StatPoint Technologies, Inc.

STATGRAPHICS® Centurion XVI

User Manual

STATGRAPHICS® CENTURION XVI USER MANUAL

Reference as: *STATGRAPHICS® Centurion XVI User Manual*

Printed in the United States of America.

Table of Contents

Preface

This book is designed to introduce users of STATGRAPHICS Centurion XVI to the basic operation of the program and its use in analyzing data. It provides a comprehensive overview of the system, including installation, data management, creating statistical analyses, and printing and publishing results. Since the book is intended to get users up to speed quickly, it concentrates on the most important features of the program, rather than trying to cover every detail. The Help menu within STATGRAPHICS Centurion XVI gives access to an extensive amount of additional information, including a separate PDF file for each of the approximately 160 statistical procedures.

The first nine chapters of this book cover basic use of the program. While you could probably figure out much of this material on your own while using the program, thorough reading of those chapters will help you get up to speed quickly and ensure that you don't miss any important features.

The last seven chapters include tutorials intended to:

1. Introduce you to some of the more commonly used statistical analyses.

2. Illustrate how the unique features of STATGRAPHICS Centurion XVI facilitate the data analysis process.

It is recommended that you explore the tutorials, since they will give you a good idea of how STATGRAPHICS Centurion XVI is best used when analyzing actual data.

NOTE: a copy of this manual in PDF format is included with the program and may be accessed from the *Help* menu. In the PDF document, all of the graphs are in color. The data files and StatFolios referenced in the manual are also provided with the program.

StatPoint Technologies, Inc.
August, 2009

Chapter 1

Getting Started

Installing STATGRAPHICS Centurion XVI, launching the program, and creating a simple data file.

1.1 Installation

STATGRAPHICS Centurion XVI is distributed in two ways: over the Internet in a single file that is downloaded to your computer, and as a set of files on a CD-ROM. To run the program, it must first be installed on your hard disk. As with most Windows programs, installation is extremely simple:

Step 1: If you received the program on a CD, insert the CD into your CD-ROM drive. After a few moments, the setup program should begin automatically. If it does not, open Windows Explorer and execute the file **sgcinstall.exe** in the root directory on the CD-ROM.

If you downloaded the program over the Internet, locate the file that you downloaded and double-click on it to begin the installation process.

Step 2: A number of dialog boxes will then be displayed. If you are running the program from a CD, the first dialog box asks you to specify the language or languages to be installed:

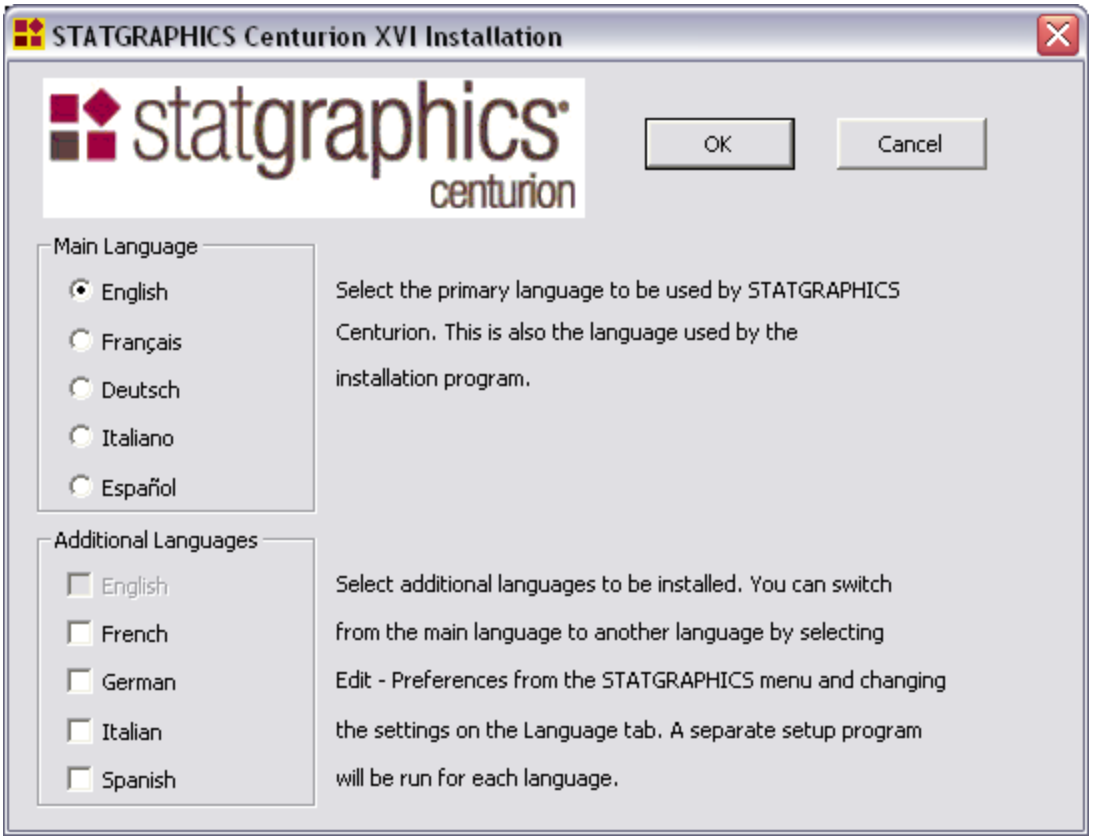


Figure 1-1. Language Selection Dialog Box

Select a main language and one or more additional languages. The main language will be used during installation and also as the default language when the program is first run. If you install additional languages, you can switch between languages while in the program by selecting *Edit – Preferences* from the main menu.

If you downloaded the program from the Internet, you will need to run a separate setup program for each language that you downloaded.

NOTE: During the evaluation period users may access any of the languages available in STATGRAPHICS Centurion XVI. Upon purchase you will be asked to designate your main and additional language (if any). Please note that only those languages specified will be available for use in STATGRAPHICS Centurion XVI.

Step 3: STATGRAPHICS Centurion XVI uses InstallShield to install the program on your computer. The InstallShield Wizard controls the installation through a series of dialog boxes. The first dialog box welcomes you to STATGRAPHICS Centurion XVI:

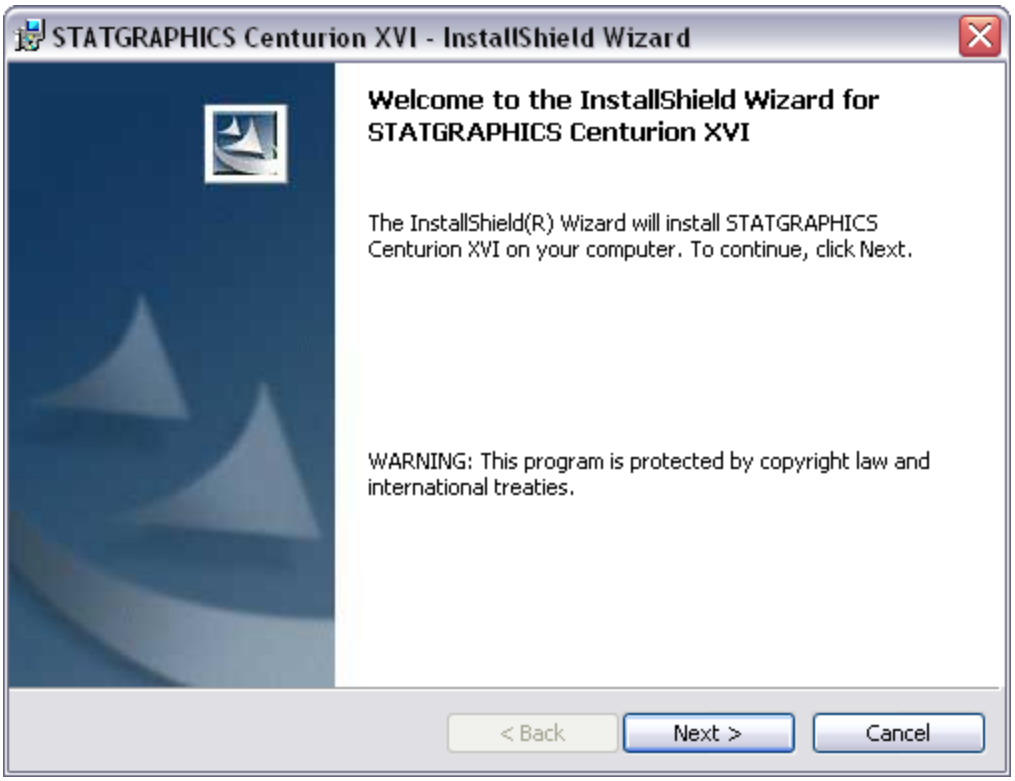


Figure 1-2. Welcome Dialog Box

Just press the *Next* button.

NOTE: NOTE: In order to install and activate STATGRAPHICS Centurion XVI you must have administrator rights to your computer. In the event that you need to have a system administrator present during the installation process, we highly recommend **installing and activating** the software while they are present.

Step 4: The second dialog box displays the license agreement for the software:

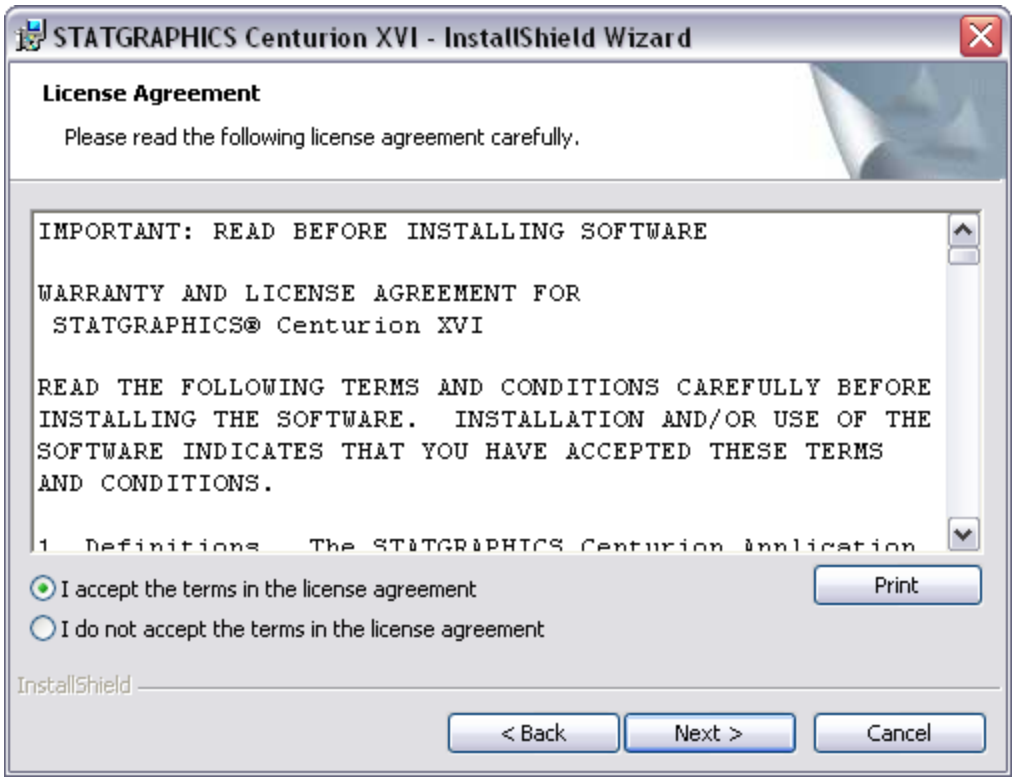


Figure 1-3. License Agreement Dialog Box

Read the license agreement carefully. If you accept the terms, click on the indicated radio button and press *Next* to continue. If you do not agree, press *Cancel*. If you do not agree with the terms, you may not use the program.

Step 5: The next dialog box requests information about the person who will use the program:

