. SnapStat: Paired Sample Comparison Input Dialog Box

The analysis performs tests for significant differences between two data samples where the data are collected as pairs. The summary includes the results of tests that determine if the mean difference is equal to zero. Also provided is a confidence interval for the mean and standard deviation of the variable, the Shapiro-Wilkes test, and lag 1 autocorrelation.

The analysis creates summary statistics for two pairs of data. The statistical information includes the number of values in each variable (Count); the average, and standard deviation; and the minimum and maximum values. It also includes values for the standardized skewness and standardized kurtosis (see Figure 18-9).



SnapStat: Paired Sample Comparison

Data variable: matA-matB
Count = 10
Average = -0 41
Median = -0 4
Standard deviation = 0 387155
Minimum = -1 1
Maximum = 0 2
Stnd skewness = -0 100482
Stnd kurtosis = 0 0690757

95% confidence intervals
Mean difference: -0 41 +/- 0 276955 [-0 686955,-0 13
Sigma: [0 266299,0 706794]

Comparison of Means
Null hypothesis: difference = 0
t statistic = 0 706794
Two-sided P-value = 0 0085

Diagnostics
Shapiro-Wilks P-value = 0 7838
Lag 1 autocorrelation = 0 260726 +/- 0 653323

Figure 18-9. SnapStat: Paired Sample Comparison

Using SnapStat 4: Multiple Sample Comparison

To Access the SnapStat Multiple Sample Comparison, from the menus, choose: SNAPSTATS!!…MULTIPLE SAMPLE COMPARISON…(see Figure 18-10).

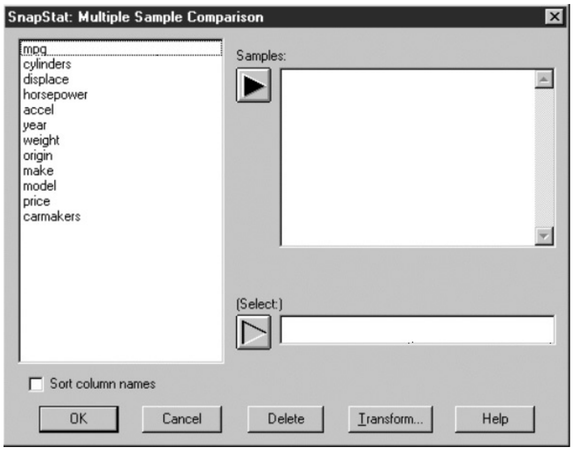


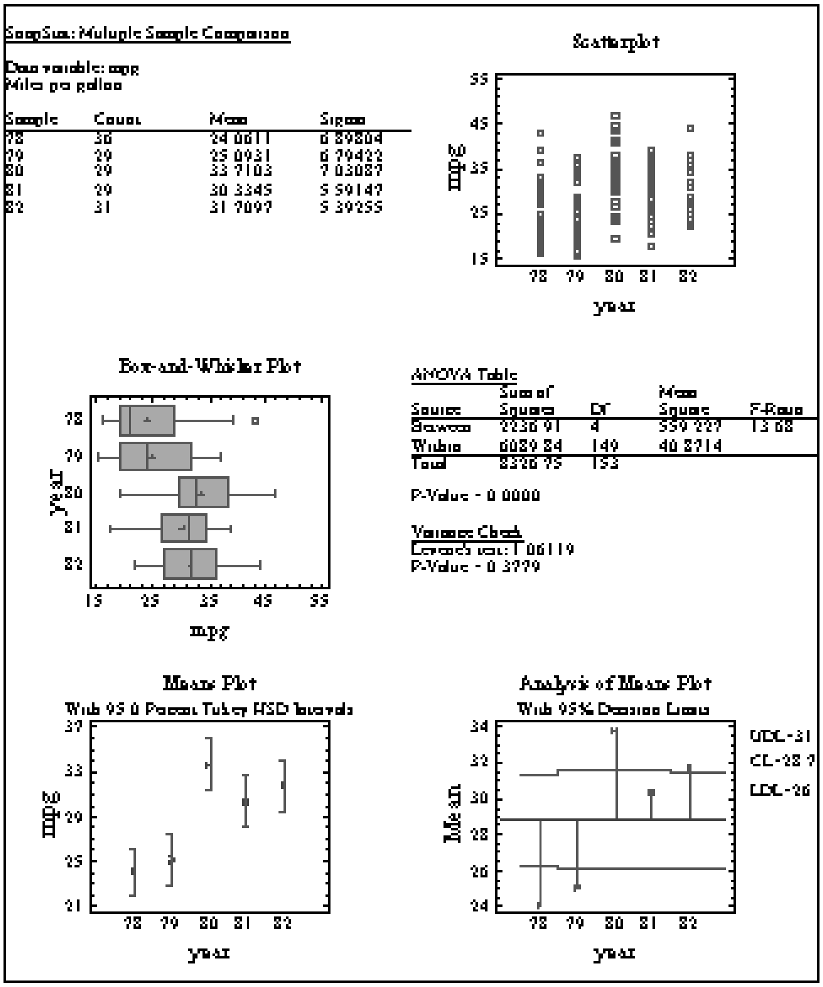
Figure 18-10 SnapStat: Multiple Sample Comparison Input Dialog Box

The *Multiple Sample Comparison* SnapStat parallels the Multiple Sample Comparison Analysis, which compares two or more samples to test the probability that when the null hypothesis is true, at least one of the observed significance levels will be less than a specified number. The more comparisons you make, the more likely it is that one or more pairs will be statistically different, even if all the population means are equal. Special tests in the analysis determine the means that are different, and/or the means that are the smallest or largest.

Note: The first dialog box allows you to choose the type of data you want to use; the second dialog box is the usual Analysis dialog box you use to enter the data you want to analyze.

The analysis creates a summary which shows the names of the variables in each sample, as well as the number of values in each variable (the Count); the average and standard deviation (see Figure 18-11).

The ANOVA Table is a standard analysis of variance (ANOVA) table and decomposes the variance of the data into two components: Between-Groups and Within Groups. The F-ratio is the mean square value for Between Groups divided by the mean square value for Within Groups. If the *p*-value is less than a given value, there is a statistically significant difference between the means of the variables at a given confidence level. The Sum of Squares - Between Groups statistic is the measure of variability among the different samples. The Sum of Squares - Within Groups statistic is the measure of variability within each of the samples. The Total is the measure of variability



SnapStat: Multiple Sample Comparison

Data variable: mpg
Miles per gallon

Sample	Count	Mean	Sigma
78	36	24 0611	6 89804
79	29	25 0931	6 79422
80	29	33 7103	7 03087
81	29	30 3345	5 59147
82	31	31 7097	5 39255

ANOVA Table

	Sum of		Mean	
Source	Squares	Df	Square	F-Ratio
Between	2236 91	4	559 227	13 68
Within	6089 84	149	40 8714	
Total	8326 75	153		

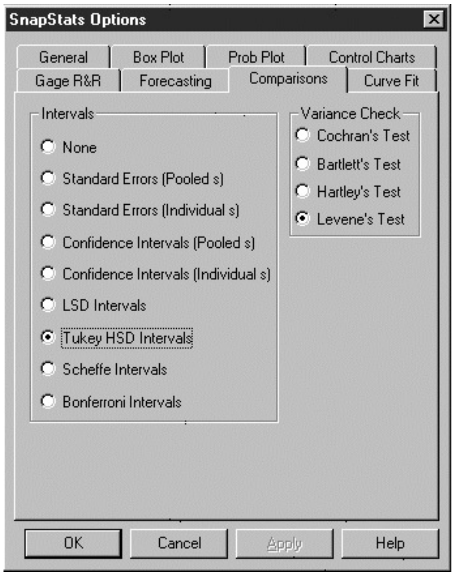
P-Value - 0 0000

Variance Check
Levene's test: 1 06119
P-Value - 0 3779

Figure 18-11. SnapStat: Multiple Sample Comparison

for all the data around the grand mean. Each Mean Square is the Sum of Squares for the source of the variation divided by the degrees of freedom (df). The *p*-values indicate the significance level. Small significance levels (less than .05 for most practical applications) indicate that the averages of the samples differ significantly.

The Option includes the Comparisons Tab (see Figure 18-12).



Intervals

○ None
○ Standard Errors (Pooled s)
○ Standard Errors (Individual s)
○ Confidence Intervals (Pooled s)
○ Confidence Intervals (Individual s)
○ LSD Intervals
● Tukey HSD Intervals
○ Scheffe Intervals
○ Bonferroni Intervals

Variance Check

○ Cochran's Test
○ Bartlett's Test
○ Hartley's Test
● Levene's Test

Figure 18-12. Comparisons Tab

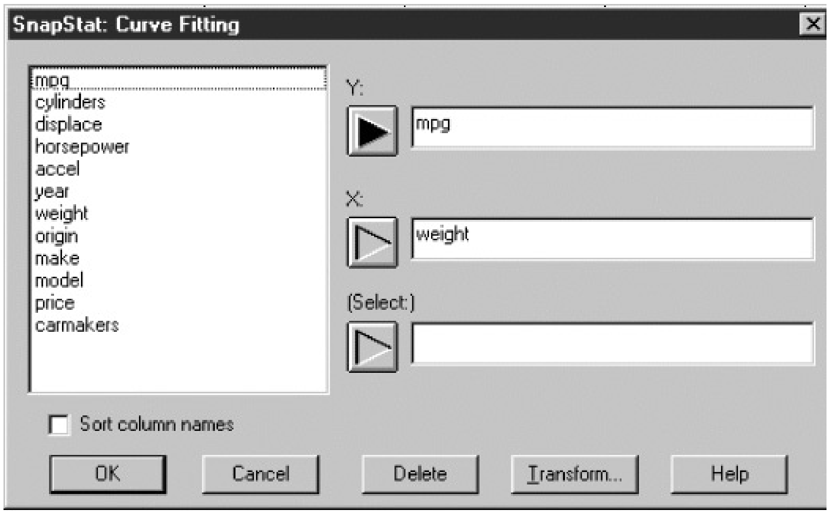
Fields include:

- Intervals - determines the type of intervals plotted.
- Variance Check - determines the test(s) displayed.

Using SnapStat 5: Curve Fitting

To Access SnapStat Curve Fitting, from the menus, choose SNAPSTATS!!…CURVE FITTING…(see Figure 18-13).

The SnapStat Curve Fitting Analysis parallels the Simple Regression procedure. The analysis fits a model that relates one dependent variable to one independent variable by minimizing the sum of the squares of the residuals for the fitted line. The summary of the analysis shows the results of fitting a linear model that describes the relationship between the two variables. The equation of the fitted linear model is: Y = a + b*X (see Figure 18-14). In the regression model, the estimate for the intercept is the value for “a”; the estimate for the slope is the value of “b. The summary includes the standard error, *t*-statistic, and *p*-value for each estimate. The *t*-statistic tests to see if the true value of the coefficient is equal to 0, which is equivalent to concluding that there is no linear relationship between the independent and dependent variables. Low probability levels (less than .05 for a 95 percent confidence level) for the intercept and slope suggest significant values.



SnapStat: Curve Fitting

Y: mpg

X: weight

(Select:)

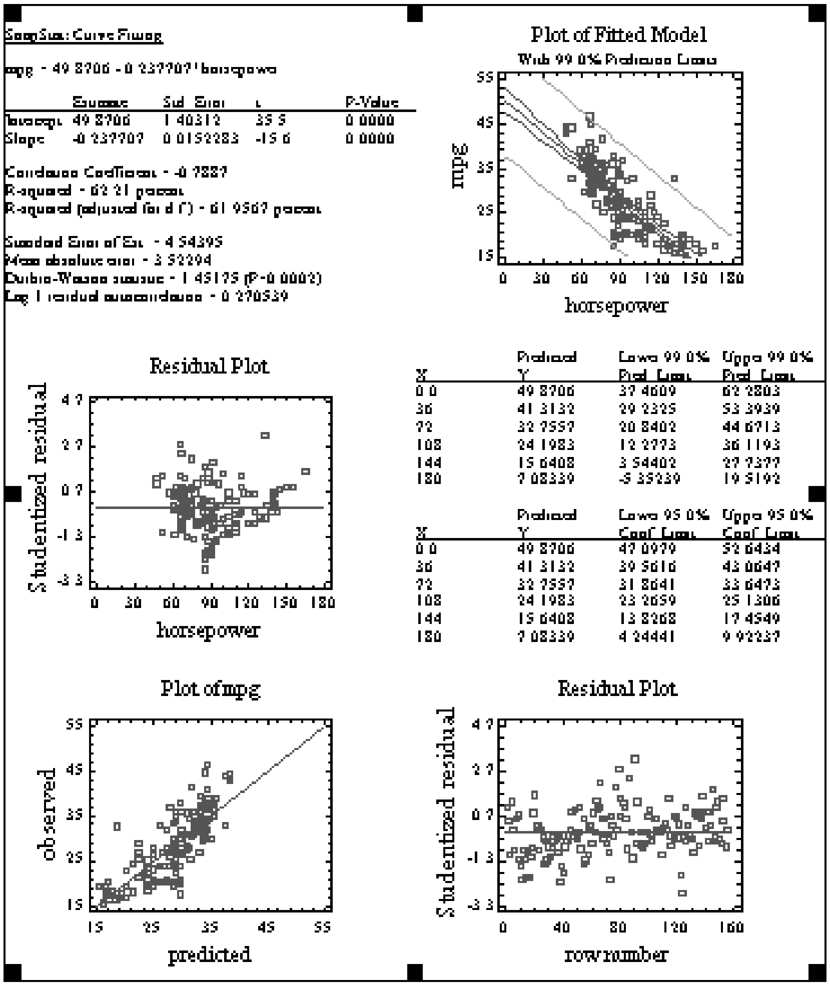
□ Sort column names

Figure 18-13. SnapStat: Curve Fitting Input Dialog Box

The summary also includes an ANOVA Table for the model. The F-ratio in the summary indicates the significance of the results; a high F-ratio suggests a significant model. The summary also shows the correlation coefficient that measures the linear relationship between the independent and dependent variables, the R-Squared value (the square of the correlation coefficient, expressed as a percentage), adjusted R-square (adjusted for degrees of freedom) and the standard error of estimation. The standard error of estimation is the square root of the residual mean square, which is the estimated standard deviation of the variable that is not explained by the estimated model. The R-Squared value shows the percentage of variability in the Y values explained by the X variable.

The forecasts table contains the predicted values for Y using the fitted model. The table displays the values for the prediction limits and confidence limits.

The prediction and confidence intervals correspond to the Inner and Outer bounds on the plot of the fitted model.



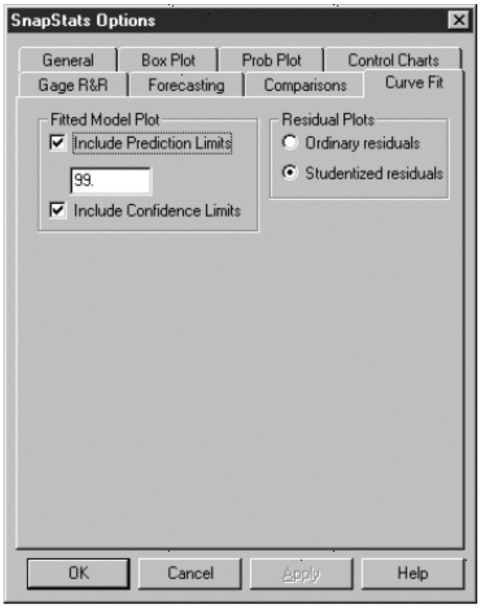
```
SnapStat: Curve Fitting

mpg = 49.8706 - 0.237707*horsepower

            Estimate     Std. Error    t        P-Value
Intercept   49.8706      1.40312       35.5     0.0000
Slope       -0.237707    0.0152283     -15.6    0.0000

Correlation Coefficient = -0.7887
R-squared = 62.21 percent
R-squared (adjusted for d.f.) = 61.9567 percent

Standard Error of Est. = 4.54395
Mean absolute error = 3.52294
Durbin-Watson statistic = 1.45175 (P=0.0002)
Lag 1 residual autocorrelation = 0.270539
```

X	Predicted Y	Lower 99.0% Pred. Limit	Upper 99.0% Pred. Limit
0.0	49.8706	37.4609	62.2803
36	41.3132	29.2325	53.3939
72	32.7557	20.8402	44.6713
108	24.1983	12.2773	36.1193
144	15.6408	3.54402	27.7377
180	7.08339	-5.35239	19.5192

X	Predicted Y	Lower 95.0% Conf. Limit	Upper 95.0% Conf. Limit
0.0	49.8706	47.0979	52.6434
36	41.3132	39.5616	43.0647
72	32.7557	31.8641	33.6473
108	24.1983	23.2659	25.1306
144	15.6408	13.8268	17.4549
180	7.08339	4.24441	9.92237

Figure 18-14. SnapStat: Curve Fitting

Option includes the Curve Fit Tab (see Figure 18-15).



Fitted Model Plot
☑ Include Prediction Limits
99.
☑ Include Confidence Limits

Residual Plots
○ Ordinary residuals
● Studentized residuals

Figure 18-15. Curve Fit Tab

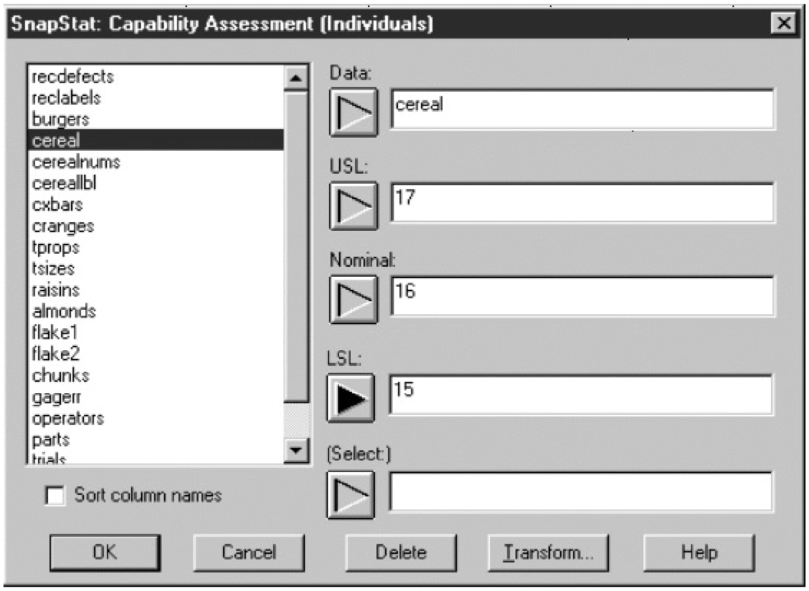
Fields include

- Fitted Model Plot - desired limits (if any) for the plot of the fitted model.
- Residual Plots - type of residuals to be plotted.

Using SnapStat 6: Capability Assessment (Individuals)

To Access the SnapStat Capability Assessment (Individuals), from the menus, choose: SNAPSTATS!! … CAPABILITY ASSESSMENT (INDIVIDUALS)…(see Figure 18-16).

The Capability Assessment SnapStat is identical to the Process Capability Analysis. The summary includes the name of the data variable, the values for



SnapStat: Capability Assessment (Individuals)

Data: cereal

USL: 17

Nominal: 16

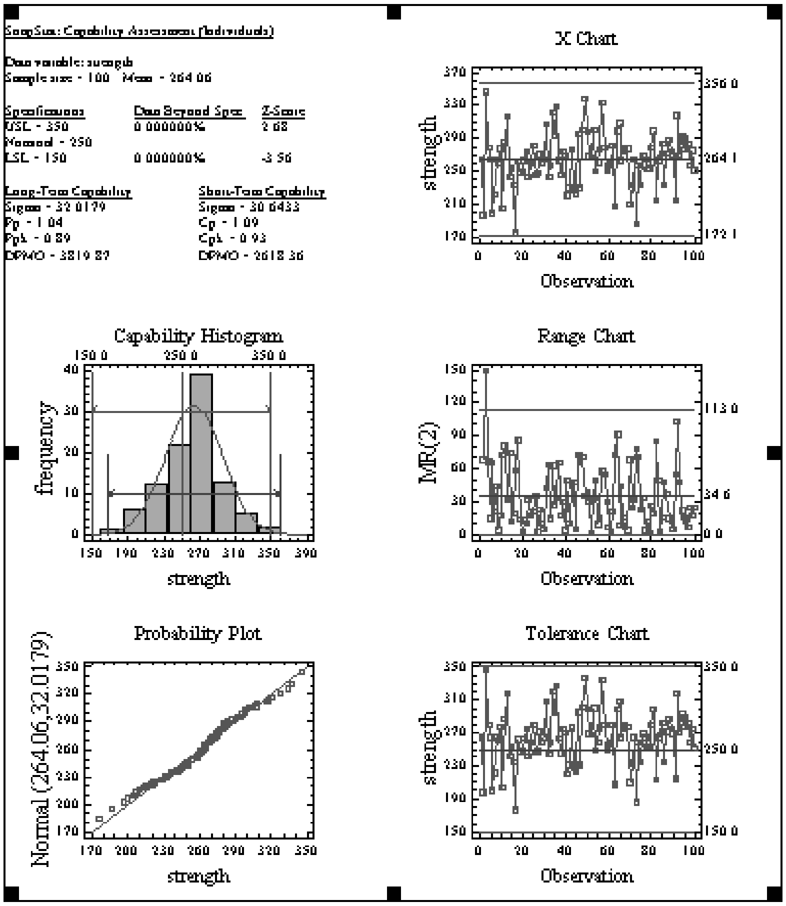
LSL: 15

(Select:)

☐ Sort column names

Figure 18-16. SnapStat: Capability Assessment (Individuals) Input Dialog Box

the sample size, and the parameters for the fitted distribution (see Figure 18-17). The table shows the percent of observations beyond the specification limits, the z-scores, and the percent of estimated observations beyond the specification limits calculated from the fitted distribution. The values for the USL, Nominal, and LSL specifications are also shown. For a normal distribution, the z-score indicates the number of standard deviations each limit is away from the sample mean.. Capable processes have USL and LSL z-scores greater than +3 and less than -3, respectively. The Observed Beyond Specifications and Estimated Beyond Specifications columns represent the total percentage of the area under the fitted normal curve that fall above and below the specification limits for the distribution.



```
SnapStat: Capability Assessment (Individuals)

Data variable: strength
Sample size = 100   Mean = 264.06

Specifications      Data Beyond Spec    Z-Score
USL = 350           0.000000%           2.68
Nominal = 250
LSL = 150           0.000000%           -3.56

Long-Term Capability          Short-Term Capability
Sigma = 32.0179               Sigma = 30.6433
Pp = 1.04                     Cp = 1.09
Ppk = 0.89                    Cpk = 0.93
DPMO = 3819.87                DPMO = 2618.36
```

Figure 18-17. SnapStat: Capability Assessment (Individuals)

Using SnapStat 7: Capability Assessment (Grouped Data)

To Access the SnapStat Capability Assessment (Grouped Data), from the menus, choose: SNAPSTATS!!…CAPABILITY ASSESSMENT (GROUPED DATA)…(see Figure 18-18).

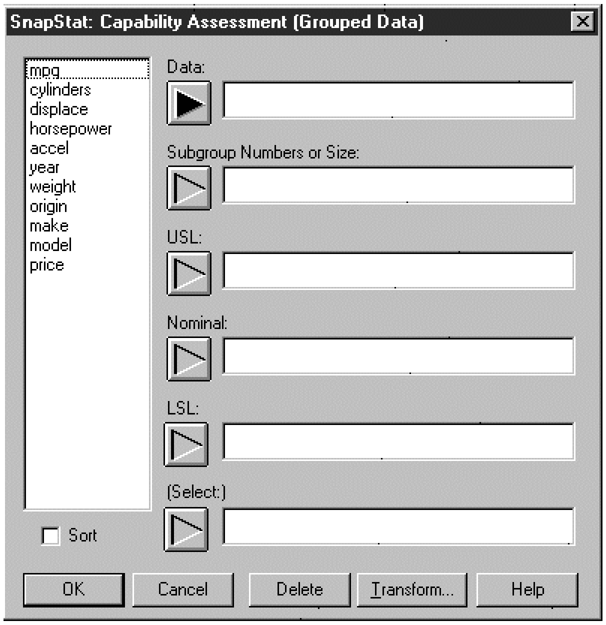
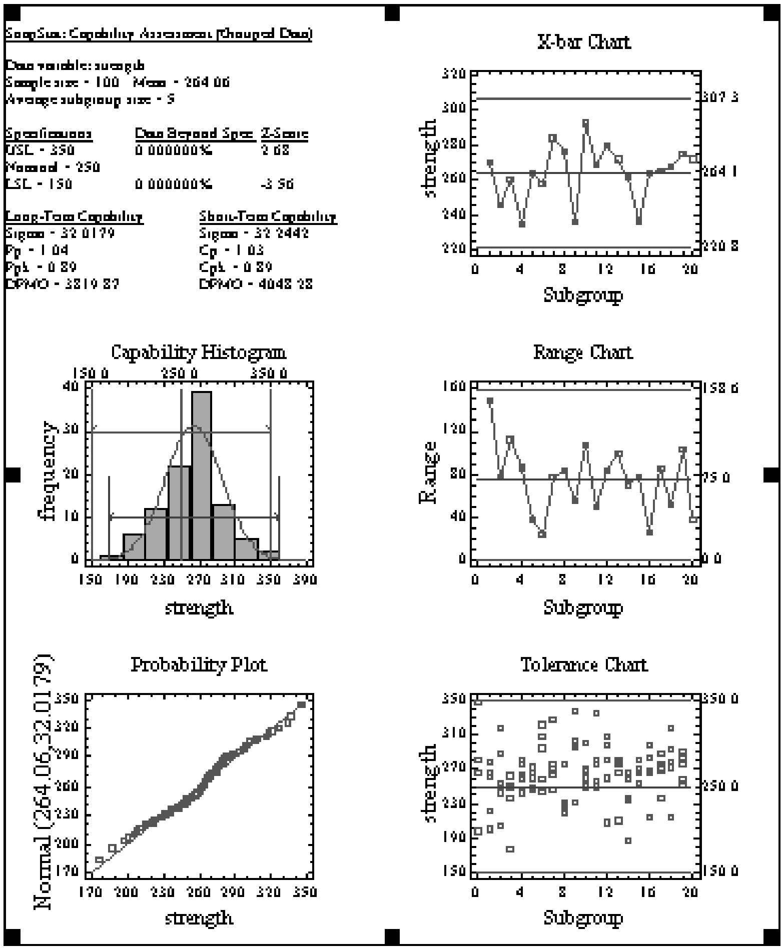


Figure 18-18. SnapStat: Capability Assessment (Grouped Data) Input Dialog Box

The data input dialog box is the same as the *Process Capability (Individuals) SnapStat* except for the “Subgroup Numbers or Size” field, which is the same as that in *X-Bar and R Charts*.

The summary includes the name of the data variable, the values for the sample size, and the parameters for the fitted distribution (see Figure 18-19). The table shows the percent of observations beyond the specification limits, the z-scores, and the percent of estimated observations beyond the specification limits calculated from the fitted distribution. The values for the USL, Nominal, and LSL specifications are also shown. The Observed Beyond Specifications and Estimated Beyond Specifications columns represent the total percentage of the area under the fitted normal curve that fall above and below the specification limits for the distribution.



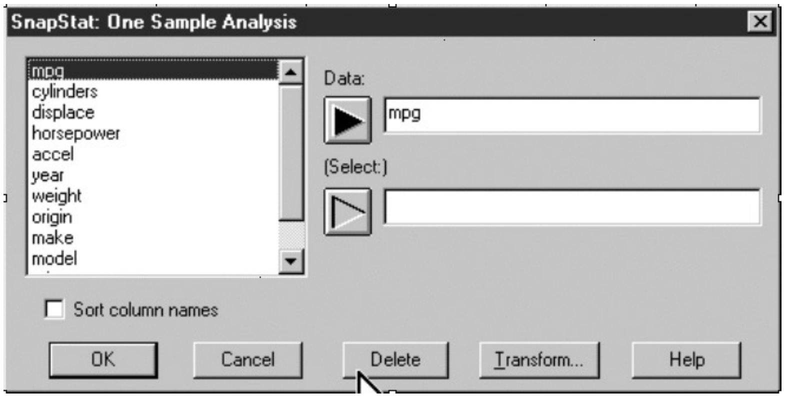
```
SnapStat: Capability Assessment (Grouped Data)

Data variable: strength
Sample size - 100   Mean - 264 06
Average subgroup size - 5

Specifications          Data Beyond Spec  Z-Score
USL - 350               0 000000%         2 68
Nominal - 250
LSL - 150               0 000000%         -3 56

Long-Term Capability         Short-Term Capability
Sigma - 32 0179              Sigma - 32 2442
Pp - 1 04                    Cp - 1 03
Ppk - 0 89                   Cpk - 0 89
DPMO - 3819 87               DPMO - 4048 28
```

Figure 18-19. SnapStat: Capability Assessment (Grouped Data)

Using SnapStat 8: Gage R&R

To Access the SnapStat Gage R&R, from the menus, choose: SNAPSTATS!!…GAGE R&R…(see Figure 18-20).

The Gage R&R SnapStat parallels the Average and Range method described in Gage R&R Analysis. This analysis displays values for the Estimated



SnapStat: One Sample Analysis

Data: mpg

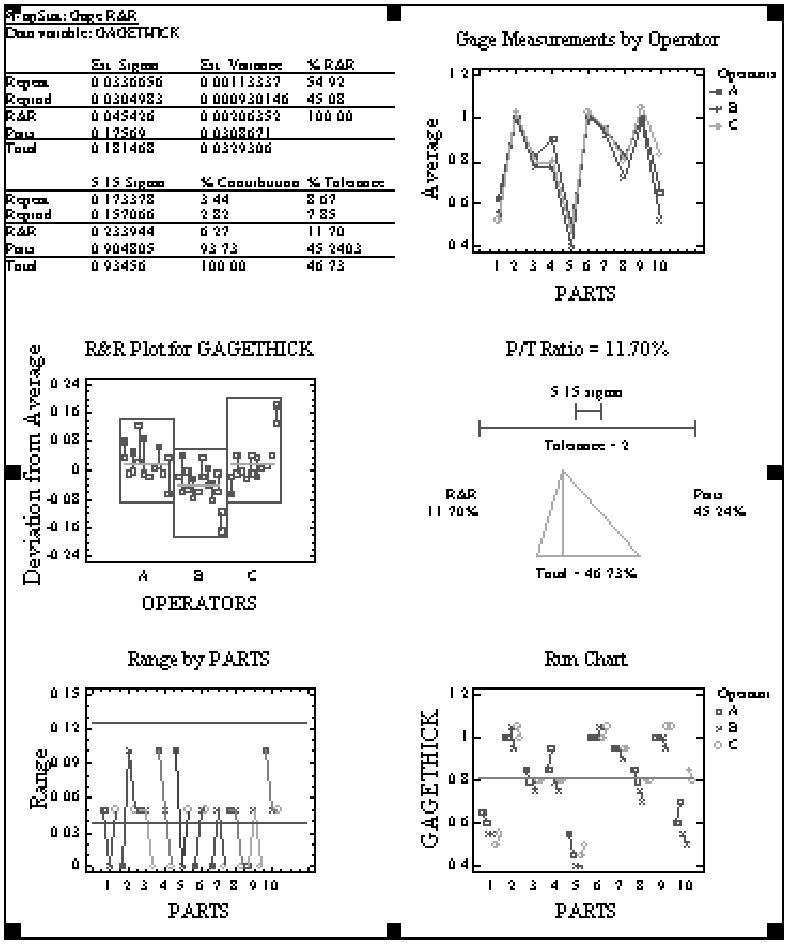
(Select:)

☐ Sort column names

Figure 18-20 SnapStat: Gage R&R Input Dialog Box

Sigma, Estimated Variance, and Percent of Total. Estimated Sigma values are the standard deviation for the measures of repeatability and reproducibility, and a combination of the two. Estimated Variance values include the measure of repeatability and reproducibility, and a combination of the two. Percent of Total is the amount of the total estimated variance that can be attributed to repeatability versus reproducibility (see Figure 18-21).

Options include the Gage R&R Options Dialog Box (see Figure 18-22) and the Gage R&R Tab (see 18-23).



SnapStat: Gage R&R
Data variable: GAGETHICK

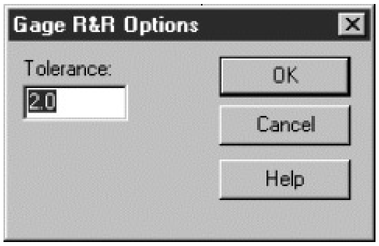
	Est. Sigma	Est. Variance	% R&R
Repeat	0 0336656	0 00113337	54 92
Reprod	0 0304983	0 000930146	45 08
R&R	0 045426	0 00206352	100 00
Part	0 17569	0 0308671	
Total	0 181468	0 0329306	

	5 15 Sigma	% Contribution	% Tolerance
Repeat	0 173378	3 44	8 67
Reprod	0 157066	2 82	7 85
R&R	0 233944	6 27	11 70
Part	0 904805	93 73	45 2403
Total	0 93456	100 00	46 73

Figure 18-21. SnapStat: Gage R&R Input Dialog Box

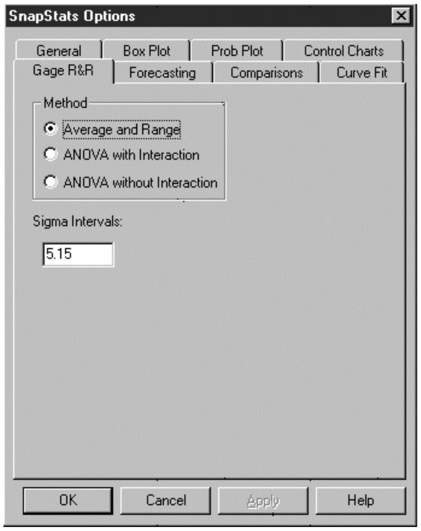
The *Tolerance* field specifies the width of the specification for the characteristic that was measured.

The Gage R&R tab selects the analysis method to be used.



Tolerance: 2.0

Figure 18-22. Gage R&R Options Dialog Box



Method

● Average and Range
○ ANOVA with Interaction
○ ANOVA without Interaction

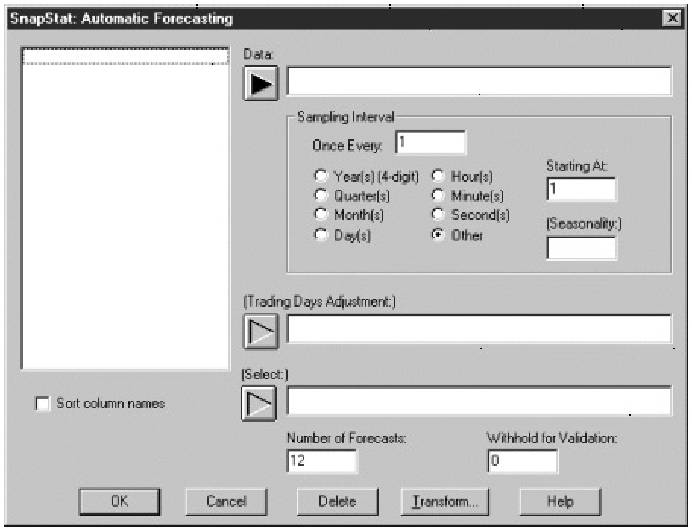
Sigma Intervals: 5.15

Figure 18-23.Gage R&R Tab

Using SnapStat 9: Automatic Forecasting

To Access the SnapStat Automatic Forecasting, from the menus, choose: SNAPSTATS!!…AUTOMATIC FORECASTING…(see Figure 18-24).

The Automatic Forecsting SnapStat is identical to the Automatic Forecsting procedure. The analysis creates a summary that displays the name of the variable and results of the validation, which are used to validate a model’s accuracy. The statistics include the following:



Once Every: 1

● Other

Starting At: 1

Number of Forecasts: 12

Withhold for Validation: 0

Figure 18-24. SnapStat: Automatic Forecasting

- *MSE (Mean Square Error)* — a measure of accuracy computed by squaring the individual error for each item in a dataset, then finding the average or mean value of the sum of those squares. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different forecasting model.

- *MAE (Mean Absolute Error)* — the average of the absolute values of the residuals; this is appropriate for linear and symmetric data. If the result is a small value, you can predict performance more accurately; if the result is a large value, you may want to use a different forecasting model.

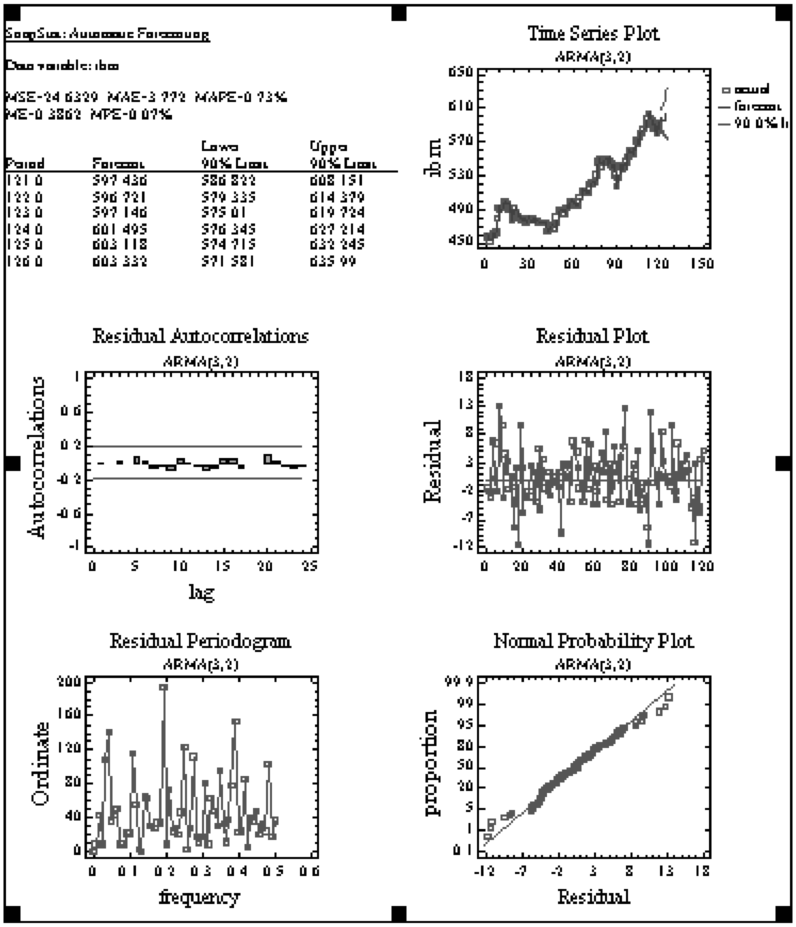
- *MAPE (Mean Absolute Percentage Error)* — the mean or average of the sum of all the percentage errors for a given dataset without regard to sign (that is, the absolute values are summed and the average is computed).

  Unlike the ME, MSE, and MAE, the size of the MAPE is independent of scale.

- *ME (Mean Error)* — the average of the residuals. The closer the ME is to 0, the less biased, or more accurate, the forecast.

- *MPE (Mean Percentage Error)* — the average of the absolute values of the residuals divided by the corresponding estimates. The one-ahead forecast errors are divided by the actual values. Like MAPE, it is independent of scale.

Figure 18-25 summarizes the forecasted values. During periods when actual data are available, the figure displays the predicted values from the fitted model. During time periods beyond the end of the series, the figure shows the prediction limits for the forecasts. The figure also shows the future forecasts and their corresponding upper and lower confidence limits.



SnapStat: Automatic Forecasting

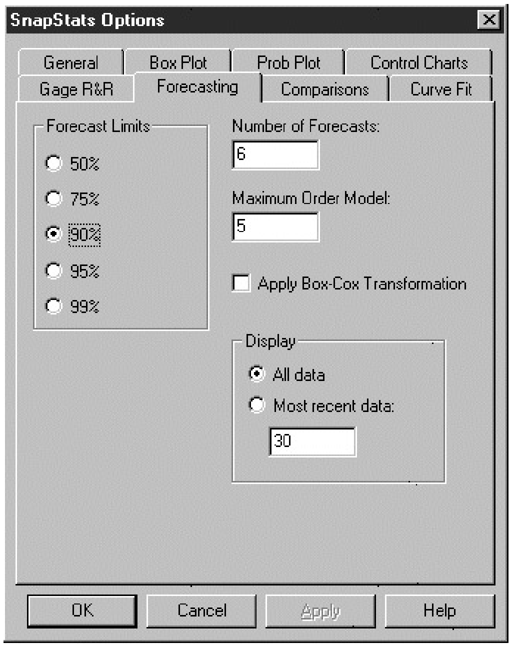
Data variable: ibm

MSE-24 6329 MAE-3 772 MAPE-0 73%
ME-0 3862 MPE-0 07%

Period	Forecast	Lower 90% Limit	Upper 90% Limit
121 0	597 436	586 822	608 151
122 0	596 721	579 335	614 379
123 0	597 146	575 01	619 724
124 0	601 495	576 345	627 214
125 0	603 118	574 715	632 245
126 0	603 332	571 581	635 99

Figure 18-25. SnapStat: Automatic Forecasting

The Option includes the Forecasting Tab (see Figure 18-26).



Forecast Limits

○ 50%

○ 75%

● 90%

○ 95%

○ 99%

Number of Forecasts: 6

Maximum Order Model: 5

□ Apply Box-Cox Transformation

Display

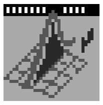
● All data

○ Most recent data: 30

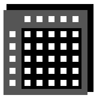
Figure 18-26.Forecasting Tab

A Recognizing Icons and Buttons

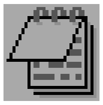
Icons



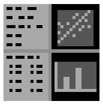
- **STATGRAPHICS Application Icon**
  Clicking this icon opens a STATGRAPHICS session. Under the Windows 95/98 operating systems, this icon also displays on the Taskbar button.



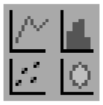
- **Data Icon**
  When you first start STATGRAPHICS *Plus*, the Application window contains the data icon. As you work with the program, this icon identifies an open data file.



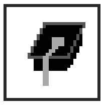
- **Comments Icon**
  When you first start STATGRAPHICS *Plus*, the Application window contains an Untitled comments icon. As you work with the program, this icon identifies a StatFolio Comments window.



- **Analysis Icon**
  Each time you use a statistical analysis, the program displays an Analysis window. When you click the Minimize button on an Analysis window, the program shrinks the window to an analysis icon in a Taskbar button. The title of the icon represents the name of the statistical analysis and may include the variables the analysis uses.

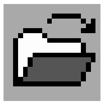


- **StatGallery Icon**
  Clicking this icon enables a function that allows you to place multiple text and graphics images on one page or multiple pages for viewing or printing.

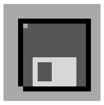


- **StatAdvisor Icon**
  Clicking this icon is one way to open the StatAdvisor, which provides a statistical interpretation of the active analysis.

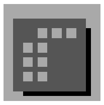
Buttons on the Application Toolbar



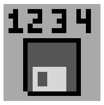
- **Open StatFolio Button**
  Clicking this button displays the Open StatFolio dialog box. It is equivalent to selecting FILE... OPEN... OPEN STATFOLIO... from the Menu bar.



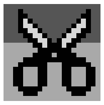
- **Save StatFolio Button**
  Clicking this button displays the Save StatFolio dialog box. It is equivalent to selecting FILE... SAVE... SAVE STATFOLIO... from the Menu bar (if the current StatFolio has a title) or FILE... SAVE AS... SAVE STATFOLIO AS... (if the current StatFolio does not have a title).



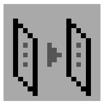
- **Open Data File Button**
  Clicking this button displays the Open Data File dialog box. It is equivalent to selecting FILE... OPEN... OPEN DATA FILE... from the Menu bar.



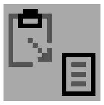
- **Save Data File Button**
  Clicking this button displays the Save Data File dialog box. It is equivalent to selecting FILE... SAVE... SAVE DATA FILE... from the Menu bar (if the current data file has a title) or FILE... SAVE AS... SAVE DATA FILE AS... (if the current data file does not have a title).



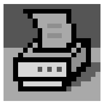
- **Cut Button**
  Clicking this button is equivalent to selecting EDIT... CUT from the Menu bar.



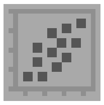
- **Copy Button**
  Clicking this button is equivalent to selecting EDIT... COPY from the Menu bar.



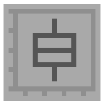
- **Paste Button**
  Clicking this button is equivalent to selecting EDIT... PASTE from the Menu bar.



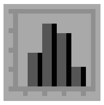
- **Print Button**
  Clicking this button is equivalent to selecting FILE... PRINT from the Menu bar.



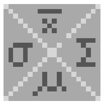
- **X-Y Plot Button**
  Clicking this button is equivalent to selecting PLOT... SCATTERPLOTS... X-Y PLOT from the Menu bar.



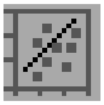
- **Box-and-Whisker Plot Button**
  Clicking this button is equivalent to selecting PLOT... EXPLORATORY PLOTS... BOX-AND-WHISKER PLOT from the Menu bar.



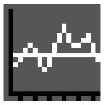
- **Frequency Histogram Button**
  Clicking this button is equivalent to selecting PLOT... EXPLORATORY PLOTS... FREQUENCY HISTOGRAM from the Menu bar.



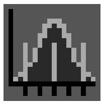
- **Multiple-Variable Analysis Button**
  Clicking this button is equivalent to selecting DESCRIBE... NUMERIC DATA... MULTIPLE-VARIABLE ANALYSIS from the Menu bar.



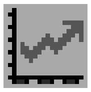
- **Multiple Regression Button**
  Clicking this button is equivalent to selecting RELATE... MULTIPLE REGRESSION from the Menu bar.



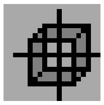
- **X-Bar and R Charts Button**
  Clicking this button is equivalent to selecting SPECIAL... QUALITY CONTROL... VARIABLES CONTROL CHARTS from the Menu bar.



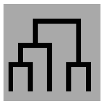
- **Process Capability Analysis Button**
  Clicking this button is equivalent to selecting SPECIAL... QUALITY CONTROL... PROCESS CAPABILITY from the Menu bar.



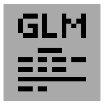
- **Forecast Analysis Button**
  Clicking this button is equivalent to selecting SPECIAL... TIME-SERIES ANALYSIS... FORECASTING from the Menu bar.



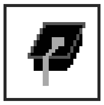
- **Experimental Design Button**
  Clicking this button is equivalent to selecting SPECIAL... EXPERIMENTAL DESIGN... OPEN DESIGN from the Menu bar.



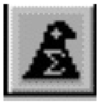
- **Multivariate Methods Button**
  Clicking this button is equivalent to selecting SPECIAL... MULTIVARIATE METHODS... CLUSTER ANALYSIS from the Menu bar.



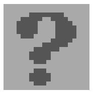
- **General Linear Models Button (GLM)**
  Clicking this button is equivalent to selecting SPECIAL... ADVANCED REGRESSION... GENERAL LINEAR MODELS from the Menu bar.



- **StatAdvisor Button**
  Clicking this button opens the StatAdvisor for the currently active analysis. The StatAdvisor provides a statistical interpretation of the analysis.

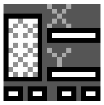


- **StatWizard Button**
  Clicking this button gives you access to the StatWizard which offers a guide to selecting analyses to use with your data. The StatWizard can be set to open whenever STATGRAPHICS is launched by selecting the checkbox on the opening screen.

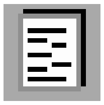


- **Help Button**
  Clicking this button gives you access to Help about the panes in an Analysis window. Online Help allows you to learn about the program without referencing the manual.

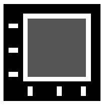
Buttons on the Analysis Toolbar



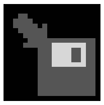
- **Return to Analysis Dialog Box Button**
  Clicking this button redisplays the Analysis or Plot dialog box for the currently active analysis.



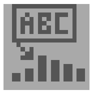
- **Tabular Options Button**
  Clicking this button displays the Tabular Options dialog box for the currently active analysis. This dialog box lists all the available text options.



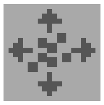
- **Graphical Options Button**
  Clicking this button displays the Graphical Options dialog box for the currently active analysis. This dialog box lists all the available graphics options.



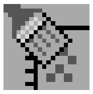
- **Save Results Button**
  Clicking this button displays the Save Results Options dialog box for the currently active analysis. This dialog box lists all the available saved results options and the names of the default variables.



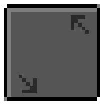
- **Add Text Button**
  (This button is not available until you maximize a graph in an Analysis window.) Clicking this button displays the Text Options dialog box. This dialog box allows you to enter the text you want to add and the direction (horizontal or vertical) in which you want the program to display the text.



- **Jittering Button**
  (This button is not available until you maximize a Scatterplot in an Analysis window.) Clicking this button displays the Jittering dialog box. This dialog box allows you to select the amount of horizontal and/or vertical jittering you want to apply to the points on a graph.

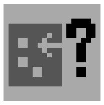


- **Brushing Button**
  (This button is not available until you maximize a Scatterplot in an Analysis window.) Clicking this button displays the Brushing dialog box. This dialog box allows you to select the variable you want to use for brushing the points.

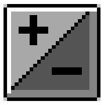


- **Smooth/Rotate Button**
  (This button is not available until you maximize a two-dimensional or three-dimensional graph in an Analysis window.) Clicking this button when a two-dimensional

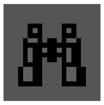
graph is maximized displays a dialog box that allows you to select the type of smoothing to be applied to the graph. Clicking this button when a three-dimensional graph is maximized displays a slidebar/button combination that allows you to rotate the graph horizontally or vertically.



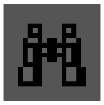
- **Set Point Labels Button**
  (This button is not available until you maximize a Scatterplot in an Analysis window.) Clicking this button displays the Point Identification dialog box. This dialog box allows you to select a variable the program will use to identify the points.



- **Include/Exclude Button**
  (This button is not available until you maximize a Scatterplot in an Analysis window for an analysis where the point values are used in a calculation, i.e., Simple Regression.) Select a point on the Scatterplot, then click on the minus section of the button to exclude a point or on the plus section to include a point that was previously excluded.



- **Locate Labels Button**
  Clicking the Locate Labels button changes the color of the corresponding points on a graph.



- **Locate Row Button**
  Clicking the Locate Row button changes the color of the corresponding point on a graph.

B Using Operators

STATGRAPHICS *Plus* contains mathematical functions, called operators, that let you perform complex calculations, generate and transform data, and otherwise manipulate data. These operators let you perform functions that range from basic math (for example, +, -, SQRT, and LOG) to complex types of calculations.

You use operators in formulas, called expressions. You can use these expressions in the Generate Data dialog box and on many other dialog boxes.

The "Understanding Operators and Expressions" section of this appendix discusses operators, expressions, and expression evaluation. The "Using Operators in Expressions" section explains general information about operators, shows the format for using operators, and provides tutorials and examples.

Understanding Operators and Expressions

Most expressions include a special symbol or word (an operator), that provides information to the program about the type of calculation to perform. For example, the symbols plus (+), minus (-), divide (/), and multiply (*) are all operators, as are the words ROUND, MEDIAN, and SUM. You can type operators in uppercase, lowercase, or a combination of the two.

There are several types of operators; they perform these functions:

- mathematical
- relational
- logical
- generation
- selection
- transformation
- time series
- distribution.

An expression is made up of one or more phrases. A phrase always contains one operator and one (or more) right argument (values or phrases that appear to the right of the operator). It may also contain a left argument (values or phrases that appear to the left of the operator), depending on the operator you use.

For example, the phrase

```
2 + 3
```

has both a right and a left argument. The plus (+) symbol is the operator, 3 is the right argument, and 2 is the left argument. The phrase

```
ABS ( -427)
```

has only a right argument. The word ABS is the operator and -427 is the right argument.

Note: A minus (-) sign directly adjacent to a number is a negative number and precedes all the other operations.

Arguments usually contain numeric values, however, some operators allow character arguments. For example, you can use the Equal To (=) relational operator to return a 1 if a row contains a specific word or character, as in NAME = “JOHN”. Note that you must surround a character argument with double quotes.

When the program evaluates an expression, it follows conventional algebraic hierarchy. The order of precedence is:

Level 1: functions in the form FUNCTION (
Level 2: ^
Level 3: * and /
Level 4: + and -
Level 5: ( )

Consider the expression:

```
2 * 4 + 6
```

In algebraic expression evaluation, which performs multiplication before addition, the expression is equivalent to

```
(2 * 4) + 6
```

therefore the result is 14.

As the above example shows, when an expression contains more than one phrase, a phrase may become a left or right argument for another phrase.

If you want to follow a certain order of operation, you must add parentheses to your expression.

For example, to evaluate the expression

$ax^2 + bx + c$

type it as:

```
(a * x ^ 2) + (b * x) + c
```

Using Operators in Expressions

This section describes the operators available in STATGRAPHICS *Plus,* defines terminology that can help you better understand the material in this appendix, provides a table that lists each operator and describes its purpose, and provides examples for you to try.

This and subsequent sections of this appendix use certain terms to represent the types of data you can use in a formula. In formulas, and in the definitions below, the terms appear in boldface italic type so you can distinguish them easily.

- ***n***
  The term ***n*** represents a single numeric value.
- ***m***
  The term ***m*** represents a single numeric value. Use this term with ***n*** when an example needs two single values.
- ***l***
  The term ***l*** represents a single numeric value. Use this term with ***m*** and ***n*** when an example needs three single values.
- ***x***
  The term ***x*** represents a numeric variable or constant.
- ***y***
  The term ***y*** represents a numeric variable or constant. Use this term with ***x*** when an example needs two numeric variables or constants.
- ***var***
  The term ***var*** represents any variable.
- ***numvar***
  The term ***numvar*** represents any numeric variable.

- ***charvar***
  The term ***charvar*** represents any character variable.
- ***datevar***
  The term ***datevar*** represents any date variable.
- ***logical***
  The term ***logical*** represents a variable that contains only 0’s and 1’s.

Table B-1 lists and describes each operator in the program. To use the table, look in the Operator column for the name of the operator you want to use, read the function description, then read the appropriate section later in this appendix for details on how to use that operator in an expression.

Table B-1. Available Operators

Operator	Function
	Mathematical
++	Adds
–	Subtracts
*	Multiplies
/	Divides
ABS	Takes the absolute value
ACOS	Calculates the inverse cosine of an angle in degrees
ACOSR	Calculates the inverse cosine of an angle in radians
ASIN	Calculates the inverse sine of an angle in degrees
ASINR	Calculates the inverse sine of an angle in radians
ATAN	Calculates the inverse tangent of an angle in degrees
ATANR	Calculates the inverse tangent of an angle in radians
AVG	Calculates the average
COS	Calculates the cosine of an angle in degrees

Table B-1. Continued

Operator	Function
	Mathematical (continued)
COSR	Calculates the cosine of an angle in radians
CV	Calculates the coefficient of variation
EXP	Raises the constant e (2.71828) to a power
EXP10	Raises 10 to a power
FACT	Calculates the factorial
GEOMEAN	Calculates the geometric mean
IQR	Calculates the interquartile range
KURTOSIS	Calculates the coefficient of kurtosis
LOG	Takes the natural logarithm
LOG10	Takes the base ten logarithm
MAX	Finds the largest value
MEDIAN	Calculates the median
MIN	Finds the smallest value
MODE	Finds the most frequent observation
PERCENTILE	Calculates the percentile
Q25	Calculates the lower quartile
Q75	Calculates the upper quartile
RANGE	Calculates the range
RANK	Calculates the rank
REPLACE	Replaces all occurrences of a number with a different number
ROUND	Rounds to the nearest integer
ROUNDTO	Rounds to specified decimal places

Table B-1. Continued

Operation	Function
	Mathematical (continued)
RUNTOT	Calculates a running total
SD	Calculates the standard deviation
SERROR	Calculates the standard error
SIN	Calculates the sine of an angle in degrees
SINR	Calculates the sine of an angle in radians
SIZE	Counts the total number of nonmissing observations
SKEWNESS	Calculates the coefficient of skewness
SKURT	Calculates the standardized kurtosis
SQRT	Calculates the square root
SSKEW	Calculates the standardized skewness
SUM	Calculates the sum of the observations
STANDARDIZE	Creates standardized variables
TAN	Calculates the tangent of an angle in degrees
TANR	Calculates the tangent of an angle in radians
TRUNCATE	Finds the largest integer greater than or equal to **n**
VARIANCE	Calculates the variance
	Relational
=	Equal to
<>	Not equal
>	Greater than
>=	Greater than or equal to
<	Less than

Table B-1. Continued

Operator	Function
	Relational
<=	Less than or equal to
	Logical
&	Both of two conditions must be true (AND)
|	Either of two conditions must be true (OR)
~	Negates a Boolean (NOT)
	Generation
COUNT	Creates a sequential numeric vector
FIRST	Generates a 1 for the first **n** rows in the file and 0 for all the other rows
LAST	Generates a 1 for the last **n** rows in the file and 0 for all the other rows
RANDOM	Generates a 1 for **n** randomly selected rows and 0 for all the other rows
REXPONENTIAL	Generates random numbers from an exponential distribution
RGAMMA	Generates random numbers from a gamma distribution
RINTEGER	Generates random numbers (integer values) from a discrete uniform distribution
RLOGNORMAL	Generates random numbers from a lognormal distribution
RNORMAL	Generates random numbers from a normal distribution
ROW	Generates a 1 for rows **n** to **m** and 0 for all the other rows
RUNIFORM	Generates random numbers from a continuous uniform distribution

Table B-1. Continued

Operator	Function
	Generation
RWEIBULL	Generates random numbers from a Weibull distribution
	Selection
CELL	Selects a row from a specific column
COMPRESS	Selects the rows that meet a condition
DROP	Selects all but the first **n** rows
DROPLAST	Selects all but the last **n** rows
EXCLUDE	Excludes a single row in the selection fields
FIRSTROWS	Selects the first **n** rows and replace the other rows with the missing value codes
LASTROWS	Selects the last **n** rows and replace the other rows with the missing value codes
SELECT	Chooses the rows that meet a condition
TAKE	Selects the first **n** rows
TAKELAST	Select the last **n** rows
	Transformation
DATENUM	Converts a date variable to a number
DIFF	Calculates the differences between the consecutive values
JOIN	Joins the columns end to end
JOIN3	Joins three variables
JOIN4	Joins four variables
JUXTAPOSE	Joins the character variables side by side, expanding the width
LAG	Shifts the values ***n*** positions forward or backward

Table B-1. Continued

Operator	Function
	Transformation
RECODE	Recodes the numeric or character values to integers
REP	Repeats each value **n** times
RESHAPE	Expands or compresses a constant or variable to have a specified number of values
STRIPBLANKS	Removes the double blanks from a character variable
	Time Series
MDIFF	Calculates the multiple backward differences
SDIFF	Calculates the seasonal differences
	Distribution
BETA	Calculates the probability of the cumulative beta distribution at a given point.
CHISQUARE	Calculates the probability of the cumulative chi-square distribution at a given point
INVBETA	Calculates the critical value of the beta distribution given a probability
INVCHISQUARE	Calculates the critical value of the chi-square distribution given a probability
INVNORMAL	Calculates the critical value of the normal distribution given a probability
INVSNEDECOR	Calculates the critical value of the Snedecor's-F distribution given a probability
INVSTUDENT	Calculates the critical value of the Student's-t distribution given a probability
NORMAL	Calculates the probability of the cumulative normal distribution at a given point

Table B-1. Continued

Operator	Function
	Distribution
SNEDECOR	Calculates the probability of the cumulative Snedecor's-F distribution at a given point
STUDENT	Calculates the probability of the cumulative Student's-t distribution at a given point

Using Operators in Examples

In the sections that follow, each description of the operators contains two parts: Type and Result. The Type portion of the example shows what you type; the Result portion shows the result after you type the expression and click the OK button.

As you review the remainder of this section, you may want to work through the examples to become comfortable with using the operators. The sample dataset (file name: **Sampdata**) includes the 10 sample variables used in the examples:

sample1 = 1 2 3 4 5
sample2 = 5 6 7 8 9
sample3 = -5 -4 -3 -2 -1
sample4 = 72 41 93 10 72
sample5 = a c e g i
sample6 = b d f h j
sample7 = Ford Chrysler Chevrolet Mazda Acura
sample8 = Ford   Mustang (a single row with extraneous blanks)
sample9 = 10/1/95 10/2/95 10/3/95
sample10 = 1.111111 2.222222 3.333333 4.444444 5.555555

Try the following two examples. The first uses the ADD operator; the second the JOIN operator. Each uses a different method to obtain the results.

Example 1

1. Select FILE... OPEN... OPEN DATA FILE from the Menu bar **or** click the **Open Data File** button (the third button from the left on the Analysis toolbar), click

the **Sampdata** file name, then click the Open button on the dialog box; the program opens the file.

2. Click the **Data Icon** (if you're using Windows 3.1) **or** the **Sampdata Taskbar** (if you're using Windows 95); the spreadsheet displays.
3. Go to the first **empty column** and highlight it by clicking on the **column name cell**.
4. Click the **right mouse** button, then select **Generate Data**; the Data Generator dialog box displays.
5. Type ***sample1*** + ***sample2*** in the text box. The results (6 8 10 12 14) will display in the column you highlighted.

Example 2

Follow Steps 1 through 4 above.

5. Select **JOIN** from the list of operators in the scroll box to the right of the dialog box; JOIN (?,?) appears in the text box.
6. Replace the question marks with ***sample1***,***sample2***; the results (1 2 3 4 5 6 7 8 9) will display in the column you highlighted.

Using Mathematical Operators

Mathematical operators allow you to perform standard mathematical operations on the constants and values in numeric variables. This section describes the mathematical operators, shows the format you use, and provides examples for you to try.

- **Add (+)**
  Use this operator to add the constants or the values in numeric variables.

  Format: ***x*** + ***y***

  Examples:

  Type: *2 + 4*
  Result: 6

  Type: ***sample1*** + ***sample2***
  Result: 6 8 10 12 14

- **Subtract (–)**
  Use this operator to subtract the constants or the values in numeric variables.

  Format: ***x* - *y***

  Examples:

  Type: *10 – 4*
  Result: 6

  Type: ***sample2* - *sample1***
  Result: 4 4 4 4 4

- **Multiply (*)**
  Use this operator to multiply the constants or the values in numeric variables.

  Format: ***x* * *y***

  Examples:

  Type: *4 * 10*
  Result: 40

  Type: ***sample1* * *sample2***
  Result: 5 12 21 32 45

- **Divide (/)**
  Use this operator to divide the constants or the values in numeric variables.

  Format: ***x*** / ***y***

  Examples:

  Type: *32 / 8*
  Result: 4

  Type: ***sample2*** / ***sample1***
  Result: 5.0 3.0 2.333333 2.0 1.8

- **Raise to a Power (^)**
  Use this operator to raise the constants or the values in numeric variables to a power.

  Format: ***x*** ^ ***y***

Examples:

Type: *9* ^ 2
Result: 81

Type: ***sample2*** ^ ***sample1***
Result: 5 36 343 4096 59049

Take the Absolute Value (ABS)

Use this operator to take the absolute values of the constant or the values in a numeric variable.

Format: ABS (***x***)

Examples:

Type: *ABS (-20)*
Results: 20

Type: *ABS* ***(sample3)***
Result: 5 4 3 2 1

Calculate the Inverse Cosine of an Angle in Degrees (ACOS)

Use this operator to calculate the inverse cosine (in degrees) of the constant or the values in a numeric variable.

Format: ACOS (***x***)

Examples:

Type: *ACOS (.7)*
Results: 45.572996

Type: *ACOS* ***(sample2*** ** .01)*
Result: 87.134016 86.560187 85.986012 85.411434 84.836393

Calculate the Inverse Cosine of an Angle in Radians (ACOSR)

Use this operator to calculate the inverse cosine (in radians) of the constant or the values in a numeric variable.

Format: ACOSR (***x***)

Examples:

Type: *ACOSR (.7)*
Results: .795399

Type: *ACOSR* ***(sample2*** ** .01)*
Result: 1.520775 1.510760 1.500739 1.490710 1.480674

- **Calculate the Inverse Sine of an Angle in Degrees (ASIN)**
  Use this operator to calculate the inverse sine (in degrees) of the constant or the values in a numeric variable.

  Format: ASIN (***x***)

  Examples:

  Type: *ASIN (.7)*
  Results: 44.427004

  Type: *ASIN (**sample2** * .01)*
  Result: 2.865983 3.439812 4.013987 4.588565 5.163607

- **Calculate the Inverse Sine of an Angle in Radians (ASINR)**
  Use this operator to calculate the inverse sine (in radians) of the constant or the values in a numeric variable.

  Format: ASINR ***(x)***

  Examples:

  Type: *ASINR (.7)*
  Results: .775397

  Type: *ASINR (**sample2** * .01)*
  Result: 0.050020 0.060036 0.070057 0.080085 0.090121

- **Calculate the Inverse Tangent of an Angle in Degrees (ATAN)**
  Use this operator to calculate the inverse tangent (in degrees) of the constant or the values in a numeric variable.

  Format: ATAN ***(x)***

  Examples:

  Type: *ATAN (.7)*
  Results: 34.992020

  Type: *ATAN (**sample2** * .01)*
  Result: 2.862405 3.433630 4.004172 4.573921 5.142764

- **Calculate the Inverse Tangent of an Angle in Radians (ATANR)**
  Use this operator to calculate the inverse tangent (in radians) of the constant or the values in a numeric variable.

  Format: ATANR (***x***)

Examples:

Type: *ATANR (.7)*
Results: 0.610725

Type: *ATANR (**sample2** * .01)*
Result: 0.049958 0.059928 0.069886 0.079829 0.089758

Calculate the Average (AVG)

Use this operator to calculate the average of the values in a numeric variable.

Format: AVG (***numvar***)

Example:

Type: *AVG (**sample1**)*
Result: 3

Calculate the Cosine of an Angle in Degrees (COS)

Use this operator to calculate the cosine (in degrees) of the constant or the values in a numeric variable.

Format: COS (***x***)

Examples:

Type: *COS (40.0)*
Result: 0.766044

Type: *COS (**sample1**)*
Result: 0.999847 0.999390 0.998629 0.997564 0.996194

Calculate the Cosine of an Angle in Radians (COSR)

Use this operator to calculate the cosine (in radians) of a constant or the values in a numeric variable.

Format: COSR (***x***)

Examples:

Type: *COSR (40.0)*
Result: 0.6669380

Type: *COSR (**sample1**)*
Result: 0.540302 -0.416146 -0.989992 -0.653643 0.283662

- **Calculate the Coefficient of Variation (CV)**
  Use this operator to calculate the coefficient of variation for the values in a numeric variable.

  Format: CV (***numvar***)

  Example:

  Type: *CV (**sample1**)*
  Result: 52.704627

- **Raise e to a Power (EXP)**
  Use this operator to raise the constant e (2.71828) to the power specified by a constant or the values in a numeric variable.

  Format: EXP (***x***)

  Examples:

  Type: *EXP (2)*
  Result: 7.38905

  Type: *EXP (**sample1**)*
  Result: 2.718281 7.389056 20.085536 54.598150 148.413159

- **Raise 10 to a Power (EXP10)**
  Use this operator to raise 10 to the power specified by a constant or the values in a numeric variable.

  Format: EXP10 (***x***)

  Examples:

  Type: *EXP10 (2)*
  Result: 100

  Type: *EXP10 (**sample1**)*
  Result: 10 100 1000 10000 100000

- **Calculate the Factorial (FACT)**
  Use this operator to calculate the factorial of a constant or the values in a numeric variable.

  Format: FACT (***x***)

  Examples:

  Type: *FACT (3)*
  Result: 6

Type: *FACT (**sample1**)*
Result: 1 2 6 24 120

- **Calculate the Geometric Mean (GEOMEAN)**

Use this operator to calculate the geometric mean of a numeric variable.

Format: GEOMEAN (***numvar***)

Example:

Type: *GEOMEAN (**sample2**)*
Result: 6.853467

- **Calculate the Interquartile Range (IQR)**

Use this operator to calculate the interquartile range of a numeric variable.

Format: IQR (***numvar***)

Example:

Type: *IQR (**sample2**)*
Result: 2

- **Calculate the Coefficient of Kurtosis (KURTOSIS)**

Use this operator to calculate the coefficient of kurtosis of a numeric variable.

Format: KURTOSIS (***numvar***)

Example:

Type: *KURTOSIS (**sample2**)*
Result: -1.2

- **Take the Natural Logarithm (LOG)**

Use this operator to take the natural logarithm of a constant or the values in a numeric variable.

Format: LOG (***x***)

Examples:

Type: *LOG (100)*
Result: 4.60517

Type: *LOG (**sample1**)*
Result: 0 0.693147 1.098612 1.386294 1.609437

- **Take the Base Ten Logarithm (LOG10)**

  Use this operator to take the base-ten logarithm of a constant or the values in a numeric variable.

  Format: LOG10 (***x***)

  Examples:

  Type: *LOG10 (100)*
  Result: 2

  Type: *LOG10 (**sample1**)*
  Result: 0 0.301029 0.477121 0.602059 0.698970

- **Find the Largest Value (MAX)**

  Use this operator to find the largest value in a numeric variable.

  Format: MAX (***numvar***)

  Example:

  Type: *MAX (**sample2**)*
  Result: 9

- **Calculate the Median (MEDIAN)**

  Use this operator to calculate the median of the values in a numeric variable.

  Format: MEDIAN (***numvar***)

  Example:

  Type: *MEDIAN (**sample2**)*
  Result: 7

- **Find the Smallest Value (MIN)**

  Use this operator to find the smallest value in a numeric variable.

  Format: MIN (***numvar***)

  Example:

  Type: *MIN (**sample1**)*
  Result: 1

- **Find the Most Frequent Observation (MODE)**

  Use this operator to find the most frequent observation in a numeric variable. If there is no unique mode, the program returns a missing value.

Format: MODE (***numvar***)

Example:

Type: *MODE (**sample4**)*
Result: 72

■ Calculate a Percentile (PERCENTILE)

Use this operator to calculate the **n**th percentile of a numeric variable.

Format: PERCENTILE (***numvar***,***n***)

Example:

Type: *PERCENTILE (**sample1**,25)*
Result: 2

■ Calculate the Lower Quartile (Q25)

Use this operator to calculate the lower quartile of a numeric variable; that is, the value at the 25th percentile.

Format: Q25 (***numvar***)

Example:

Type: *Q25 (**sample2**)*
Result: 6

■ Calculate the Upper Quartile (Q75)

Use this operator to calculate the upper quartile of a numeric variable; that is, the value at the 75th percentile.

Format: Q75 (***numvar***)

Example:

Type: *Q75 (**sample2**)*
Result: 8

■ Calculate the Range (RANGE)

Use this operator to calculate the range of values in a numeric variable; that is, the difference between the largest and smallest values.

Format: RANGE (***numvar***)

Example:

Type: *RANGE (**sample2**)*
Result: 4

■ Calculate the Rank of Each Row (RANK)

Use this operator to calculate the rank of the values in a numeric variable; that is, rank the values from lowest to highest, and where there are multiple occurrences of the same value, return the median rank for each of those values.

Format: RANK (***varname***)

Example:

Type: *RANK* ***(sample4)***
Result: 3.5 2 5 1 3.5

■ Replace All Occurrences of a Number with a Different Number (REPLACE)

Use this operator to change all occurrences of a number with a different number.

Format: REPLACE (***varname,old,new)***

Example:

Type: *REPLACE* ***(sample1****,5,6)*

Result: 1 2 3 4 6

■ Round to the Nearest Integer (ROUND)

Use this operator to round a constant or the values in a numeric variable to the nearest integer.

Format: ROUND (***x***)

Examples:

Type: *ROUND 1234.5678*
Result: 1235

Type: *ROUND* ***(sample1*** *+ .6)*
Result: 2 3 4 5 6

■ Round to Specified Number of Decimal Places (ROUNDTO)

Use this operator to round the values in a numeric variable to the number of decimal places you specify.

Format: ROUNDTO ***(varname,decimal)***

Example:

Type: *ROUNDTO* ***(sample10,4)***
Result: 1.1111 2.2222 3.3333 4.4444 5.5556

- **Calculate a Running Total (RUNTOT)**
  Use this operator to calculate a running total for a numeric variable.

  Format: RUNTOT(***numvar***)

  Example:

  Type: *RUNTOT* ***(sample1)***
  Result: 1 3 6 10 15

- **Calculate the Standard Deviation (SD)**
  Use this operator to calculate the standard deviation of a numeric variable.

  Format: SD (***numvar***)

  Example:

  Type: *SD* ***(sample2)***
  Result: 1.581138

- **Calculate the Standard Error (SERROR)**
  Use this operator to calculate the standard error of a numeric variable.

  Format: SERROR (***numvar***)

  Example:

  Type: *SERROR* ***(sample1)***
  Result: 0.707106

- **Calculate the Sine of an Angle in Degrees (SIN)**
  Use this operator to calculate the sine (in degrees) of a constant or the values in a numeric variable.

  Format: SIN (***x***)

  Examples:

  Type: *SIN (30.0)*
  Result: .5

  Type: *SIN* ***(sample1)***
  Result: 0.017452 0.034899 0.052335 0.069756 0.087155

- **Calculate the Sine of an Angle in Radians (SINR)**
  Use this operator to calculate the sine (in radians) of a constant or the values in a numeric variable.

  Format: SINR (***x***)

  Examples:

  Type: *SINR (30.0)*
  Result: -0.988032

  Type: *SINR **(sample1)***
  Result: 0.841470 0.909297 0.141120 -0.756802 -0.958924

- **Count the Number of Nonmissing Observations (SIZE)**
  Use this operator to calculate the size of a numeric variable; that is, the total number of nonmissing observations.

  Format: SIZE (***numvar***)

  Example:

  Type: *SIZE **(sample1)***
  Result: 5

- **Calculate the Coefficient of Skewness (SKEWNESS)**
  Use this operator to calculate the coefficient of skewness. If the number of observations is < = 2, the program returns a missing value.

  Format: SKEWNESS (***numvar***)

  Example:

  Type: *SKEWNESS **(sample4)***
  Result: -.757888

- **Calculate the Standardized Kurtosis (SKURT)**
  Use this operator to calculate the standardized kurtosis. If the number of observations is < = 7, the program returns a missing value.

  Format: SKURT (***numvar***)

  Example:

  Type: *SKURT **(sample1)***
  Result: -0.547723

- **Calculate the Square Root (SQRT)**
  Use this operator to calculate the square root of a constant or the values in a numeric variable.

  Format: SQRT (***x***)

  Examples:

  Type: *SQRT 10000*
  Result: 100

  Type: *SQRT* ***(sample1)***
  Result: 1 1.414213 1.732050 2 2.236067

- **Calculate the Standardized Skewness (SSKEW)**
  Use this operator to calculate the standardized skewness. If the number of observations is < = 7, the program returns a missing value.

  Format: SSKEW (***numvar***)

  Example:

  Type: *SSKEW* ***(sample4)***
  Result: -.691854

- **Calculate Standardized Variables (STANDARDIZE)**
  Use this operator to subtract the average from each value and then divide by the standard deviation.

  Format: STANDARDIZE (***numvar***)

  Example:

  Type: *STANDARDIZE*
  (sample1)
  Result: -1.26491 -0.632456 0.632456 1.26491

- **Calculate the Sum of the Observations (SUM)**
  Use this operator to total the values in a numeric variable.

  Format: SUM (***numvar***)

  Example:

  Type: *SUM* ***(sample1)***
  Result: 15

- **Calculate the Tangent of an Angle in Degrees (TAN)**
  Use this operator to calculate the tangent (in degrees) of a constant or the values in a numeric variable.

  Format: TAN (***x***)

  Examples:

  Type: *TAN (30.0)*
  Result: 0.577350

  Type: *TAN (**sample1**)*
  Result: 0.017455 0.034920 0.052407 0.069926 0.087488

- **Calculate the Tangent of an Angle in Radians (TANR)**
  Use this operator to calculate the tangent (in radians) of a constant or the values in a numeric variable.

  Format: TANR (***x***)

  Examples:

  Type: *TANR (30.0)*
  Result: -6.405331

  Type: *TANR (**sample1**)*
  Result: 1.557407 -2.185039 -0.142546 1.157821 -3.380515

- **Find the Largest Integer (TRUNCATE)**
  Use this operator to find the largest integer less than or equal to a constant or the values in a numeric variable.

  Format: TRUNCATE (***x***)

  Examples:

  Type: *TRUNCATE (4.98315)*
  Result: 4

  Type: *TRUNCATE (**sample1** + .06)*
  Result: 1 2 3 4 5

- **Calculate the Variance (VARIANCE)**
  Use this operator to calculate the variance of a numeric variable.

  Format: VARIANCE (***numvar***)

Example:

Type: *VARIANCE* ***(sample2)***
Result: 2.5

Using Relational Operators

Relational operators determine if an expression is true or false. The program matches each value in the left argument with a value in the right argument. If the calculation for that pair of values is true, the program returns a 1; if it is false, a 0. If one of the values is missing, the program returns the missing value.

This section describes the relational operators, shows the format you use, and provides examples for you to try. You typically use these operators with the SELECT and COMPRESS operators to select values that meet certain conditions. See SELECT and COMPRESS later in this appendix for examples.

- **Equal To (=)**
  Use this operator to determine if the constants or the values in numeric variables are equal.

  Format: ***x*** = ***y***

  Examples:

  Type: ***sample1*** = *1*
  Result: 1 0 0 0 0

  Type: ***sample1*** = ***sample2***
  Result: 0 0 0 0 0

- **Not Equal (<>)**
  Use this operator to determine if the constants or the values in numeric variables are not equal.

  Format: ***x*** <> ***y***

  Examples:

  Type: ***sample1*** <> *1*
  Result: 0 1 1 1 1

  Type: ***sample1*** <> ***sample2***
  Result: 1 1 1 1 1

- **Greater Than (>)**

  Use this operator to determine if the values in the left argument are greater than the values in the right argument.

  Format: ***x* > *y***

  Examples:

  Type: ***sample1*** > *3*
  Result: 0 0 0 1 1

  Type: ***sample2*** > ***sample1***
  Result: 1 1 1 1 1

- **Greater Than or Equal To (>=)**

  Use this operator to determine if the values in the left argument are greater than or equal to the values in the right argument.

  Format: ***x* >= *y***

  Examples:

  Type: ***sample1*** >= *3*
  Result: 0 0 1 1 1

  Type: ***sample2*** >= ***sample1***
  Result: 1 1 1 1 1

- **Less Than (<)**

  Use this operator to determine if the values in the left argument are less than the values in the right argument.

  Format: ***x* < *y***

  Examples:

  Type: ***sample1*** < *3*
  Result: 1 1 0 0 0

  Type: ***sample1*** < ***sample2***
  Result: 1 1 1 1 1

- **Less Than or Equal To (<=)**

  Use this operator to determine if the values in the left argument are less than or equal to the values in the right argument.

  Format: ***x* <= *y***

Examples:

Type:	***sample1*** *<= 3*
Result:	1 1 1 0 0
Type:	***sample1*** <= ***sample2***
Result:	1 1 1 1 1

Using Logical Operators

Logical operators select values that meet both (&) or either (|) of two conditions, or that negate (~) a variable that contains only 0s and 1s.

You use logical operators in conjunction with relational operators. The relational operators test for a condition and return either a 1 (true) or a 0 (false). For example, when you type the expression

40 > 45

the program returns:

0

You use the logical operators both (&) and either (|) to make two relational tests on two variables. The logical operators compare the 0s and 1s that the relational tests return to determine if a value meets both conditions (&) or either condition (|). You use the logical operator (~) to negate a variable that contains only 0s and 1s; that is, to change 0s to 1s and vice versa.

This section describes the logical operators and provides examples for you to try. Like the relational operators, you typically use logical operators with the SELECT and COMPRESS operators to select values that meet certain conditions. See the information on SELECT and COMPRESS for examples.

- **Both of Two Conditions Must Be True (&)**
  Use this operator to determine if the values meet both of two conditions. The program calculates this result by testing for each condition separately, then comparing the results. If the value meets both conditions, the program returns a 1 (true). If the value meets only one or neither condition, the program returns a 0 (false). If one of the values is missing, the program returns the missing value.

  Format: (***logical***) & (***logical***)

Examples:

Type: *(40 > 10) & (10 = 50)*
Result: 0. The result of the first condition (40 > 10) is 1. The result of the second condition (10 = 50) is 0. The value meets only one condition, so the program returns a 0 (false).

Type: ***(sample1** = 3) & (**sample2** > 5)*
Result: The program returns a 1 for cases where a value in the variable **sample1** is equal to 3 and the corresponding value in **sample2** is greater than 5. For cases that meet only one or neither of these conditions, the program returns a 0.

- **Negate a Boolean Constant or Variable (~)**

Use this operator to negate a Boolean constant or the values in a variable; that is, return a 1 for each item of the argument that is 0, and return a 0 for each item of the argument that is 1.

Format: ~ ***x***

Examples:

Type: *~ 0*
Result: 1

Type: *~ (**sample1** = 4)*
Result: The program evaluates the statement in parentheses, returning a 1 (true) for every value in the **sample1** variable that meets the condition and a 0 (false) for every value that does not meet the condition. Then the program negates the Boolean result, in effect returning a 1 for every value in the **sample1** variable that is not equal to 4.

- **Either of Two Conditions Must be True (|)**

Use this operator to determine if the values meet either of two conditions. The program calculates this result by testing for each condition separately, then comparing the results. If the value meets one or both conditions, the program returns a 1 (true). If the value meets neither condition, the program returns a 0 (false). If one of the values is the missing value code 32768, the program returns the missing value code.

Format: **x | y**

Examples:

Type: *(40 > 10) | (40 = 20)*
Result: The result of the first condition (40 > 10) is 1. The result of the second condition (40 = 20) is 0. The value meets one condition, so the program returns a 1 (true).

Type: ***(sample2** = 7) | (**sample1** > 2)*
Result: The program returns a 1 for cases where a value in the variable **sample2** is equal to 7 or the corresponding value in **sample1** is greater than 2. For cases that meet neither condition, the program returns a 0.

Using Generation Operators

Generation operators create values quickly and easily. This section describes the data-generation operators, shows the format you use, and provides examples for you to try.

- **Create a Sequential Numeric Vector (COUNT)**
  Use this operator to create a sequential vector of integers, from ***start*** through ***end*** by ***step***.

  Format: COUNT (***start***,***end***,***step***)

  Examples:

  Type: *COUNT (10,20,2)*
  Result: 10 12 14 16 18 20

  Type: *10 + (COUNT (1,5,1))*
  Result: 11 12 13 14 15

- **Generate a 1 for the First n Rows (FIRST)**
  Use this operator to generate a 1 for the first **n** rows in the file, and 0 for all the other rows. (Note that this operator is designed especially for use in the Select text box on Analysis dialog boxes.)

  Format: FIRST (***n***)

Example:

Type:	*FIRST (3)*
Result:	In a file with four rows, the program generates three 1s followed by one 0.

- **Generate a 1 for the Last n Rows (LAST)**

Use this operator to generate a 1 for the last **n** rows in the file, and 0 for all the other rows. (Note that this operator is designed especially for use in the Select text box on Analysis dialog boxes.)

Format: LAST (***n***)

Example:

Type:	*LAST (3)*
Result:	In a file with four rows, the program generates one 0 followed by three 1s.

- **Generate a 1 for n Randomly Selected Rows (RANDOM)**

Use this operator to generate a 1 for **n** randomly selected rows in the file, and 0 for all the other rows. (Note that this operator is designed especially for use in the Select text box on Analysis dialog boxes.)

Format: RANDOM (***n***)

Example:

Type:	*RANDOM (3)*
Result:	In a file with four rows, the program generates three 1s in random rows and 0s in all the other rows.

- **Generate Random Numbers from an Exponential Distribution (REXPONENTIAL)**

Use this operator to generate **n** random numbers from an exponential distribution with a specified **mean**. (Note that your results may not match those shown below.)

Format: REXPONENTIAL (***n***,***mean***)

Example:

Type:	*REXPONENTIAL (5,20)*
Result:	35.445269 42.880350 48.049617 18.078479 2.923928

- **Generate Random Numbers from a Gamma Distribution (RGAMMA)**
  Use this operator to generate **n** random numbers from a gamma distribution with a specified **alpha** and **beta**. (Note that your results may not match those shown below.)

  Format: RGAMMA (***n***,***alpha***,***beta***)

  Example:

  Type: *RGAMMA (5,1,3)*
  Result: 2.231758 2.729751 3.263105 1.364058 2.697922

- **Generate Random Numbers (Integer Values) from a Discrete Uniform Distribution (RINTEGER)**
  Use this operator to generate **n** random numbers from a discrete uniform distribution with specified **lower** and **upper** values.

  Format: RINTEGER (***n***,***lower***,***upper***)

  Example:

  Type: *RINTEGER (5,10,15)*
  Result: 15 15 14 13 11

- **Generate Random Numbers from a Lognormal Distribution (RLOGNORMAL)**
  Use this operator to generate **n** random numbers from a lognormal distribution with a specified **mean** and **standard deviation.**

  Format: RLOGNORMAL (***n***,***mean,standard deviation***)

  Example:

  Type: *RLOGNORMAL (5,7,9)*
  Result: 51372.3221 0.002422 2156.244872 8.874807 0.000172

- **Generate Random Numbers from a Normal Distribution (RNORMAL)**
  Use this operator to generate **n** random numbers from a normal distribution with a specified **mean** and **standard deviation**.

  Format: RNORMAL (***n***,***mean, standard deviation***)

  Example:

  Type: *RNORMAL (5,3.2,4.9)*
  Result: -2.957235 6.426126 -0.497581 9.598366 -3.141531

■ **Generate a 1 for Rows n to m (ROWS)**
Use this operator to generate a 1 for rows **n** to **m** in the file, and 0 for all the other rows. (Note that this operator is designed especially for use in the Select text box on Analysis dialog boxes.)

Format: ROWS (***n***,***m***)

Example:

Type: *ROWS (2,4)*
Result: In a file with 10 rows, the program generates one 0, followed by three 1s and six 0s

■ **Generate Random Numbers from a Continuous Uniform Distribution (RUNIFORM)**
Use this operator to generate **n** random numbers from a continuous uniform distribution from a specified **lower** limit to a specified **upper** limit.

Format: RUNIFORM (***n***,***lower***,***upper***)

Example:

Type: *RUNIFORM (5,20,100)*
Result: 39.514873 59.491310 35.746625 79.181132 51.351437

■ **Generate Random Numbers from a Weibull Distribution (RWEIBULL)**
Use this operator to generate **n** random numbers from a Weibull distribution with a specified **alpha** and **beta**.

Format: RWEIBULL (***n***,***alpha***,***beta***)

Example:

Type: *RWEIBULL (5,.2,.4)*
Result: 0.000016 2.331905 0.000462 2.183750 81.772798

Using Selection Operators

Selection operators select certain values from variables. This section describes the selection operators, shows the format you use, and provides examples for you to try.

- **Select Specified Row (CELL)**
  Use this operator to select a specified row from a numeric or a character variable.

  Format: CELL (***var***,***n***)

  Example:

  Type: *CELL (**sample2**,3)*
  Result: 7

- **Select Rows that Meet a Condition (COMPRESS)**
  Use this operator to select only the values that meet a certain condition. The COMPRESS operator is similar to the SELECT operator except that it does not replace the values that do not meet the condition with missing values; it discards such values entirely from the selection. You can use this operator with either a character or a numeric variable.

  Format: COMPRESS (***var***,***logical***)

  Examples:

  Type: *COMPRESS (**sample2**,**sample1** > 2)*
  Result: 7 8 9

  Type: *COMPRESS* (***sample2,sample7*** > *"Acura"*)
  Result: 5

- **Select All but First n Rows (DROP)**
  Use this operator to drop the values from the beginning of a numeric or a character variable.

  Format: DROP (***var***,***n***)

  Example:

  Type: *DROP (**sample1**,2)*
  Result: 3 4 5

- **Select All but Last n Rows (DROPLAST)**
  Use this operator to drop the values from the end of a numeric or a character variable.

  Format: DROPLAST (***var***,***n***)

  Example:

  Type: *DROPLAST (**sample2**,2)*
  Result: 5 6 7

■ **Exclude single rows in selection fields (EXCLUDE)**
Use this operator to exclude a single row in the selection fields on the data input boxes.

Format: EXCLUDE (***n***)

Example:

Type: *EXCLUDE (**26**)*
Result: The program does not use the value in row 26.

■ **Select First n Rows, Using Missing Values for Other Rows (FIRSTROWS)**
Use this operator to select the first **n** rows in a numeric or a character variable and replace the other rows with missing values.

Format: FIRSTROWS (***var***,***n***)

Example:

Type: *FIRSTROWS (**sample2**,4)*
Result: 5 6 7 8 (missing)

■ **Select Last n Rows, Using Missing Values for Other Rows (LASTROWS)**
Use this operator to select the last **n** rows in a numeric or a character variable and replace the other rows with missing values.

Format: LASTROWS (**var**,***n***)

Example:

Type: *LASTROWS (**sample2**,4)*
Result: (missing) 6 7 8 9

■ **Select Rows that Meet a Condition (SELECT)**
Use this operator to select all the values that meet one or more conditions, and replace all the other values with missing values if the values are character.

Format: SELECT ***(var,logical)***

Examples:

Type: *SELECT (**sample1**,**sample2** = 6*)
Result: The program selects the values from **sample1** for cases where the corresponding values in **sample2** are greater than or equal to 6. For cases where the

corresponding values in **sample2** are not greater than or equal to 6, the program returns a missing value. The variables you use in this type of expression must contain the same number of values.

Type:	*SELECT (**sample1**,(**sample2*** < 7) | (***sample2*** > 8)
Result:	The program selects values from **sample1** for cases where the corresponding values in **sample2** are less than 7 or greater than 8. For cases that do not meet the conditions, the program returns a missing value.

- **Select First n Rows (TAKE)**
  Use this operator to select values from the beginning of a numeric or a character variable.

Format: TAKE (***var***,***n***)

Example:

Type: *TAKE (**sample2**,3)*
Result: 5 6 7

- **Select Last n Rows (TAKELAST)**
  Use this operator to select the values from the end of a numeric or a character variable.

Format: TAKELAST (***var***,***n***)

Example:

Type: *TAKELAST (**sample2**,3)*
Result: 7 8 9

Transformation Operators

Transformation operators manipulate values and variables. This section describes the transformation operators, shows the format you use, and provides examples for you to try.

- **Convert a Date Variable to a Number (DATENUM)**
  Use this operator to make it easier to use dates in a calculation. It converts the data to a STATGRAPHICS *Plus* date code. The right argument must be a date variable. The DATENUM operator converts the values to counting numbers that are equivalent to changing the date type in the editor to integer.

Format: DATENUM (***datevar)***

Examples:

Type: *DATENUM* ***(sample9)***
Result: 16710 16711 16712

- **Calculate Differences between Consecutive Values (DIFF)**

  Use this operator to calculate the differences between the consecutive values in a numeric variable. The result of DIFF has one less value than the original variable.

  Format: DIFF (***numvar***)

  Examples:

  Type: *DIFF* ***(sample1)***
  Result: (*missing*) 1 1 1 1

  Type: *DIFF* ***(sample1*sample2)***
  Result: (*missing*) 7 9 11 13

- **Join Columns End to End (JOIN)**

  Use this operator to join two numeric or two character variables end to end to form one continuous vector of values.

  Format: JOIN (***var***,***var***)

  Example:

  Type: *JOIN* ***(sample1****,* ***sample2)***
  Result: 1 2 3 4 5 5 6 7 8 9

- **Join Three Variables (JOIN3)**

  Use this operator to join three numeric or three character variables end to end.

  Format: JOIN3 (***var,var,var***)

  Example:

  Type: *JOIN3* ***(sample1,sample2,sample3)***
  Result: 1 2 3 4 5 5 6 7 8 9 -5 -4 -3 -2 -1

- **Join Four Variables (JOIN4)**

  Use this operator to join four numeric or four character variables end to end.

  Format: JOIN4 (***var,var,var,var***)

Example:

Type: *JOIN4 (**sample1,sample3,sample1,sample3**)*
Result: 1 2 3 4 5 -5 -4 -3 -2 -1 1 2 3 4 5 -5 -4 -3 -2 -1

- **Join Character Variables Side by Side (JUXTAPOSE)**
  Use this operator to join character variables side by side, expanding the width.

  Format: JUXTAPOSE (***charvar***,***charvar***)

  Example:

  Type: *JUXTAPOSE (**sample5**,**sample6**)*
  Result: ab cd ef gh ij

- **Shift Values n Positions Forward or Backward (LAG)**
  Use this operator to shift the values in a numeric variable a specified number of positions forward or backward. Use a positive number to shift the values forward, a negative number to shift the values backward. The program removes the values you shift out of the variable at the appropriate end and adds a missing value to fill the empty observations at the other end.

  Format: LAG (***numvar***,***n***)

  Examples:

  Type: *LAG (**sample1**,2)*
  Result: The program inserts two missing values at the beginning of the variable **sample1** and moves the first nonmissing value in **sample1** to the third position.

  Type: *LAG (**sample1**,2)*
  Result: The program drops the first two values (observations) in **sample1** and moves the third value to the first position.

- **Recode Numeric or Character Values to Integers (RECODE)**
  Use this operator to recode the values in a numeric or a character variable. The program assigns a new value, starting with 1, to each unique ascending value in the variable.

  Format: RECODE (***var***)

Examples:

Type: *RECODE (**sample4**)*
Result: 3 2 4 1 3

Type: *RECODE (**sample7**)*
Result: 4 3 2 5 1

- **Repeat Each Value n Times (REP)**
  Use this operator to repeat each value in a numeric or a character variable the number of times you specify.

  Format: REP (***var***,***n***)

  Example:

  Type: *REP (**sample1**,2)*
  Result: 1 1 2 2 3 3 4 4 5 5

- **Expand or Compress a Constant or Variable (RESHAPE)**
  Use this operator to expand or compress a numeric or a character variable so it has the number of rows you specify. If the number of rows requires more values than the variable or expression contains, the program cyclically repeats the values. If the number of rows requires fewer values, the program drops the values from the end of the variable.

  Format: RESHAPE (***var***,***n***)

  Examples:

  Type: *RESHAPE (COUNT (1,5,1),12*)
  Result: 1 2 3 4 5 1 2 3 4 5 1 2

  Type: *RESHAPE (**sample2**,3)*
  Result: 5 6 7

- **Remove Double Blanks (STRIPBLANKS)**
  Use this operator to remove the double blanks from a character variable, shifting the characters to the left.

  Format: STRIPBLANKS (***charvar***)

  Example:

  Type: *STRIPBLANKS (**sample8**)*
  Result: Ford Mustang

Time Series Operators

Time series operators calculate seasonal and backward differences. This section describes the time series operators, shows the format you use, and provides examples for you to try.

- **Calculate Multiple Backward Differences (MDIFF)**
  Use this operator to calculate the multiple backward differences, where **n** is the number of repeated differences. Using this operator is equivalent to using the DIFF operator **n** times.

  Format: MDIFF (***numvar***,***n***)

  Example:

  Type: *MDIFF* ***(sample4****,2)*
  Result: (*missing*) (*missing*) 83.0 -135.0 145.0

- **Calculate Seasonal Differences (SDIFF)**
  Use this operator to calculate the seasonal differences, where **n** is the order of differencing, such as 4 for quarterly data. The value at index **i** is ($x_{\mathbf{i}}$ - $x_{\mathbf{i-n}}$).

  Format: SDIFF ***(numvar,n)***

  Example:

  Type: *SDIFF* ***(sample4****,2)*
  Result: *(missing) (missing)* 21.0 -31.0 -21.0

Using Distribution Operators

Distribution operators calculate the probability of a distribution at a given point or the value of a distribution given a probability. This section describes the distribution operators, shows the format you use, and provides examples for you to try.

- **Calculate the Probability of the Cumulative Beta Distribution (BETA)**
  Use this operator to calculate the probability of the cumulative beta distribution with parameters **m** and **l** evaluated at **n** (0 <= **n** <= 1).

  Format: BETA (**n**,***m***,***l***)

Example:

Type: *BETA (0.5,2,5)*
Result: 0.890624

- **Calculate the Probability of the Cumulative Chi-Square Distribution (CHISQUARE)**
  Use this operator to calculate the probability of the cumulative chi-square distribution with **df** degrees of freedom evaluated at
  n (0 <= **n**).

  Format: CHISQUARE (***n***,***df***)

  Example:

  Type: *CHISQUARE (30,20)*
  Result: 0.930146

- **Calculate the Critical Value of the Beta Distribution (INVBETA)**
  Use this operator to calculate the critical value of the beta distribution with parameters **m** and **l** evaluated at
  n (0 <= **n** <= 1).

  Format: INVBETA (***n***,***m***,***l***)

  Example:

  Type: *INVBETA (0.9,2,5)*
  Result: 0.510318

- **Calculate the Critical Value of the Chi-Square Distribution (INVCHISQUARE)**
  Use this operator to calculate the critical value of the chi-square distribution with **df** degrees of freedom evaluated at **n** (0 <= **n** < 1).

  Format: INVCHISQUARE (***n***,***df***)

  Example:

  Type: *INVCHISQUARE (.9,20)*
  Result: 28.3989

- **Calculate the Critical Value of the Normal Distribution (INVNORMAL)**
  Use this operator to calculate the critical value of the normal distribution, evaluated at **n** (0 < **n** < 1).

  Format: INVNORMAL (***n***,***mean***,***sdev***)

Example:

Type: *INVNORMAL (.9,0,1)*
Result: 1.28155

- **Calculate the Critical Value of the Snedecor's-F Distribution (INVSNEDECOR)**
  Use this operator to calculate the critical value of the Snedecor's-F distribution with **df1** and **df2** degrees of freedom evaluated at **n** (0 <= **n** < 1).

  Format: INVSNEDECOR (***n***,***df1***,***df2***)

  Example:

  Type: *INVSNEDECOR (.9,3,20)*
  Result: 2.38053

- **Calculate the Critical Value of the Student's-t Distribution (INVSTUDENT)**
  Use this operator to calculate the critical value of the student's-t distribution with **df** degrees of freedom evaluated at **n** (0 < **n** < 1).

  Format: INVSTUDENT (***n***,***df***)

  Example:

  Type: *INVSTUDENT (.9,20)*
  Result: 1.32534

- **Calculate the Probability of the Cumulative Normal Distribution (NORMAL)**
  Use this operator to calculate the probability of the cumulative normal distribution evaluated at **n**.

  Format: NORMAL (***n***,***mean***,***sdev***)

  Example:

  Type: *NORMAL (2,0,1)*
  Result: 0.97725

- **Calculate the Probability of the Cumulative Snedecor's-F Distribution (SNEDECOR)**
  Use this operator to calculate the probability of the cumulative Snedecor's-F distribution with **df1** and **df2** degrees of freedom evaluated at **n** (0 <= **n**).

Format: SNEDECOR (***n***,***df1***,***df2***)

Example:

Type: *SNEDECOR (4,3,20)*
Result: 0.977923

■ Calculate the Probability of the Cumulative Student's-t Distribution (STUDENT)

Use this operator to calculate the probability of the cumulative student's-t distribution with **df** degrees of freedom evaluated at **n**.

Format: STUDENT (***n***,***df***)

Example:

Type: *STUDENT (2,20)*
Result: 0.970367

C Keyboard Equivalents

Keyboard equivalents provide shortcuts you can use in place of mouse actions. Tables C-1 and C-2 group keyboard equivalents into General Application keystrokes and Spreadsheet keystrokes. Most keystrokes appear next to their corresponding items on the pull-down menus.

Table C-1. General Application Keystrokes

Keystroke	Purpose
F1	Access Online Help
F10	Run the StatFolio Start-Up Sfcript
Alt	Access File Menu
Ctrl-F11	Access Open StatFolio Dialog Box
Shift-F11	Access Save StatFolio Dialog Box or automatically saves if the StatFolio has a name
F11	Access Save StatFolio As Dialog Box
Ctrl-F12	Access Open Data File Dialog Box
Shift-F12	Access Save Data File Dialog Box or automatically saves if the data file has a name
F12	Access Save Data File As Dialog Box
F3	Access Save Graph Dialog Box
Shift-F3	Access Print Preview
F4	Access Print Dialog Box
Shift-F4	Access Print Setup Dialog Box

Table C-1. Continued

Keystroke	Purpose
Alt-F4	Access Exit STATGRAPHICS *Plus* Dialog Box
F7	Redisplay Analysis Dialog Box
F5	Access Tabular Options Dialog Box
F8	Access Graphical Options Dialog Box
F9	Access Save Results Dialog Box
Shift-F5	Access Modify Column Dialog Box (when column selected)
Shift-F7	Access Generate Data Dialog Box (when column selected)
Ctrl-X	Cut Highlighted Text
Ctrl-C	Copy Highlighted Text
Ctrl-V	Insert Highlighted Text (Paste)
Ctrl-Z	Undo Previous Change
F2	Access Font Dialog Box
Ctrl-F4	Delete an Analysis
Ctrl-F5	Restore the highlighted Taskbar/Icon
Ctrl-F6	Navigates among the Taskbars/Icons
Ctrl-F7	Displays Move Icon; Use Arrow Keys to Move
Ctrl-F8	Displays Size Icon; Use Arrow Keys to Resize
Ctrl-F9	Minimize a Window
Ctrl-F10	Maximize a Window

Table C-2. Spreadsheet Keystrokes

Keystroke	Purpose
Ins	Inserts a blank block of cells before a tagged block which provides the row and column specifications
Del	Deletes a tagged block
Arrow Keys	Moves one cell in the specified direction
PgUp	Moves up one page
PgDn	Moves down one page
Home	Move to the first column of the current row
End	Move to the last column of the current row
Ctrl-PgUp	Move one page to the left
Ctrl-PgDn	Move one page to the right
Ctrl-Home	Move to the first column in the first row
Ctrl-End	Move to the last column in the last row
Ctrl-Arrow Key	Move to the last nonblank cell in the specified direction
Tab	Moves one cell to the right
Shift-Tab	Moves one cell to the left
Shift-Arrow Key	Select cells in the specified direction
Shift-PgUp/PgDn	Select one page of cells in the specified direction
Shift-Spacebar	Select the current row
Ctrl-Spacebar	Select the current column
Shift-Ctrl-Spacebar	Select the entire spreadsheet

Table C-2. Continued

Keystroke	Purpose
Shift-Ctrl-Arrow Key	Select cells between the active cell and the end of the spreadsheet in the specified direction

D References

Abernethy, R. B., Breneman, J. E., Medlin, C. H., and Reinman, G. L. 1983. *Weibull Analysis Handbook*. Final Report for Period 1 July 1982 to 31 August 1983. Wright-Patterson AFB, Ohio: United States Air Force.

Anderson, T. W. 1958. *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Ansell, J. I. and Phillips, M. J. 1994. *Practical Methods for Reliability Data Analysis*. London: Oxford Science Publications, Clarendon Press.

Armitage, P. and Berry G. 1987. *Statistical Methods in Medical Research*, second edition. Oxford: Blackwell Scientific Publications.

Belsley, D. A., Kuh, E., and Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity*. New York: Wiley.

Berkson, B. J. and Gage, R. P. 1950. "Calculation of Survival Rates for Cancer," *Proc. Staff Meet.* Mayo Clinic, **25:**270-286.

Box, G. E. P. and Draper, N. R. 1987. *Empirical Model-Building and Response Surfaces*. New York: Wiley.

Box, G. E. P. and Jenkins, G. M. 1976. *Time Series Analysis, Forecasting and Control*, second edition. San Francisco: Holden-Day.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. New York: Wiley.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. 1983. *Graphical Methods for Data Analysis*. *The Wadsworth Statistics/Probability Series*, ed. by P. J. Bickel, W. S. Cleveland, and R. M. Dudley. Co-publishing project of Wadsworth International Group and Duxbury Press, divisions of Wadsworth, Inc. Belmont, California: Wadsworth International Group.

Cleveland, W. S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of American Statistical Association*, **74**:829-836.

Cleveland, W. S. 1981. "LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression," *The American Statistician*, **35**:54.

Collett, D. 1996. *Modelling Survival Data in Medical Research*. London: Chapman & Hall.

Cornell, J. A. 1973. "Experiments with Mixtures: A Review," *Technometrics*, **15**:437-455.

Cornell, J. A. 1990. *Experiments with Mixtures*, second edition. New York: Wiley & Sons.

Cornell, J. A. and Piepel, G. F. 1993. *Design and Analysis of Mixture Experiments*. Computer Associates.

Cox, D.R. and Lewis, P.A.W. 1966. *The Statistical Analysis of Series of Events*. London: Methuen and Company.

D'Agostino, R. B. and Stephens, M. A. 1986. *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

Desu, M. M. and Raghavarao, D. 1990. *Sample Size Methodology*. In *Statistical Modeling and Decision Science*, ed. by G. J. Lieberman and I. Olkin. San Diego, California: Academic Press.

Draper, N. and Smith, H. 1981. *Applied Regression Analysis*, second edition. New York: Wiley.

Durbin, J. and Watson, G. S. 1951. "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, **38**.

Escobar, Luis A. and Meeker, William Q. 1998. *Statisstical Methods for Reliability Data*. New York: Wiley.

Everitt, B. S. 1977. *The Analysis of Contingency Tables.* New York: Routledge Chapman & Hall.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fogiel, M. and the Staff of the Research and Education Association. 1978. *The Statistical Problem Solver*. Piscataway, New Jersey: Research and Education Association.

Freund, J. E. and Williams, F. J. 1977. *Elementary Business Statistics, The Modern Approach*. New Jersey: Prentice-Hall.

Frigge, M., Hoagland, D. C., and Iglewicz, B. 1989. "Some Implementations of the Boxplot," *American Statistician*, **43**:50-54.

Gibbons, Jean D. 1976. *Nonparametric Methods for Quantitative Analysis*. New York: Holt, Rinehart and Winston.

Guttman, I., Wilks, S. S., and Hunter, J. S. 1982. *Introductory Engineering Statistics*, third edition. New York: Wiley.

Haaland, P. 1989. *Experimental Design in Biotechnology.* New York: Marcel Dekker.

Hastings, N. A. J. and Peacock, J. B. 1975. *Statistical Distributions*. London: Butterworth and Co.

Hays, W. L. 1981. *Statistics*, third edition. New York: Holt, Rinehart and Winston.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., eds. 1991. *Fundamentals of Exploratory Analysis of Variance*. New York: Wiley.

Hollander, M. and Wolfe, D. A. 1973. *Nonparametric Statistical Methods*. New York: John Wiley and Sons, Inc.

Iglewicz, Boris and Hoaglin, David C. 1993. *How to Detect and Handle Outliers.* Milwaukee: ASQ Quality Press.

Jackson, J. E. 1959. “Quality Control Methods for Several Related Variables,” *Technometrics*, 1:4.

Johnson, N. L. and Kotz, S. 1970. *Continuous Univariate Distributions - 1*. New York: Wiley.

Johnson, R. A. and Wichern, D. W. 1982. *Applied Multivariate Statistical Analysis*. New Jersey: Prentice-Hall.

Kalbfleisch, J. D. and Prentice, R. L. 1980. *Statistical Analysis of Failure Time Data*. New York: Wiley.

Kempthorne, O. and Folks, L. 1971. *Probability, Statistics, and Data Analysis*. Ames, Iowa: Iowa University Press.

Kvalseth, T. O. 1985. “Cautionary Note about R^2,” *The American Statistician*, **39**.

Lapin, L. L. 1987. *Statistics for Modern Business Decisions*, fourth edition. Orlando, Florida: Harcourt Brace Jovanovich, Inc.

Law, A. M. and Kelton, W. D. 1982. *Simultation Modeling and Analysis*. New York: McGraw-Hill.

Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons.

Levene, H. 1960. *Robust Tests for Equality of Variances*. In: *Contributions to Probability and Statistics*, ed. by Olkin, et. al. Stanford University Press.

Madansky, A. 1988. *Prescriptions for Working Statisticians.* New York: Springer-Verlag.

McGill, R., Tukey, J. W., and Larsen, W. A. 1978. "Variation of Box Plots," *American Statistician*, **32**:12-16.

Milliken, G. A. and Johnson, D. E. 1984. *Analysis of Messy Data,* Vol. I: *Designed Experiments*. New York: Van Nostrand Reinhold Company.

Montgomery, D. C. 1991. *Design and Analysis of Experiments*, third edition. New York: Wiley.

Montgomery, D. C. 1991. *Introduction to Statistical Quality Control*, second edition. New York: John Wiley & Sons.

Montgomery, D. C. and Peck, E. A. 1992. *Introduction to Linear Regression Analysis*, second edition. New York: Wiley & Sons.

Morrison, D. F. 1990. *Multivariate Statistical Methods*, third edition. New York: McGraw-Hill Publishing Co.

Myers, R. H. 1990. *Classical and Modern Regression with Applications*, second edition. Belmont, California: Duxbury Press.

Nelson, W. B. 1982. *Applied Life Data*. New York: Wiley.

Neter, J., Wasserman, W., and Kutner, M. 1985. *Applied Linear Statistical Methods.* Homewood, Illinois: Richard E. Erwin, Inc.

Olkin, I., ed., *Statistical Modeling and Decision Science*. San Diego, California: Academic Press.

Ott, E. R. 1983. "Analysis of Means: A Graphical Procedure," *Journal of Quality Technology*, **15**:10-18.

Ott, E. R. and E. G. Schilling. 1990. *Process Quality Control*, second edition. New York: McGraw-Hill Book Company.

Owen, D. B. 1962. *Handbook of Statistical Tables*. Redding, Massachusetts: Addison-Wesley.

Searle, S. R., Casella, G., and McCulloch, C. E. 1991. *Variance Components*. In: *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*, ed. by V. Barnett, R. A. Bradley, N. I. Fisher, J. S. Hunter, J. B. Kadane, D. G. Kendall, A. F. M. Smith, S. M. Stigler, J. Teugels, and G. S. Watson. New York: John Wiley & Sons, Inc.

Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Siegel, S. and Castellan, N. J. 1988. *Nonparametric statistics for the behavioral Sciences*, second editon. New York: McGraw-Hill.

Snedecor, G. W. and Cochran, W. G. 1967. *Statistical Methods*, sixth edition. Ames, Iowa: Iowa State University Press.

Snee, R. D. 1974. “Graphical Display of Two-way Contingency Tables,” *American Statistician*, **28**:9-12.

Somers, R. H. 1962. “A New Symmetric Measure of Association for Ordinal Variables.” *American Sociological Review*, **27**:799-811.

Tatsuoka, M. M. 1971. *Multivariate Analysis*. New York: Wiley.

The Math Forum. 1998. “Ask Dr. Math,” Forum SmartPage Web Tool.

Tobias, P. A. and Trindade, D. C. 1995. *Applied Reliability*, second edition. London: Chapman & Hall.

Tufte, E. R. 1990. *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Velleman, P. F. and Hoaglin, D. C. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Belmont, California: Duxbury Press.

Vogt, Paul. 1993. *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences*. Newbury Park, California: Sage Publications, Inc.

Weiss, N. A. and Hassett, M. J. 1991. *Introductory Statistics*, third edition. New York: Addison-Wesley Publishing Company.

Winer, B. J. 1971. *Statistical Principles in Experimental Design*, second edition. New York: McGraw-Hill.

Wonnacott, T. H. and Wonnacott, R. J. 1972. *Introductory Statistics*, second edition. New York: Wiley.

E Calculations

Arrhenius Plot

The Arrhenius model fits the form:

$$P = A \exp(\Delta / kt)$$

where

$k = 8.617\,10^{-5}$ EV ⁄ degrees Kelvin (Boltzmann's Constant)

t = temperature

Box-Cox Transformation

The Box-Cox Transformation fits a linear regression model relating the transformed values of a dependent variable y to a single predictor variable x, according to the equation:

$$w_i = \beta_0 + \beta_1 x_i \qquad (i = 1, 2, ..., n)$$

where w transforms the dependant variable according to:

$$w = \begin{cases} \dfrac{(y + \lambda_2)^{\lambda 1} - 1}{\lambda_1 \, g_m^{\lambda 1\text{-}1}} & \text{if } \lambda_1 \neq 0 \\ g_m \ln(y + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases}$$

where g_m is the geometric mean of $(y + \lambda_2)$.

λ_1 is estimated from the data so as to minimize the residual mean square error.

λ_2 is specified by the user.

Box-and-Whisker Plot

Center of the Boxplot:

The center of the boxplot is the sample median. When there are an even number of observations, the program calculates the median as the average of observation n/2 and observation (n/2)+1 after sorting the observations in numerical order.

Sample Quartiles Q_1 and Q_3:

First Quartile (Q_1):

$$i = \text{ceiling}\left[\frac{(n+1)}{4}\right]$$

$$j = \text{ceiling}\left[\frac{n}{4}\right]$$

$$\text{First Quartile} = \frac{\left(X_{[i]} + X_{[j]}\right)}{2}$$

where $x_{[i]}$ is the ith order statistic of the variable x.

Third Quartile (Q_3):

$$i = \text{ceiling}\left[\frac{3(n+1)}{4}\right]$$

$$j = \text{ceiling}\left[\frac{3n}{4}\right]$$

$$\text{Third Quartile} = \frac{\left(X_{[i]} + X_{[j]}\right)}{2}$$

where $x_{[i]}$ is the ith order statistic of the variable x.

Notches for the Median:

$$\text{median} \pm \left[\frac{(1.25\ \text{IQR})}{\left(1.35\sqrt{n}\right)}\right]\left[\left(\frac{Z_{\alpha/2}}{\sqrt{2}}\right) + Z_{\alpha/2}\right] / 2$$

Comparison of Proportions

Notation:

k = number of groups

p_i = proportions for group i

n_i = sample size for group i

h_α = critical value for multivariate t distribution with ∞ degrees of freedom

$$\overline{p} = \frac{\Sigma\ n_i c_i}{\Sigma\ n_i} \qquad \text{for } i = 1, ..., k$$

$$\overline{n} = \frac{\Sigma\ n_i}{k}$$

$$s = \sqrt{\frac{\overline{p}(1 - \overline{p})}{\overline{n}}}$$

$$\text{LDL} = \overline{p} - h_\alpha s\sqrt{\frac{(k-1)}{k}}$$

$$\text{UDL} = \overline{p} + h_\alpha\ s\sqrt{\frac{(k-1)}{k}}$$

Comparison of Rates

Notation:

K = number of groups

H = the critical value of a multivariate t distribution for k groups

$\hat{\lambda}_j$ = sample rates

T_j = sampling interval

To test the hypothesis that the rates are equal, a likelihood ratio test statistic is computed from the rates $\hat{\lambda}_j$ and T_j, according to

$$H = 2\left\{ \sum_{j=1}^{m} \hat{l}_f T_j \log\left(\hat{l}_j\right) - \left(\sum_{j=1}^{m} \hat{l}_f T_j \right) \log\left(\sum_{j=1}^{m} \hat{l}_f T_j \Big/ \sum_{j=1}^{m} T_j \right) \right\}$$

The ANOM table and plot:

The centerline of the chart is placed at the following:

$$CL = \hat{\lambda} = \frac{\sum_{j=1}^{m} T_j \hat{\lambda}_j}{\sum_{j=1}^{m} T_j}$$

And the decision limit is placed at the following:

$$CL \pm h \sqrt{\frac{\hat{\lambda}(m-1)}{\sum_{j=1}^{m} T_j}}$$

Crosstabulation

Notation:

r = number of rows in the table

c = number of columns in the table

f_{ij} = observed frequency in position (row i, column j)

x_i = distinct values of the row variable arranged in ascending order, i=1,...,r

y_j = distinct values of the column variable arranged in ascending order, j=1,...,c

E_{ij} = expected frequency in position (row i, column j)

Totals:

$$R_i = \sum_{j=1}^{c} f_{ij} \quad C_j = \sum_{i=1}^{r} f_{ij} \quad N = \sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij}$$

Note: Any row or column that totals 0 is eliminated from the table before the program performs the calculations.

Chi-square:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - E_{ij})^2}{E_{ij}}$$

where E_{ij} is the expected value for cell ij under H_o.

$$E_{ij} = R_i \frac{C_j}{N}$$

Fisher's Exact Test:

Run for a two-by-two table, when N is less than or equal to 100. For calculation details, see standard references such as *The Analysis of Contingency Tables* by B. S. Everitt.

Lambda:

with rows dependent

$$\lambda = \frac{\left(\sum_{j=1}^{c} f_{max,\ j} - R_{max} \right)}{N - R_{max}}$$

with columns dependent

$$\lambda = \frac{\left(\sum_{i=1}^{r} f_{i,\ max} - C_{max} \right)}{N - C_{max}}$$

when symmetric

$$\lambda = \frac{\left(\sum_{i=1}^{r} f_{i,\ max} + \sum_{j=1}^{c} f_{max,\ j} - R_{max} - C_{max} \right)}{2N - R_{max} - C_{max}}$$

where

$f_{i,max}$ = largest value in row i

$f_{max,j}$ = largest value in column j

R_{max} = largest row total

C_{max} = largest column total

Uncertainty Coefficient:

with rows dependent

$$U_R = \frac{U(R) + U(C) - U(RC)}{U(R)}$$

with columns dependent

$$U_C = \frac{U(R) + U(C) - U(RC)}{U(C)}$$

when symmetric

$$U = 2\left(\frac{U(R) + U(C) - U(RC)}{U(R) + U(C)} \right)$$

where

$$U(R) = -\sum_{i=1}^{r} \frac{R_i}{N} \ln \frac{R_i}{N}$$

$$U(C) = -\sum_{j=1}^{c} \frac{C_j}{N} \ln \frac{C_j}{N}$$

$$U(RC) = -\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{f_{ij}}{N} \ln \frac{f_{ij}}{N} \qquad \text{for } f_{ij} > 0$$

Somer’s D:

rows dependent

$$D_R = \frac{2(P_C - P_D)}{\left(N^2 - \sum_{j=1}^{c} C_j^2\right)}$$

columns dependent

$$D_C = \frac{2(P_C - P_D)}{\left(N^2 - \sum_{i=1}^{r} R_i^2\right)}$$

when symmetric

$$D = \frac{4(P_C - P_D)}{\left(N^2 - \sum_{i=1}^{r} R_i^2\right) + \left(N^2 - \sum_{j=1}^{c} C_j^2\right)}$$

where the number of concordant pairs is

$$P_C = \sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij} \sum_{h>i} \sum_{k>j} f_{hk}$$

and the number of discordant pairs is

$$P_D = \sum_{i=1}^{r} \sum_{j=1}^{c} f_{ij} \sum_{h>i} \sum_{k<j} f_{hk}$$

Eta:

with rows dependent

$$E_R = \sqrt{1 - \frac{SS_{CN}}{SS_C}}$$

where the total corrected sum of squares for the rows is

$$SS_R = \sum_{i=1}^{r} \sum_{j=1}^{c} x_i^2 f_{ij} - \frac{\left(\sum_{i=1}^{r} \sum_{j=1}^{c} x_i^2 f_{ij}\right)^2}{N}$$

and the sum of squares of rows within categories of columns is

$$SS_{RN} = \sum_{j=1}^{r} \left(\sum_{i=1}^{c} x_i^{2} f_{ij} - \frac{\left(\sum_{i=1}^{r} x_i^{2} f_{ij} \right)^{2}}{C_j} \right)$$

with columns dependent

$$E_C = \sqrt{1 - \frac{SS_{CN}}{SS_C}}$$

where the total corrected sum of squares for the columns is

$$SS_C = \sum_{i=1}^{r} \sum_{j=1}^{c} y_i^{2} f_{ij} - \frac{\left(\sum_{i=1}^{r} \sum_{j=1}^{c} y_j \; f_{ij} \right)^{2}}{N}$$

and the sum of squares of columns within categories of rows is

$$SS_{CN} = \sum_{i=1}^{r} \left(\sum_{j=1}^{c} y_j^{2} f_{ij} - \frac{\left(\sum_{j=1}^{c} y_j^{2} f_{ij} \right)^{2}}{R_i} \right)$$

Contingency Coefficient:

$$C = \sqrt{\frac{\chi^{2} / N}{1 + \chi^{2} / N}}$$

Cramer's V:

$$V = \sqrt{\frac{\chi^{2}}{N}}$$ for a two-by-two table

$$V = \sqrt{\frac{\chi^2}{N(m-1)}}$$ for all others where m = min (r-1,c-1)

Conditional Gamma:

$$G = \frac{P_C - P_D}{P_C + P_D}$$

Pearson's R:

$$R = \sum_{j=1}^{c}\sum_{i=1}^{r} x_i y_j f_{ij} - \frac{\frac{\left(\sum_{j=1}^{c}\sum_{i=1}^{r} x_j f_{ij}\right)\left(\sum_{j=1}^{c}\sum_{i=1}^{r} y_j f_{ij}\right)}{N}}{\sqrt{SS_R SS_C}}$$

If R=1, no significance is printed. Otherwise, the one-sided significance is based on

$$t = R\sqrt{\frac{N-2}{1-R^2}}$$

Kendall's Tau b:

$$\tau_b = \frac{2(P_C - P_D)}{\sqrt{\left(N^2 - \sum_{i=1}^{r} R_i{}^2\right)\left(N^2 - \sum_{j=1}^{c} C_j{}^2\right)}}$$

Kendall's Tau c:

$$\tau_c = \frac{2m(P_C - P_D)}{(m-1)N^2}$$

Distribution Fitting

Exponential Distribution:

$$f(x) \;=\; \frac{1}{\beta} e^{-x/\beta} \qquad , \qquad x>0$$

$$E(x) \;=\; \beta \qquad , \qquad \operatorname{var}(x) = \beta^2$$

Extreme Value Distribution:

$$f(x) = \left(\frac{1}{\beta}\right) \exp\left[\left(\frac{1}{\beta}\right)(x-\alpha) - \exp\left[\left(\frac{1}{\beta}\right)(x-\alpha)\right]\right] ,$$

all real x

$$E(x) = a + b\Gamma(1) \qquad , \qquad \operatorname{var}(x) = \frac{b^2\pi^2}{6}$$

where $\Gamma(1) = -0.57721$

Lognormal Distribution:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/2\sigma^2} \qquad , \qquad x > 0$$

$$E(x) = e^{\mu + \sigma^2/2}$$

$$\operatorname{var}(x) = \exp(2\mu + \sigma^2) \; [\exp(\sigma^2) - 1]$$

Normal Distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} \qquad , \qquad \text{all real x}$$

$$E(x) = \mu \qquad , \qquad \operatorname{var}(x) \;=\; \sigma^2$$

Weibull Distribution:

$$f(x) \; = \; \alpha\beta^{-\alpha} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} \qquad , \qquad x>0$$

$$E(x) \; = \; \beta\Gamma\left(1+\frac{1}{\alpha}\right)$$

$$var(x) \; = \; \left(\frac{\beta^{2}}{\alpha}\right)\left\{2\,\Gamma\left(\frac{2}{\alpha}\right)-\left(\frac{1}{\alpha}\right)\left[\Gamma\left(\frac{1}{\alpha}\right)\right]^{2}\right\}$$

where $\Gamma\ (1/\alpha)$ is the gamma function with parameter $1/\alpha$.

Distribution Plotting

Bernoulli Distribution:

$$p(x) = p^{x}\left(1-p\right)^{1-x} \qquad , \qquad x=0,1$$

Binomial Distribution:

$$p(x) = \binom{n}{x} p^{x}\,(1-p)^{n-x} \qquad , \qquad x=0,1,2,...,n$$

Discrete Uniform Distribution:

$$p(x) = \frac{1}{b-a+1} \qquad , \qquad x=a,a+1,...,b$$

Geometric Distribution:

$$p(x) = p(1-p)^{x} \qquad , \qquad x=1,2,3,...$$

Hypergeometric Distribution:

$$f\,(x;\,n,\,M,\,N) = \frac{\begin{pmatrix} M \\ x \end{pmatrix}\begin{pmatrix} N-M \\ n-x \end{pmatrix}}{\begin{matrix} N \\ n \end{matrix}} \qquad x=0,1,2,...,n$$

Negative Binomial Distribution:

$$p(x)=\begin{pmatrix} x\text{ - }1 \\ k\text{ - }1 \end{pmatrix} p^{\,k}(1-p)^{x-k} \qquad , \qquad x=k, k+1, k+2,...$$

Poisson Distribution:

$$P(x) = \frac{\beta^{x}e^{-\beta}}{x!} \qquad , \qquad x=0,1,2,...$$

Beta Distribution:

$$f(x) \;=\; \frac{x^{\alpha\text{-}1}(1\text{-}x)^{\beta\text{-}1}}{\beta(\alpha,\beta)} \qquad , \qquad 0 < x < 1$$

where $\beta\,(\alpha, \beta)$ is the beta function defined by

$$\beta\,(Z_{1,}\;Z_2) = \int_0^1 t^{Z_1-1}(1-t)^{Z_2-1}\,dt$$

for any real numbers

$Z_1 > 0$ and $Z_2 > 0$

Cauchy Distribution:

$$f(x) = \left\{ \pi\,\beta \left[1 + \left(\frac{x-\alpha}{\beta} \right)^{2} \right] \right\}^{-1} \qquad , \qquad \beta > 0$$

Chi-Square Distribution:

$$f(x) \;=\; \frac{x^{(v-2)/2} \quad e^{\,(-x/2)}}{2^{(v/2)}\,\Gamma\!\left(\frac{v}{2}\right)} \qquad , \qquad x > 0$$

Erlang Distribution:

$$f(x) \;= \frac{\left(\frac{x}{b}\right)^{c-1} e^{\left(-x/b\right)}}{[b(c-1)!]}$$

Exponential Distribution:

$$f(x) \;=\; \frac{1}{\beta}\,e^{-x/\beta} \qquad , \qquad x > 0$$

Extreme Value Distribution:

$$f(x) = \left(\frac{1}{\beta}\right)\exp\left[\left(\frac{1}{\beta}\right)(x-\alpha) - \exp\left[\left(\frac{1}{\beta}\right)(x-\alpha)\right]\right], \qquad \text{all real x}$$

F Distribution:

$$f(x) = \frac{\Gamma\left[\left(\frac{v+w}{2}\right)\right]\left(\frac{v}{w}\right)^{(v-2)} x^{(v-2)/2}}{\Gamma\left(\frac{v}{2}\right)\Gamma\left(\frac{w}{2}\right)\left[1+\left(\frac{v}{w}\right)\right] x^{(v+w)/2}} \qquad , \qquad x > 0$$

where $\Gamma\ (v\,/\,2)$ is the gamma function with parameter v / 2
and where v and w are the numerator and denominator degrees of freedom.

Gamma Distribution:

$$f(x) = \frac{\beta^{-\alpha} x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)} \qquad , \qquad x > 0$$

Laplace Distribution:

$$f(x) = \left(\frac{\beta}{2}\right) e^{\left(-\beta|x-\alpha|\right)}$$

Logistic Distribution:

$$f(x) = \frac{e^{\frac{x-a}{k}}}{k\left[1 + \exp^{\left[\frac{x-a}{k}\right]}\right]^2}$$

Lognormal Distribution:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\left(\ln x - \mu\right)^2/2\sigma^2} \qquad , \qquad x > 0$$

Normal Distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(x-\mu\right)^2/2\sigma^2} \qquad , \qquad \text{all real x}$$

Pareto Distribution:

$$f(x) = \frac{k}{\beta\left(1 + \frac{x}{\beta}\right)^{k+1}} \qquad , \qquad x > 0$$

Student’s t Distribution:

$$f(x) \ = \frac{\Gamma\left[\left(\frac{v+1}{2}\right) 1 + \left(\frac{x^2}{v}\right)\right]^{-(v+1)/2}}{\left(v\pi\right)^{1/2} \Gamma\left(\frac{v}{2}\right)} \qquad , \qquad \text{all real x}$$

Triangular Distribution:

$$f(x) = \frac{2(x-a)}{(b-a)(c-a)} \quad , \quad \text{for } a \le x \le c$$

$$f(x) = \frac{2(b-x)}{(b-a)(b-c)} \quad , \quad \text{for } c < x \le b$$

Uniform Distribution:

$$f(x) = \frac{1}{(b-a)} \quad , \quad a \le x \le b$$

Weibull Distribution:

$$f(x) \; = \; \alpha\beta^{-\alpha} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} \quad , \quad x > 0$$

Frequency Histogram

For information on the calculations used in plotting frequency histograms, see Lapin (1987).

Hypothesis Tests (Compare)

Difference of Means (equal sigma):

$100(1-\alpha)$**%** Confidence Interval for $\mu_1 - \mu_2$**.**

If it is assumed that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, where σ^2 is unknown

$$\left[\left(\bar{x}_1 - \bar{x}_2 \right) \pm t_{n1+n2\text{-}2;\ a/2} s_W \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

$$s_W \; = \; \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Test statistic

$$\frac{\overline{x}_1 - \overline{x}_2}{s_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n1\ -\ n2\ -\ 2}$$

Difference of Means (unequal sigma):

Compute the quantities c and m where

$$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

and the modified degrees of freedom for t

$$\frac{1}{m} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}$$

then

$$\left[\left(\overline{x}_1 - \overline{x}_2\right) \pm t_{m;\,\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

is an approximate 100(1 – α)**%** Confidence Interval for $\mu_1 - \mu_2$**.**

Ratio of Variances:

Confidence interval

If s_1^2 and s_2^2 are variances of independent samples of size \n $_{1\backslash}$ and n_2 from the normal distributions $N(\mu_1,\sigma_1^2)$ and $N\left(\mu_2,\sigma_2^2\right)$ respectively, where μ_1, μ_2, σ_1^2, and σ_2^2 are unknown, then

$$\left[\left(\frac{s_1^2}{s_2^2}\right)\left(\frac{1}{F_{n1\text{-}1,\,n2\text{-}1;\alpha/2}}\right),\left(\frac{s_1^2}{s_2^2}\right)\left(\frac{1}{F_{n1\text{-}1,\,n2\text{-}1;\alpha/2}}\right)\right]$$

Test statistic

$$F = \frac{s_1^2 / s_2^2}{\sigma_1^2 / \sigma_2^2} \sim F_{n1 - 1,\ n2 - 1}$$

Power:

Difference of means

$t_c = t_{\alpha/2,\, v}$ where *v* = degrees of freedom shown for confidence interval

$$P = 1 - F\left(t_c - \frac{(\Delta - \Delta_0)}{\text{s.e.}} \right) + F\left(-t_c - \frac{(\Delta - \Delta_0)}{\text{s.e.}} \right)$$

where

$$\text{s.e.} = \sigma_0 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{if equal sigma}$$

or

$$\text{s.e.} = \sqrt{\frac{\sigma_{0,1}^2}{n_1} + \frac{\sigma_{0,2}^2}{n_2}} \quad \text{if unequal sigma}$$

Ratio of Variances:

The power functions of the left-sided, right-sided, and two-sided tests are respectively:

$$\gamma(\lambda) = P\left(F_{n_1\text{-}1,\, n_2\text{-}1} < \frac{1}{\lambda} F_{n_1\text{-}1,\, n_2\text{-}1;\, 1\text{-}\alpha} \right)$$

$$\gamma(\lambda) = P\left(F_{n_1\text{-}1,\, n_2\text{-}1} < \frac{1}{\lambda} F_{n_1\text{-}1,\, n_2\text{-}1;\alpha} \right)$$

and

$$\gamma(\lambda) = P\left(F_{n_1\text{-}1,\, n_2\text{-}1} < \frac{1}{\lambda} F_{n_1-1,\, n_2-1;1-\alpha/2} \right)$$

$$+ P\left(F_{n_1\text{-}1,\, n_2\text{-}1} > \frac{1}{\lambda} F_{n_1-1,\, n_2-1;\alpha/2} \right)$$

where

$$\lambda = \frac{\sigma_1^2 / \sigma_2^2}{\sigma_{0,1}^2 / \sigma_{0,2}^2}$$

Difference between Proportions:

Confidence interval

Normal approximation

$$\left(\hat{P}_1 - \hat{P}_2\right) \pm z_{\alpha/2} \sqrt{\frac{\hat{P}_1\left(1-\hat{P}_1\right)}{n_1} + \frac{\hat{P}_2\left(1-\hat{P}_2\right)}{n_2}}$$

Test statistic

$$Z = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{\hat{P}_1\left(1-\hat{P}_1\right)}{n_1} + \frac{\hat{P}_2\left(1-\hat{P}_2\right)}{n_2}}}$$

Hypothesis Tests (Describe)

Normal Mean:

Confidence interval

$$\overline{x} \pm t_{n\text{-}1,\alpha/2} \frac{s}{\sqrt{n}}$$

Test statistic

$$t = \frac{\overline{x} - \mu_0}{s / \sqrt{n}} \sim t_{n-1}$$

Normal Sigma:

Confidence interval

$$\left[\frac{s\sqrt{n-1}}{\chi_{n-1;\alpha/2}}, \frac{s\sqrt{n-1}}{\chi_{n-1;1-\alpha/2}}\right]$$

Test statistic

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2} \sim \chi_n^2 - 1$$

Binomial parameter:

Confidence interval

$$p_1 = \frac{\nu_1 F_{\nu 1,\, \nu 2,\, \alpha/2}}{\left(\nu_2 + \nu_1 F_{\nu 1,\, \nu 2,\, \alpha/2}\right)}$$

where

$$\nu_1 = 2x$$
$$\nu_2 = 2(n - x + 1)$$

$$p_u = \frac{\nu_1 F_{\nu_1, \nu_2, 1-\alpha/2}}{\left(\nu_2 + \nu_1 F_{\nu_1, \nu_2, 1-\alpha/2}\right)}$$

where

$$\nu_1 = 2(x + 1)$$
$$\nu_2 = 2(n - x)$$

Test statistic

If $np(1-p) > 25$

$$Z = \frac{X + \frac{1}{2} - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0, 1)$$

else

$$p = 2 \times \operatorname{Min}\left[\sum_{\leq x} p_b(X), \sum_{\geq x} p_b(X)\right]$$

Poisson:

Confidence interval

$$\Theta_L = \frac{1}{2}\chi^2_{2x, \alpha/2}$$

$$\Theta_L = \frac{1}{2}\chi^2_{2(x+1), 1-\alpha/2}$$

Test statistic

If $n\Theta > 25$,

$$Z = \frac{X + \frac{1}{2} - n\lambda_0}{\sqrt{n\lambda_0}} \sim N(0,1)$$

$$p = 2 \times \operatorname{Min}\left[\sum_{\leq x} p_p(X), \sum_{\geq x} p_p(X)\right]$$

Power

Normal Mean:

$$t_c = t_{\alpha/2,\, n-1}$$

$$P = 1 - F\left(t_c - \frac{(\mu - \mu_0)}{\sigma_0 / \sqrt{n}} \right)$$

$$1 + F\left(-t_c - \frac{(\mu - \mu_0)}{\sigma_0 / \sqrt{n}} \right)$$

Normal Sigma:

$$\chi_1^2 = \chi^2_{1-\alpha/2,\, n-1}$$

$$\chi_2^2 = \chi^2_{\alpha/2,\, n-1}$$

$$P = 1 - F\left(\chi_2^2 \frac{\sigma_0^2}{\sigma^2} \right) + F\left(\chi_1^2 \frac{\sigma_0^2}{\sigma^2} \right)$$

Binomial Parameter and Poisson Rate:

$$P = 1.0 - F\left(X_u - 1, P, n \right) + F\left(X_{L,}\ P, n \right)$$

where X_u and X_L are critical values leaving no more than $\alpha/2$ in the tail of the curve.

Life Tables

For information on the calculations used in the Life Tables analyses see: Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

Multifactor ANOVA

For information on the calculations used in multifactor analysis of variance, see Neter, et al. (1985).

Multiple Regression

The program uses Gram-Schmidt decomposition (with tolerance = 1.0E – 08) to estimate the coefficients.

Notation:

$\underset{\sim}{Y}$ = vector of n observations for the dependent variable

$\underset{\sim}{X}$ = n-by-p matrix of observations for p-1 independent variables and the constant term, if any

$\sim$ = a variable that is a vector or matrix

Mean:

$$\overline{Y} = \frac{\sum\limits_{i=1}^{n} Y_i}{n}$$

where n is the number of observations.

Estimated Coefficients:

$$\underset{\sim}{b} = (\underset{\sim}{X}' \ \underset{\sim}{X})^{-1} \underset{\sim}{X}' \underset{\sim}{Y}$$

where X' is the transpose of $\underset{\sim}{X}$.

Standard Errors:

$$S(\underset{\sim}{b}) = \sqrt{\text{diagonal elements of } (\underset{\sim}{X}' \underset{\sim}{X})^{-1} \text{ MSE}}$$

where

$$\text{SSE} = \underset{\sim}{Y}' \underset{\sim}{Y} - \underset{\sim}{b}' \underset{\sim}{X}' \underset{\sim}{Y}$$

$$\text{MSE} = \frac{\text{SSE}}{n - p}$$

and where p is the number of coefficients estimated.

t-Values:

$$\underset{\sim}{t} = \frac{\underset{\sim}{b}}{S(\underset{\sim}{b})}$$

Significance Level:

t-values follow the Student’s-t distribution with n – p degrees of freedom.

R-Squared:

$$R^2 = \frac{SSTO - SSE}{SSTO}$$

where

$$SSTO = \begin{cases} \underset{\sim}{Y}'\underset{\sim}{Y} - n\overline{Y}^2 & \text{if constant in model} \\ \underset{\sim}{Y}'\underset{\sim}{Y} & \text{if no constant} \end{cases}$$

Adjusted R-Squared:

$$1 - \left(\frac{n-1}{n-p} \right) \left(1 - R^2 \right)$$

Standard Error of Estimate:

$$SE = \sqrt{MSE}$$

Predicted Values:

$$\underset{\sim}{\hat{Y}} = \underset{\sim}{X}\,\underset{\sim}{b}$$

Residuals:

$$\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}}$$

Durbin-Watson Statistic:

$$D = \frac{\sum_{i=1}^{n-1}\left(e_{i+1} - e_{i}\right)^{2}}{\sum_{i=1}^{n} e_{i}^{2}}$$

Mean Absolute Error:

$$\frac{\left(\sum_{i=1}^{n}\left|e_{i}\right|\right)}{n}$$

Predictions:

$\underset{\sim}{X}_h$ = m-by-p matrix of independent variables
for m predictions

Predicted Value:

$$\hat{\underset{\sim}{Y}}_h = \underset{\sim}{X}_h \underset{\sim}{b}$$

Standard Error of Prediction:

$$S(\hat{\underset{\sim}{Y}}_{h(new)}) = \sqrt{\text{diagonal elements of MSE } (1 + \underset{\sim}{X}_h(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'_h)}$$

Standard Error of Mean Response:

$$S(\hat{\underset{\sim}{Y}}_{h}) = \sqrt{\text{diagonal elements of MSE } (\underset{\sim}{X}_h(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'_h)}$$

Multiple-Sample Comparison

Sample Mean:

$$\overline{x}_t = \frac{\sum_{i=1}^{n_t} x_{it}}{n_t}$$

MSE = Mean Square Error:

$$\frac{\sum_{t=1}^{k}(n_t - 1)s_t^2}{\left(\sum_{t=1}^{k} n_t\right) - k}$$

where

$$s_t^2 = \sum_{i=1}^{n_t} \frac{(x_{it} - \overline{x}_t)^2}{n_t - 1}$$

Df = degrees of freedom for the error term:

$$n - k$$

Standard Error (internal):

$$\sqrt{\frac{s_t^2}{n_t}}$$

Standard Error (pooled):

$$\sqrt{\frac{\text{MSE}}{n_t}}$$

where

k = number of samples

n = number of observations

n_t = number of observations for sample t

x_{it} = ith observation for sample t

Kruskal-Wallis Test:

Assign ranks 1 to N to each observation in the combined sample, with the smallest value in the sample having Rank 1, and the largest value having Rank N.

Calculate the average rank of sample t as

$$R_t = \frac{\sum_{i=1}^{n_t} R_{it}}{n_t}$$

where

n_t = number of observations in sample t

R_{it} = rank of observation i in sample t, using the overall combined ranking

N = total number of observations

If there are no ties, the test statistic is

$$w = \left[\frac{12}{N(N+1)} \Sigma\, n_t\ R_t^2 \right] - 3(N+ 1)$$

where k is the number of samples.

If there are ties, let

u_j = the number of observations tied at any rank for
j = 1,2,3,...,m

where

m = the number of tied groups in the sample.

The test statistic corrected for ties is

$$W = \frac{w}{1 - \frac{\sum_{j=1}^{m} u_j^{\,3} - \sum_{j=1}^{m} u_j}{n(n^2 - 1)}}$$

Note that W reduces to w when there are no ties.

When $n_t > 5$ for all t or when there are more than three samples, W and w follow a chi-square distribution with k − 1 degrees of freedom.

Multiple-Variable Analysis

Average:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Geometric Mean:

$$\left(\prod_{I=1}^{n} x_i \right)^{\frac{1}{n}}$$

Variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} = 1\left(x_i - \overline{x} \right)^2$$

Standard Deviation:

$$s = \sqrt{s^2}$$

Standard Error:

$$\frac{s}{\sqrt{n}}$$

Lower Quartile:

$$i = \text{ceiling}\left[\frac{(n+1)}{4}\right]$$

$$j = \text{ceiling}\left[\frac{n}{4}\right]$$

$$\text{Lower Quartile} = \frac{(X_{[i]} + X_{[j]})}{2}$$

Upper Quartile:

$$i = \text{ceiling}\left[\frac{3(n+1)}{4}\right]$$

$$j = \text{ceiling}\left[\frac{3n}{4}\right]$$

$$\text{Upper Quartile} = \frac{(X_{[i]} + X_{[j]})}{2}$$

Skewness:

(missing if s=0 or n<3)

$$\frac{n\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3}{(n-1)(n-2)s^3}$$

Standardized Skewness:

~N(0,1) for n > 150

$$\frac{\text{skewness}}{\sqrt{\frac{6}{n}}}$$

Kurtosis:

(missing if s=0 or n<4)

$$\frac{n(n+1)\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Standard Kurtosis:

$$\frac{\text{Kurtosis}}{\sqrt{\frac{24}{n}}}$$

Coefficient of Variation:

$$\frac{s}{\overline{x}} \times 100$$

Sum:

$$\sum_{i=1}^{n} x_i$$

The sample correlation coefficient between x and y is where s_x is the sample standard deviation for x and s_y is the sample standard deviation for y.

$$r(x, y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{s_x s_y}$$

To determine the significance level, use the fact that

$$z = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right)$$

$$\mu_z = \frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right)$$

and variance

$$\sigma_z^2 = \frac{1}{n-3}$$

where ρ is the true (population) correction coefficient.

The sample covariance between x and y is

$$\mathrm{Cov}(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})$$

where

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

One-Variable Analysis

For information on the calculations for the density trace and quantile plot, see Chambers, et al. (1983).

n = number of nonmissing observations

x_i = value of the ith observation

Average:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Standard Deviation:

$$s = \sqrt{s^2}$$

Median:

The median is the middle value in a set of numbers arranged in order of magnitude.

100(1 - α)% Confidence Interval for Mean:

$$\overline{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

$$t_{n-1} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where

μ_0 is the hypothesized mean

100(1   α)%Confidence Interval for Standard Deviation:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{n\text{-}1,\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{n\text{-}1,1\text{-}\alpha/2}}} \right]$$

One-Way ANOVA

Treatment Mean:

$$\overline{x}_t = \frac{\sum_{i=1}^{n_t} x_{it}}{n_t}$$

MSE = Mean Square Error:

$$\frac{\sum_{t=1}^{k} (n_t - 1) s_t^2}{\left(\sum_{t=1}^{k} n_t \right) - k}$$

where

$$s_t^2 = \sum_{i=1}^{n_t} \frac{(x_{it} - \overline{x}_t)^2}{n_t - 1}$$

Df = degrees of freedom for the error term:

$$n - k$$

Standard Error (internal):

$$\sqrt{\frac{s_t^2}{n_t}}$$

Standard Error (pooled):

$$\sqrt{\frac{MSE}{n_t}}$$

where

k = number of treatments

n = number of observations

n_t = number of observations for treatment t

x_{it} = ith observation for treatment t

Kruskal-Wallis Test:

Assign ranks 1 to N to each observation in the combined sample, with the smallest value in the sample having Rank 1, and the largest value having Rank N.

Calculate the average rank of treatment t as

$$\overline{R}_t = \frac{\sum_{i=1}^{n_t} R_{it}}{n_t}$$

where

n_t = number of observations in treatment t

R_{it} = rank of observation i in treatment t, using the overall combined ranking

N = total number of oberservations

If there are no ties, the test statistic is

$$w = \left[\frac{12}{N(N+1)} \sum n_t \; R_t^2 \right] - 3(N+1)$$

where k is the number of treatments.

If there are ties, let

u_j = number of observations tied at any rank for
j = 1,2,3,...,m

where

m = number of tied groups in the sample.

The test statistic corrected for ties is:

$$W = \frac{w}{1 - \frac{\sum_{j=1}^{m} u_j^{\,3} - \sum_{j=1}^{m} u_j}{n(n^2 - 1)}}$$

Note that W reduces to w when there are no ties.

When $n_t > 5$ for all t or when there are more than three treatments, W and w follow a chi-square distribution with k − 1 degrees of freedom.

Outlier Identification

Notation:

n = number of samples

x_i = observed value for sample i, i=1,2,...,n

μ = population mean

σ = population standard deviation

$\bar{x}$ = sample mean

$\tilde{x}$ = sample median

IQR = sample interquartile range

s = sample standard deviation

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Studentized values without deletion - using the sample mean and standard deviation, each data value is standardized by

$$t_i = \frac{x_i - \bar{x}}{s}$$

Studentized values with deletion - each data value is removed from the sample one at a time and the mean $\bar{x}_{(i)}$ and standard deviation $s_{(i)}$ are calcuated using the remaining n-1 data values. Each data value is then standardized by

$$t_i = \frac{\bar{x}_i - x_{(i)}}{s_{(i)}}$$

Modified MAD Z-score - using the sample median and median absolute deviation defined by

$$MAD = median_i\left\{\left[x_i - \tilde{x}\right]\right\}$$

each data value is standardized by

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

Grubbs' Test:

Grubbs Test is calculated if n 3. Also called the Extreme Studentized Deviate Test (ESD), it is based on the largest Studentized value (without deletion) t_{max}. The test statistic T is computed according to

$$T = \sqrt{\frac{n(n-2)t_{max}^2}{(n-1)^2 - nt_{max}^2}}$$

An approximate two-sided P-value is obtained by comparing *T* to Student's t-distribution with n-2 degrees of freedom. A small P-value (as in the table above) leads to the conclusion that the most extreme point is indeed an outlier. For small samples, one can refer instead to Iglewicz and Hoaglin

(1993) who give 5% and 1% values for t_{max} in Appendix A of their monograph, as well as for a generalized test involving r > 1 potential outliers.

Dixon's Test:

Dixon's test is performed if 4 n 30. This test begins by ordering the data values from smallest to largest. Letting x(j) denote the j-th smallest value, statistics are then computed to test for 5 potential situations:

Situation 1: **1 outlier on the right**. Compute:

$$r = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$$

Situation 2: **1 outlier on the left**. Compute:

$$r = \frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$$

Situation 3: **2 outliers on the right**. Compute:

$$r = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}}$$

Situation 4: **2 outliers on the left**. Compute:

$$r = \frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$$

Situation 5: **1 outlier on either side**. Compute:

$$r = \max\left[\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}, \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}\right]$$

The calculated statistic *r* is then compared to critical values in tables such as Appendix A.3 of Iglewicz and Hoaglin (1993). For each test, STATGRAPHICS indicates whether or not the result is statistically significant at the 5% level and at the 1% level. A significant result indicates the presence of the situation indicated.

Resistant Estimators for the Mean:

To estimate the mean, STATGRAPHICS calculates:

1. the **sample mean** $\bar{x}$

2. the **sample median** $\tilde{x}$

3. a 100α% **trimmed mean** defined by

$$T(\alpha) = \frac{1}{n(1-2\alpha)}\left[k\left(x_{(r+1)} + x_{(n-r)}\right) + \sum_{i=r+2}^{n-r-1} x_{(i)}\right]$$

where $r = [\alpha n]$ and $k = 1-(\alpha n - r)$. This is basically the mean of the sample after removing a fraction each of the smallest and largest data values. By default, STATGRAPHICS trims 15% from each end, although that value may be changed using *Pane Options*.

4. a **Winsorized mean** in which copies of $x_{(r+1)}$ and $x_{(n-r)}$ replace the data values which would be trimmed away by a trimmed mean:

$$T_W = \frac{1}{n}\left\{\sum_{i=r+1}^{n-r} x_{(i)} + r\left[x_{(r+1)} + x_{(n-r)}\right]\right\}$$

Note that the latter three statistics, which are less sensitive to the one unusually low value, all suggest that the true population mean in the current example is higher than that estimated by $\bar{x}$.

Resistant Estimators for Sigma:

To estimate the standard deviation, STATGRAPHICS computes:

1. the **sample standard deviation** s.

2. an estimate based on the **mean absolute deviation** (MAD) discussed earlier:

$$\hat{\sigma} = \frac{MAD}{0.6745} = \frac{\text{median}\left\{\left|x_i - \bar{x}\right|\right\}}{0.6745}$$

3. an estimate based on a **weighted sum of squares around the sample median**:

$$S_{bi} = \frac{\sqrt{n\sum_{i=1}^{n}(x_i - \tilde{x})2(1-u_i^2)4}}{\sum_{i=1}^{n}(1-u_i^2)(1-5u_i^2)}$$

where

$$u_i = \frac{x_i - \tilde{x}}{9MAD}$$

4. a **Winsorized sigma** defined by:

$$S_W = \sqrt{\frac{n\left\{\sum_{i=r+1}^{n-r}(x_{(i)} - T_W)^2 + r[(x_{(r+1)} - T_W)^2 + (x_{(n-r)} - T_W)^2]\right\}}{(n-2r)(n-2r-1)}}$$

Confidence Intervals:

The resistant estimators can also be used to obtain a confidence interval for the population mean μ. Recall that the usual formula involving the sample mean and standard deviation is

$$x \pm t_{n-1,1-\alpha/2}\frac{2}{\sqrt{n}}$$

where $t_{n-1,1-\alpha/2}$ is the value of Student's t distribution with n-1 degrees of freedom which is exceeded with probability /2. STATGRAPHICS displays the above interval together with an interval based on the Winsorized mean and standard deviation:

$$T_W \pm t_{n-2r-1,1-\alpha/2}\frac{S_W}{\sqrt{n}}$$

Power Transformations

The Power Transformations transform the values of a single numeric variable according to:

$$w = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda 1\text{-}1}}{\lambda_1 g_m^{\lambda 1\text{-}1}} & \text{, if } \lambda_1 \neq 0 \\ g_m \ln(y+\lambda_2) & \text{, if } \lambda_1 = 0 \end{cases}$$

where g_m is the geometric mean of $(y + \lambda_2)$.

λ_1 is estimated from the data so as to minimize the sample variance.

λ_2 is specified by the user.

Probability Plot

The program first sorts the input data from the smallest value to the largest value to compute order statistics. It then generates a scatterplot where

horizontal position $= X_{(i)}$

vertical position $= \varphi\left(\dfrac{i - 3/8}{n + 1/4}\right)$

where

$\varphi(p) =$ value of the inverse cumulative standard normal distribution at probability p.

The labels for the vertical axis are based on the probability scale using:

$$100\left(\frac{i\text{-}\ 3/8}{n + 1/4}\right)$$

Sample Size

Normal Mean:

Absolute error = Δ

$$n = \left(\frac{t_{n-1,\ \alpha/2}\,\sigma}{\Delta}\right)^2$$

Relative error = r (proportion)

$$n = \left(\frac{t_{n-1, \alpha/2} \sigma}{r\mu} \right)^2 \text{ if } \mu \neq 0$$

Power

$$n = \left(\frac{\left(t_{n-1, \alpha/2} + t_{n-1, \beta} \right) \sigma}{\Delta} \right)^2$$

Normal Sigma:

Relative error = r

Find n such that

$$G_{n-1}\left((1+r)^2 (n-1) \right) - G_{n-1}\left((1-r)^2 (n-1) \right) \geq 1 - \alpha$$

where

$G_{n-1}(\,.\,)$ is the cumulative χ^2_{n-1} distribution.

Absolute error Δ

$$r = \frac{\Delta}{\sigma}$$

$$G_{n-1}\left((1+r)^2 (n-1) \right) - G_{n-1}\left((1-r)^2 (n-1) \right) \geq 1 - \alpha$$

where

$G_{n-1}(\,.\,)$ is the cumulative χ^2_{n-1} distribution.

Power

$$n = 1 + \frac{1}{2}\left(\frac{\lambda\left(Z_{\alpha/2} + Z_{\beta}\right)}{\lambda - 1}\right)^{2} + \frac{1}{2}$$

where

$$\lambda = \frac{\sigma}{\sigma + \Delta}$$

Binomial:

Absolute error = Δ

$$n = \left(Z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{\Delta}}\right)^{2}$$

Relative error = r

$$n = \left(Z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{r\theta}}\right)^{2}$$

Power

$$n = \left(\frac{\left(Z_{\alpha/2} + Z_{\beta}\right)}{2(\operatorname{asin}\sqrt{\theta + \Delta} - \operatorname{asin}\sqrt{\theta}}\right)^{2}$$

Poisson:

Absolute error = Δ

$$n = \left(\frac{Z_{\alpha/2}\sqrt{\lambda}}{\Delta}\right)^{2}$$

Relative error = r

$$n = \left(\frac{Z_{\alpha/2}\sqrt{\lambda}}{r\lambda} \right)^2$$

Power

$$n = \left(\frac{Z_{\alpha/2} + Z_{\beta}\sqrt{\lambda}}{\Delta} \right)^2$$

Scatterplot Smoothing

Locally Weighted Regression (lowess):

The program fits a straight line to the k data points closest to X using weighted least squares; the data closest to X gets the most weight.

Once the fitted values

$$\hat{Y}_i$$

have been determined, the residuals,

$$r_i = (y_i - \hat{y}_i)$$

are used to determine new weights (Robust Locally Weighted). The critical component in lowess is the choice of smoothing parameters; that is, the number of points used in the calculation. The more points selected, the smoother the fitted curve.

Two-Sample Comparison

For information on the calculations for the density traces and quantile plots, see Chambers, et al. (1983).

n_1 = number of nonmissing observations in sample 1

n_2 = number of nonmissing observations in sample 2

x_{ij} = value of the ith observation in the jth sample

Average:

$$\bar{x}_j = \frac{1}{nj}\sum_{i=1}^{n} x_{ij} \qquad j = 1,2$$

Variance:

$$s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 \qquad j = 1,2$$

Standard Deviation:

$$s_j = \sqrt{s_j^2} \qquad j = 1,2$$

Standard Error:

$$s_{\bar{x}j} = \frac{s_j}{\sqrt{n_j}} \qquad j = 1,2$$

Median:

The median is the value to which half the values in the sample exceed and half do not. When there are an even number of samples, the median is the average of the value to which half the values exceed and the value to which half do not.

Pooled Standard Deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where

$$s_1^2 \;=\; \text{the variance of sample 1}$$
$$s_2^2 \;=\; \text{the variance of sample 2}$$

100(1 - α)% Confidence Interval for Difference in Means, Equal Variance:

$$(\overline{x}_1 - \overline{x}_2\ ) \pm t_{n_1+n_2-2;a/2}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

100(1 – α)% Confidence Interval for Difference in Means, Unequal Variance:

$$\left[(\overline{x}_1 - \overline{x}_2\ ) \pm t_{m;a/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

where

m = degrees of freedom

and where

$$m \;=\; \frac{\left(\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1-1}\left(\dfrac{s_1^2}{n_1}\right)^2+\dfrac{1}{n_2-1}\left(\dfrac{s_2^2}{n_2}\right)^2}$$

100(1-α)% Confidence Interval for Ratio of Variances:

$$\left(\frac{s_1^2}{s_2^2}\right)\left(\frac{1}{F_{n_1-1,\,n_2-1;a/2}}\right),\left(\frac{s_1^2}{s_2^2}\right)\left(\frac{1}{F_{n_1-1,\,n_2-1;a/2}}\right)$$

t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$\bar{x}_1$ = mean of sample 1

$\bar{x}_2$ = mean of sample 2

Δ = hypothesized difference between the means

Weibull Analysis

Notation:

n = sample size (>0)
r = number of failures (>0)
$x_{(i)}$ = ordered data values (failure times and censored times)
α = shape parameter
β = scale parameter
Δ = origin of Weibull distribution
$f(x)$ = Weibull density function
$F(x)$ = cumulative Weibull density function

Rank Regression Estimation

Determines the value of α and β by regressing $\ln(x_{(i)} - \Delta)$ against specific estimates of the reliability R_i. The reliabilities or plotting positions may be defined in any of four different ways:

Median Ranks
Uses an approximation to the median plotting position as calculated from the medians of the order statistics from a uniform distribution on (0,1):

$$R_i = \frac{j_i - .3}{(n + .4)}$$

where j_i is a failure order number defined by $j_i = j_{i-1} + \delta$ and $j_0 = 0$.

δ is initially set to 1 and modified whenever R_i corresponds to a censoring time according to:

$$\delta = \frac{(n + 1) - \text{prior failure order number}}{1 + \text{number of times before prior censored time}}$$

Expected Ranks

Uses the expected values of the order statistics. This computes reliabilities according to the recursive relation:

$$R_i = \left[\frac{r_i}{r_i + 1}\right] R_{i-1}$$

where r_i equals the reverse ranks of the data values and $R_0 = 1$.

Kaplan-Meier

Uses the Kaplan-Meier method for computing the empirical cumulative distribution function. This option computes reliabilities according to:

$$R_i = \left[\frac{r_i - 1}{r_i}\right] R_{i-1}$$

Modified Kaplan-Meier

Uses the modified Kaplan-Meier method in which reliabilities are computed from the Kaplan-Meier reliabilities according to:

$$R'_i = \frac{(R_i + R_i - 1)}{2}$$

Maximum Likelihood Estimation

Determines the values of α and β by maximizing the likelihood function of the data. The likelihood function is given by the product of $f(x_{(i)})$ for all failure times and $(1 - F(x_{(i)}))$ for all censored times.

Weibayes

Assuming that the shape parameter α is specified by the user, the maximum likelihood estimator of the scale parameter β is given by

$$\hat{\beta} = \Sigma \left(\frac{x_i^{\alpha}}{r}\right)^{1/\alpha}$$

Confidence Bands for Percentiles

The Confidence Inteval is given by the set of values Q_0 for which

$$\Lambda \leq \chi^2_{(1),\gamma}$$

where

$$\Lambda = -2 \log L(\tilde{\alpha}, \tilde{\beta}) + 2 \log L(\hat{\alpha}, \hat{\beta})$$

and $\tilde{\alpha}$ is the solution to

$$\frac{r}{\alpha} - r \log Q_0 + \sum_{i \in D} \log(x_i - \Delta) + \log(1 - p) \sum_{i=1}^{n} \left(\frac{(x_i - \Delta)}{Q_0} \right)^{\alpha} \log \left(\frac{(x_i - \Delta)}{Q_0} \right) = 0$$

and

$$\tilde{\beta} = \frac{Q_0}{[-\log(1-p)]^{1/\alpha}}$$

Note: $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimators of the Weibull parameters. D represents the set of all observed failure times.

F Glossary

- **Adjusted R-Squared**
  A statistic that is suitable for comparing models that have different numbers of independent variables; indicates the percentage of variability for which the model accounts.

- **Analysis of Means (ANOM) Plot**
  A graphical display comparing sample means with the grand mean and the user-defined decision limits.

- **ANOVA Table**
  A standard analysis of variance table that presents the values for the variability between groups and within groups.

- **Arrhenius Plot**
  A graphical display fitting an Arrhenius model to a set of percentiles taken at two or more study temperatures. The fitted model can then be used to predict the corresponding percentile at a normal temperature.

- **Barchart**
  A graphical display of a frequency table where each bar represents an individual row or cell in the table.

- **Baseline**
  The starting point for the bars on a Barchart.

- **Bernoulli Distribution**
  A distribution whose outcome has only two possibilities: success or failure.

- **Beta Distribution**
  A distribution that is useful for random variables constrained to lie between 0 and 1; characterized by two parameters: Shape 1 and Shape 2.

- **Binomial Distribution**
  A distribution that gives the probability of observing successes in a fixed number of independent or Bernoulli trials.

- **Bonferroni Method**
  A method for determining statistically significant differences between means when the F-ratio is not significant, but a small number of comparisons are planned.

- **Bootstrap Intervals**
  A method to determine bounds for the mean, standard deviation, and median. Select interval type and bootstrap options using the Pane Options dialog box. The intervals are formed by selecting k random subsamples of n observations each with replacement from the n data values, calculating each statistic, and displaying percentiles for the k values of the computed statistics.

- **Box-Cox Transformations**
  An analysis that automatically identifies a transformation from a family of power transformations. The analysis determines the appropriate transformation parameter, Lambda1, which minimizes the mean squared error (MSE) or the fitted model.

- **Boxcar Method**
  A method that assigns a constant density to each small interval around a value, which smooths the density function into a series of overlapping rectangles or boxcars, similar to a simple moving average.

- **Box-and-Whisker Plot**
  A graphical summary of the presence of outliers in data for a single variable.

- **Brushing**
  A technique used to change the color of points on a Scatterplot to indicate that an additional variable has been added to the analysis.

- **Bubble Chart Analysis**
  An analysis that creates an X-Y Scatterplot consisting of circles, where the size of the circles is determined by a third variable. The plot is helpful when you want to view data in higher dimensions.

- **Casement Plot**
  A plot of the X and Y variables in groups determined by the Z variable.

- **Cauchy Distribution**
  A distribution that fits data that follow a Cauchy distribution. The distribution's probability density function has no mean and an infinite variance. It is characterized by two parameters: Mode and Scale. Data fit to this distribution should be continuous data with a Mode between -infinity and +infinity and a Scale greater than 0.

- **CDF (Cumulative Distribution Function)**
  A function that provides cumulative probabilities as an alternative way for describing the pdf of a random variable. The height of the function indicates the probability of obtaining a value less than or equal to a specified value.

- **Censored Data**
  Data that are incomplete in some way; for example, having certain values that are unknown, or having a situation in which an event does not occur during the period of observation.

- **Censored Data Analysis**
  An analysis that allows you to create a relevant summary of the data by fitting a probability distribution function to determine if the data follow a normal or another type of distribution.

- **Chi-Square Distribution**
  A distribution that is useful for random variables constrained to be greater than 0; characterized by one parameter: Degrees of Freedom.

- **Chi-Square Test for Independence**
  A method that performs a hypothesis test to determine whether or not to reject the notion that two variables are independent.

- **Cochrane-Orcutt Transformation**
  This procedure handles situations in which the residuals are autocorrelated, which frequently occurs when the data is collected over time.

- **Coded X-Y Scatterplot**
  A scatterplot that uses coded points to show the levels for a classification factor.

- **Comments Window**
  A window you use to record information about a StatFolio.

- **Comparison of Alternative Models**
  A table of the correlation coefficients and R-Squared values for several transformed model types.

- **Comparison of Means**
  An analysis whose results display the values for the 95 percent confidence intervals for the mean of two variables and for the difference between the means, assuming equal variances and not assuming equal variances.

- **Comparison of Medians**
  The results of a Mann-Whitney (Wilcoxon) test (a Rank test).

- **Comparison of Proportions**
  An analysis that tests the hypothesis that the mean proportions of the samples are all identical. It generates an Analysis of Means (ANOM) Plot to determine which samples are significantly different from the grand mean.

- **Comparison of Rates**
  An analysis that tests the hypothesis that the mean rates of the samples are all identical. It generates an Analysis of Means (ANOM) Plot to determine which samples are significantly different from the grand mean.

- **Comparison of Standard Deviations**
  A display of the results from an F-test, which compares the ratio of the variance for the first population with the variance of the second population.

- **Component Effects Plot**
  A plot of the residuals that is helpful in judging the relative magnitude of the residuals with respect to the explanatory power of the variable.

- **Component Line Chart**
  A graphical display of several data values versus time in a cumulative or noncumulative plot.

- **Contingency Coefficient**
  A version of a Chi-Square test that measures the strength of the dependency between two characteristics.

- **Contingency Tables Analysis**
  An analysis for determining if two classification factors are related and, if so, how closely.

- **Contour Plot**
  A two-dimensional plot of a response surface where the contours of the plot represent the height of the surface; useful for determining optimal regions.

- **Correlation Matrix**
  A table of the Pearson correlation coefficients for the estimated coefficients in a model.

- **Correlations**
  A measurement of the strength of the linear relationship between random variables.

- **Cosine Method**
  A method for smoothing local densities into a series of overlapping, bell-shaped intervals similar to a weighted moving average, which results in a smoother density trace.

- **Covariances**
  A matrix of covariance measures for a set of observed values. The covariance measures the linear association between two variables.

- **Cramer's V and Conditional Gamma**
  Chi-square based measures of association that are helpful in determining the relationship between variables.

- **Critical Values**
  The smallest values for the area that falls under the distribution curve; the values are no less than the values you select for a probability.

- **Crosstabulation Analysis**
  An analysis that summarizes the joint distribution of two discrete variables through a crosstabulation. The crosstabulation counts the number of times each unique value occurs in the first variable, then in the second.

- **Cumulative Hazard Function Plot**
  A plot of the estimated cumulative hazard function that is helpful when you need suggestions for possible parametric models when you use failure data.

- **Custom Chart Analysis**
  An analysis that allows you to create customized charts.

- **Death Density Function Plot**
  A plot of the estimated death density function that corresponds to the probability distribution of item lifetimes.

- **Density Function**
  A function that produces a plot that gives the rate at which the cumulative probability increases at a given point.

- **Density/Mass Function Plot**
  A plot of the probability density function for a distribution.

- **Density Trace Plot**
  A plot of the shape or distribution of the data, especially the variations in density over the range of the data.

- **Discrete Uniform Distribution**
  A distribution that allocates equal probabilities to all integer values between a lower and an upper limit.

- **Distribution Fitting Analysis**
  An analysis that determines if data follow a normal or another type of distribution. The analysis fits one of several distribution functions.

- **Distribution Functions 1 and 2**
  Plots that display one of several types of distribution functions: Density Function, CDF, Survivor Function, Log Survivor Function, and Hazard Function.

- **Distribution-Free Tolerance Limits**
  A method to create and display nonparametric tolerance limits, which should be used when data are non-normal. See Normal Tolerance Limits.

- **Dot Diagram Analysis**
  An analysis that provides an additional type of graphical display of numeric data. The dots on the plot represent every value in a numeric column.

- **Double Reciprocal Model**
  A model that fits the reciprocal of the dependent variable and the reciprocal of the independent variable.

- **Draftsman's Plot**
  A series of two-variable plots for all combinations of three variables.

- **Duncan Method**
  A method for determining statistically significant differences among means if you want to protect the results against Type I errors for a small number of comparisons.

- **Durbin-Watson Statistic**
  A measure of the serial correlation of the residuals.

- **Dynamic Data Exchange (DDE)**
  A channel through which applications pass information between each other.

- **Erlang Distribution**
  A distribution useful for random variables constrained to be greater than 0; characterized by the Shape and Scale parameters.

- **Eta**
  An approximate measure of association for two variables when you measure the dependent variable on an interval scale and the independent variable on a nominal or ordinal scale.

- **Exclude Function**
  A function that can be used to exclude a single row in the selection fields on the data input dialog boxes.

- **Exponential Distribution**
  A continuous distribution useful for characterizing random variables that can take only positive values; completely determined by its mean.

- **Extreme Value Distribution**
  A distribution useful for random variables constrained to be greater than 0; characterized by two parameters: Mode and Scale. A distribution for fitting the limiting distribution for the maximum number of samples collected from a process. This distribution applies when extremes rather than means are collected from samples from an unknown or complex underlying distribution.

- **F (Variance Ratio) Distribution**
  A distribution useful for random variables constrained to be greater than 0; characterized by two parameters: Numerator Degrees of Freedom and Denominator Degrees of Freedom.

- **Factor Means Plot**
  An analysis that creates a matrix plot for data from designed experiments. It consists of main effects plots on the diagonal and interaction plots off the diagonal.

- **Frequency Histogram**
  A graphical display of the values in a sample created by dividing the range of the data into nonoverlapping intervals and counting the number of values that fall into each interval. The counts are called frequencies.

- **Frequency Table**
  A table of the number of observations that fall within each class and the relative, cumulative, and cumulative relative frequencies.

- **Frequency Tabulation**
  A table created by grouping data into a default number of class intervals of frequencies and cumulative frequencies that summarize the distribution of the data.

- **Gamma Distribution**
  A distribution useful for random variables constrained to be greater than 0; characterized by the Shape and Scale parameters.

- **Geometric Distribution**
  A distribution that characterizes the number of failures before the first success in a series of Bernoulli trials; a special case of Negative Binomial distributions, where $k = 1$.

- **Goodness-of-Fit Tests**
  A standardized chi-square statistic that compares observed frequencies with the expected frequencies predicted by the fitted distribution. The program performs this analysis along with the Kolmogorov-Smirnov one-sample test.

- **Group Comparisons**
  A table that shows the results of comparing groups in a study using Logrank and Wilcoxon tests.

- **Hazard Function**
  A function that produces a plot that shows the probability of failure during a small increment of time when no failure occurs before a time you select. The hazard function is equal to the probability density function divided by the survivor function. When you model lifetime data, the function represents the instantaneous failure rate.

- **High-Low-Close Plot**
  A graphical display of market data that fluctuate from hour-to-hour, day-to-day, or week-to-week. A vertical line extends from each low value to its corresponding high value. A horizontal line marks the final value for the period.

- **Histogram**
  A plot of data shown as a series of successive bars that represent classes. The width of the bars is equal to the class interval.

- **Hypergeometric Distribution**
  A distribution that arises when a random selection is made between objects of two distinct types (success, fail). The sampling occurs without replacement; this is, each time an item is drawn and studied, it is not placed back into the population. The distribution gives the probability for the number of successes. Data fit to this distribution should be integers greater than or equal to 0.

- **Hypothesis Tests (Compare) Analysis**
  An analysis that estimates and tests hypotheses about the sample mean, sample variance, the binomial proportion, or the rate for a single random variable.

- **Hypothesis Tests (Describe) Analysis**
  An analysis that tests hypotheses about the sample means and variances for two or more random samples.

- **Hypothesis Tests**
  Tabular results of the *t*-statistic, its *p*-value, and the recommendation to either accept or not accept the null hypothesis.

- **Influential Points**
  A table of the observations that have leverage values above three times the average leverage value; the row numbers of the observations flagged as high leverage or DFITS.

- **Interaction Plot**
  A plot of two-factor interactions.

- **Interquartile Range**
  A range in which the middle 50 percent of the data will fall.

- **Interval Plot**
  A display of the dependent variable plotted against the predicted values, their row numbers, or any of the independent variables.

- **Kendall's Tau**
  An option (in addition to Spearman's R) added to the Rank Correlations tabular option. The correlation coefficients range between -1 and +1, and measure the strength of the associations between the variables.

- **Kendall's Tau b and c**
  Statistics that measure the relative degree of agreement or disagreement between two variables.

- **Key Glyph**
  A plot of a single glyph with individual rays labeled with the names of the corresponding variables.

- **Kolmogorov-Smirnov Test**
  A method that tests the null hypothesis that two samples can reasonably be assumed to come from the same distribution.

- **Kruskal-Wallis and Friedman Tests**
  Previously called Kruskal-Wallis Test. A nonparametric method that tests the assumption that the medians of samples are equal. Perform a Friedman Test for balanced data in which each row with any data has data for all of the columns. The hypothesis tested, that of equal means in each column, is the same for both tests. However, the Friedman test is only meaningful if the data is blocked by row.

- **Kurtosis**
  A value used to measure the flatness or steepness of a distribution with respect to a Gaussian or Normal distribution.

- **Lack-of-Fit Test**
  A test that determines if the regression model adequately fits the data.

- **Lambda Values**
  Values that measure the degree of association on a scale from 0 to 1. The closer Lambda is to 1, the more useful one factor is in predicting the other factor.

- **Laplace Distribution**
  A distribution useful for random variables from a distribution that is more peaked than normal; characterized by two parameters: Mean and Scale.

- **Levene's Test**
  Compares the sample variances by performing an analysis of variance on the absolute deviations of the data values from their respective sample means. It is less sensitive than Bartlett's test to departures from normality of the underlying populations.

- **Life Tables (Intervals) Analysis**
  An analysis that creates life tables from the counts of failures in a set of intervals.

- **Life Tables (Times) Analysis**
  An analysis that creates life tables from the failure times of items, which provides an estimate of a survival function.

- **Linear Model**
  A model that fits a simple linear regression equation for Y in terms of X.

- **Locally Weighted Regression**
  An option that smooths data by fitting a straight line to points using weighted least squares.

- **Logistic Distribution**
  A distribution useful for random variables that are not constrained to be greater than or equal to 0; characterized by two parameters: Mean and Standard Deviation.

- **Lognormal Distribution**
  A distribution useful for random variables constrained to be greater than 0; characterized by the Mean and Scale parameters.

- **Log Survival Function Plot**
  A plot that shows the estimated log survival function. The plot shows the probability of observing a value greater than a specified value (in log units).

- **Log Survivor Function Plot**
  A plot that shows the probability of observing a value greater than a value you select (in log units).

- **Lower Quartile**
  The value below which 25 percent of the data will fall.

- **LSD (Least Significant Differences) Method**
  A method for testing the statistically significant differences between means when the F-ratio is significant and comparisons are planned.

- **Matrix Plot**
  A plot that displays three or more numeric variables.

- **Matrix Plot Analysis**
  An analysis that creates a matrix of two-variable plots for a set of numeric variables, with box-and-whisker plots on the diagonal.

- **Maximum Likelihood**
  A method used to calculate an estimation. It first calculates the probability that the sample statistic (observed value) would occur if it were the true value of the parameter, then chooses the parameter that has the greatest probability of being the actual observation.

- **Mean Absolute Error**
  The average of the absolute values of the residuals.

- **Mean Marker**
  A symbol that denotes the location of the average on a Box-and-Whisker Plot.

- **Means Plot**
  A plot of the mean for each sample and the intervals for the means.

- **Means Table**
  A table of the sample means and standard errors for the levels in the data.

- **Median Notch**
  A notch in the box around the median on a Box-and-Whisker Plot. The length of the notch represents an approximate confidence interval for the median.

- **Median Plot**
  A plot of the sample medians for the levels in the variables.

- **Mode**
  The value with the highest frequency.

- **Mosaic Plot**
  A plot that represents the frequencies in the cells of a Crosstabulation.

- **Multifactor ANOVA Analysis**
  An analysis that examines the effects of two or more factors on one variable; typically uses data collected for a designed experiment.

- **Multiple Bar Chart Analysis**
  An analysis that creates a plot of one or more frequency bars for each classification factor (row). The plot graphically represents a secondary classification factor within a primary classification factor.

- **Multiple Box-and-Whisker Plot**
  A graphical summary of the outliers in data for two or more variables.

- **Multiple Dot Diagram**
  Creates a dot diagram for data that are divided into more than one group and displays summary statistics and confidence intervals for each group.

- **Multiple Range Tests**
  A table that shows the number of observations and the mean for each sample, the variables in homogeneous groups that are not significantly different, the average values for the difference and plus/minus limits for each pair of samples and the size of the interval about this difference.

- **Multiple Regression Analysis**
  An analysis that analyzes the relationship among one dependent variable and one or more independent variables.

- **Multiple-Sample Comparison Analysis**
  An analysis used to compare several sets of data, where the observations in the samples are assumed to be independent.

- **Multiple-Variable Analysis**
  An analysis used to analyze the relationship among several numeric variables.

- **Multiple X-Y Plot**
  A scatterplot with one variable on the X-axis and one or more variables on the Y-axis.

- **Multiple X-Y-Z Plot**
  A scatterplot for three or more variables. One variable each is plotted on the X- and Y-axes; one or more variables on the Z-axis.

- **Multiplicative Model**
  A model used to fit a simple regression that is equivalent to taking the natural logs of Y and X.

- **Negative Binomial Distribution**
  A distribution that characterizes the number of failures before the *k*th success in a series of Bernoulli trials.

- **Newman-Keuls Method**
  A method for determining statistically significant differences among means if the F-ratio is significant and you want to test multiple hypotheses.

- **Normal Distribution**
  A continuous probability distribution that is useful in characterizing a large variety and type of data. It is a symmetric, bell-shaped distribution completely determined by its Mean and Standard Deviation.

- **Normal Probability Plot**
  A plot that consists of an arithmetic (interval) horizontal axis scaled for the data and a vertical axis scaled so the cumulative distribution function of a Normal distribution plots as a straight line.

- **Normal Tolerance Limits**
  A method for creating and displaying normal tolerance limits, which you use with normally distributed data. See Distribution-Free Tolerance Limits.

- **Object Linking and Embedding (OLE)**
  A set of program interfaces that enable you to combine objects that are supported by different applications.

- **One-Variable Analysis**
  An analysis of the data in a numeric variable.

- **One-Way ANOVA Analysis**
  An analysis for determining the effect of one qualitative factor on one response variable.

- **Open Database Connectivity (ODBC)**
  A database-independent interface from Microsoft Corporation that allows you to access data in a variety of SQL-compliant database servers.

- **Outliers**
  Points more than 3 interquartile ranges below the first quartile or above the third quartile.

- **Outlier Identification (Numeric Data) Analysis**
  An analysis that identifies unusual data values (outliers) in a sample of numeric values using Studentized values, modified Z-scores, Grubb’s test, and Dixon’s test.

- **Paired-Sample Comparison Analysis**
  An analysis that calculates the difference between each value in each pair of observations.

- **Pareto Distribution**
  A distribution with a decreasing density function. One parameter, Shape, is necessary to specify the distribution.

- **Partial Correlations**
  A matrix of estimated partial correlation coefficients for a set of observed values. A partial correlation coefficient measures the relationship between two variables while controlling for possible effects of the other variables.

- **Percentiles**
  Tabular results that show the percentage of the values that are equal to or less than a value you select.

- **Piechart**
  A circle-shaped chart segmented into shaded or colored wedges. The chart is useful for displaying percentage breakdowns for up to 20 classification factors.

- **Plot of Fitted Model**
  A plot of the curve or fitted line that includes the confidence limits for the means and the prediction limits.

- **Poisson Distribution**
  A distribution that expresses probabilities that concern the number of events per unit time.

- **Polar Coordinates Plot**
  A two-dimensional scatterplot or line plot for pairs of points that are defined by radius and angle positions. The values for the radius variable are plotted against the values for the angle variable.

- **Polynomial Regression Analysis**
  An analysis that calculates a model between a single dependent variable Y and a single independent variable X.

- **Power Curve**
  A plot that illustrates the probability of the null hypothesis being rejected.

- **Power Transformations**
  An analysis designed to compare the effect of various power transformations on the normality of the distribution.

- **Probability Distributions**
  Functions that allow you to perform three basic operations: (1) calculate probabilities, (2) create plots of the probability and cumulative distributions, (3) generate random numbers. STATGRAPHICS *Plus* contains 22 distributions. See Bernoulli, Binomial, Discrete Uniform, Geometric, Negative Binomial, Poisson, Beta, Chi-Square, Erlang, Exponential, Extreme Value, F (Variance Ratio), Gamma, Laplace, Logistic, Lognormal, Normal, Pareto, Student's *t*, Triangular, Uniform, and Weibull.

- **Quantile Plot**
  A plot that displays the quantiles for the data. The Y-axis represents the proportion of values that are below a particular value.

- **Quantile/Quantile Plot**
  A plot of the quantile values plotted as pairs and useful for comparing the cumulative distributions for two samples.

- **R-Squared**
  A statistic that measures the proportion of variability in a model for the dependent variable.

- **Radar/Spider Plot**
  A graphical display that shows the data in each row, with the size of each of the variables plotted along one of the spokes.

- **Range Plot**
  A plot of the sample ranges for the levels in the variables.

- **Rank Correlations**
  A matrix of Spearman rank correlation coefficients similar to a correlation coefficients matrix.

- **Rank Regression**
  A method used to calculate an estimation. The best-fit line is drawn through the data plotted on the logarithmic (Weibull) grid to determine two parameters: Shape and Scale.

- **Rank Test**
  A test that combines the data for two samples, sorts the data from smallest to largest, and compares the average ranks for the two samples.

- **Recode Data**
  A method for changing the values in a spreadsheet column.

- **Residuals versus Factor Levels Plot**
  A plot of the residuals versus the factor levels.

- **Residuals versus Observation Plot**
  A plot of the residuals versus the observations.

- **Residuals versus Predicted Plot**
  A plot of the residuals versus the predicted values.

- **Residuals versus Row Number Plot**
  A plot of the residuals versus the row numbers.

- **Residuals versus Samples Plot**
  A plot of the residuals versus the samples.

- **Residuals versus *X* Plot**
  A plot of the residuals versus the independent variable.

- **Response Surfaces Analysis**
  An analysis that creates Surface and Contour plots for functions. The plots show the shape of a response surface when it is necessary to locate optimal settings.

- **Robust Lowess**
  An option that smooths data by repeating the smoothing process a second time, giving reduced weight to points that are the furthest from the first smoothed line.

- **Row-Wise Statistics Analysis**
  An analysis that provides the capability to calculate and save statistics on a row-wise basis for data that are in two or more columns.

- **Running Lines**
  An option that smooths data by fitting a straight line to points.

- **Running Means**
  An option that smooths data using a running average or a running mean.

- **Sample Size Determination (Compare) Analysis**
  An analysis that helps to determine the number of observations required to provide sufficiently powerful estimates, and to analyze the power curve for samples already drawn. The analysis calculates sample sizes for testing five types of parameters.

- **Sample Size Determination (Describe) Analysis**
  An analysis that calculates the sampling sizes for four parameters: Normal Mean, Normal Sigma, Binomial Proportion, and Poisson Rate.

- **Scatterplot**
  A plot of the values as point symbols along a single horizontal axis.

- **Scatterplot Matrix**
  A matrix plot of the variables with points for the data; visually presents the relationships among the pairs of variables.

- **Scheffe Method**
  A method for determining statistically significant differences among means when the F-ratio is significant and you want to make unplanned comparisons between pairs of means.

- **Sigma Plot**
  A plot of the sample ranges for the levels of the variables.

- **Simple Regression Analysis**
  An analysis that fits a model that relates one dependent variable to one independent variable by minimizing the sum of the squares of the residuals for the fitted line.

- **Skewness**
  A value used to measure the symmetry or shape of the data.

- **Skychart**
  A three-dimensional histogram that shows the relationship among two variables and the distribution of the data.

- **SnapStats**
  New analyses designed to provide one-page summaries for commonly encountered data analysis problems.

- **Somer’s D Statistics**
  Statistics that show a symmetric measure of association for variables that are measured on an ordinal scale.

- **Spearman’s R**
  An option (in addition toKendall’s Tau) added to the Rank Correlations tabular option. The correlation coefficients range between -1 and +1, and measure the strength of the associations between the variables.

- **Star Plots**
  A graph of multivariate data that compares different observations so you can see the relative values for all the variables at once. Each star represents one observation in the data.

- **Start-Up Scripts**
  A special feature that allows you to perform analyses, assign values to data variables, execute Windows commands, print results, publish results to a web site, and automatically end program execution. This also allows batch-type processing.

- **StatAdvisor**
  A special feature unique to STATGRAPHICS *Plus* that allows you to display and print a short and easy-to-understand interpretation of the text and graphics options in an analysis.

- **StatFolio**
  A special feature unique to STATGRAPHICS *Plus* that is similar to a statistical project; a file in which you can save data, analyses, a StatGallery, and comments.

- **StatFolio Start-Up Scripts**
  A new option under the Edit menu that allows you to define a script that will be run whenever the current StatFolio is loaded, either manually through the File menu or on program startup via the command line.

- **StatGallery**
  A special feature unique to STATGRAPHICS *Plus* that allows you to copy text and graphics panes to multiple pages so you can view or print them.

- **StatLink**
  A special feature unique to STATGRAPHICS *Plus* that allows you to tie StatFolios directly to data from spreadsheets, databases, or measuring devices such as digital micrometers (via linking software). The data source is queried when a StatFolio is opened and can be polled at user-specified intervals.

- **StatPublish**
  A special feature unique to STATGRAPHICS *Plus* that allows you to publish a StatFolio to a web site for access by web browsers.

- **StatReporter**
  A special feature unique to STATGRAPHICS *Plus* that allows you to publish reports directly from STATGRAPHICS *Plus*.

- **StatWizard**
  A special feature unique to STATGRAPHICS *Plus* that assists new or casual users in the selection of the appropriate analysis for collecting and analyzing data.

- **Statistical Tolerance Limits**
  An analysis that takes user-specified population, mean and standard deviation information to determine confidence intervals for a proportion of the population.

- **Stem-and-Leaf Display**
  A table that shows the range and concentration of the values, the symmetry of the data, and if there are gaps or outliers.

- **Student's *t* Distribution**
  A distribution useful in forming confidence intervals for the mean when the variance is unknown, testing to determine if two sample means are significantly different, or testing to determine the significance of coefficients in a regression. The distribution is similar in shape to a Normal distribution.

- **Subset Analysis**
  An analysis that allows you to calculate statistics for a single column of data at each level of a second code variable; formerly known as the Codebook Analysis in STATGRAPHICS *Plus*.

- **Sunray Plot**
  A plot that lets you visually compare observations with the means for the variables. A separate glyph is plotted for each observation in the data.

- **Suppress Ticmark Gap**
  A field in the Edit Preferences dialog box that allows you to suppress the tickmark gap on all charts. When selected, the gap between the intersection of the axes and the first tickmark is eliminated.

- **Surface Plot**
  A three-dimensional plot of a response surface that is useful in locating optimal regions.

- **Survival Function Plot**
  A plot that shows the estimated survival function, which is the probability that an item will not fail before a specified length of time.

- **Survivor Function**
  A function that produces a plot of the survivor function for a distribution you select. The function indicates the probability of obtaining a value greater than a value you select. The survivor function is 1 - CDF. In reliability data, the survivor function shows the probability of surviving to time *x* without failure.

- **Symmetry Plot**
  A plot helpful in determining the symmetry of the data; useful in analyses that are best applied to a symmetrical distribution.

- **Table of Means**
  A table of the number of observations in each sample, the means, the pooled standard errors, and the lower and upper limits for the means.

- **Tabulation Analysis**
  An analysis that summarizes the distribution of a single categorical variable through a frequency tabulation.

- **Tail Areas**
  The area under the density curve at specified points calculated by using the cumulative distribution function. The function calculates the probability that a random variable will fall below a given value.

- **Tests for Normality**
  An analysis of the data to determine if they are normally distributed. The results include the Chi-Square Goodness-of-Fit statistic, the Shapiro-Wilk's W-statistic, the *z*-Score for Skewness, and the *z*-score for Kurtosis.

- **Time Series Operators**
  Operators that calculate seasonal and backward differences.

- **Tolerance Limits**
  The values between which you can expect to find a specified proportion of a population.

- **Triangular Distribution**
  A distribution useful for random variables constrained to lie between two fixed limits. This distribution peaks at some value between two limits and is characterized by three parameters: Lower Limit, Central Value (Mode), and Upper Limit.

- **Tukey HSD (Honestly Significant Difference) Method**
  A method for determining statistically significant differences among means, based on the studentized range distribution.

- **Two-Sample Comparison Analysis**
  An analysis used to calculate the confidence intervals for the difference between the population means and the ratio of the population variances, and to calculate confidence intervals for the hypotheses means, variances, and medians.

- **Type I Sums of Squares**
  A method for computing the sums of squares for each factor in the order in which they were entered.

- **Type II Sums of Squares**
  A method for computing additional sums of squares for each factor as though that factor was the last one added to the model.

- **Uncensored Data**
  Data that are present throughout the entire duration of an experiment or data that do not have to be excluded from an experiment for any reason.

- **Uncensored Data Analysis**
  An analysis that allows you to create a relevant summary of data by fitting a probability distribution function to determine if the data follow a Normal or another type of distribution.

- **Uniform Distribution**
  A distribution useful for characterizing data that range over an interval of values, each of which is equally likely.

- **Univariate Plot**
  A plot that allows you to study the distribution of cases for one variable.

- **Unusual Residuals**
  A table of the observations with studentized residuals less than -2 or greater than 2.

- **Upper Quartile**
  The value above which 25 percent of the data will lie.

- **Variance Check**
  The results of three statistical tests: Cochran's C test, Bartlett's test, and Hartley's test. These tests confirm the assumption that the variance of samples is equal.

- **Variance Components Analysis**
  An analysis used to analyze the effect of one or more qualitative factors on one response variable when the data are completely nested or hierarchical.

- **View Published Results**
  A menu option that starts up the default browser and loads the table of contents for the published StatFolio.

- **Weibull Analysis**
  An analysis that uses the Weibull distribution modeling method to create failure-rate curves.

- **Weibull Distribution**
  A distribution useful for random variables constrained to be greater than 0; characterized by the Shape and Scale parameters.

- **Weibull Plot**
  A plot of the Weibull distribution that is helpful in determining if the distribution is a good fit for a set of data. See Weibull Distribution.

- **X-Y Lineplot**
  An X-Y Plot that connects lines without points.

- **X-Y-Z Lineplot**
  A three-dimensional X-Y-Z Plot for three or more variables that connects lines without points.

- **X-Y Plot**
  A scatterplot that shows points only (no connecting lines). It plots one variable versus another to examine the relationships.

- **X-Y-Z Plot**
  A three-dimensional scatterplot for three or more variables that shows points only (no connecting lines).

Index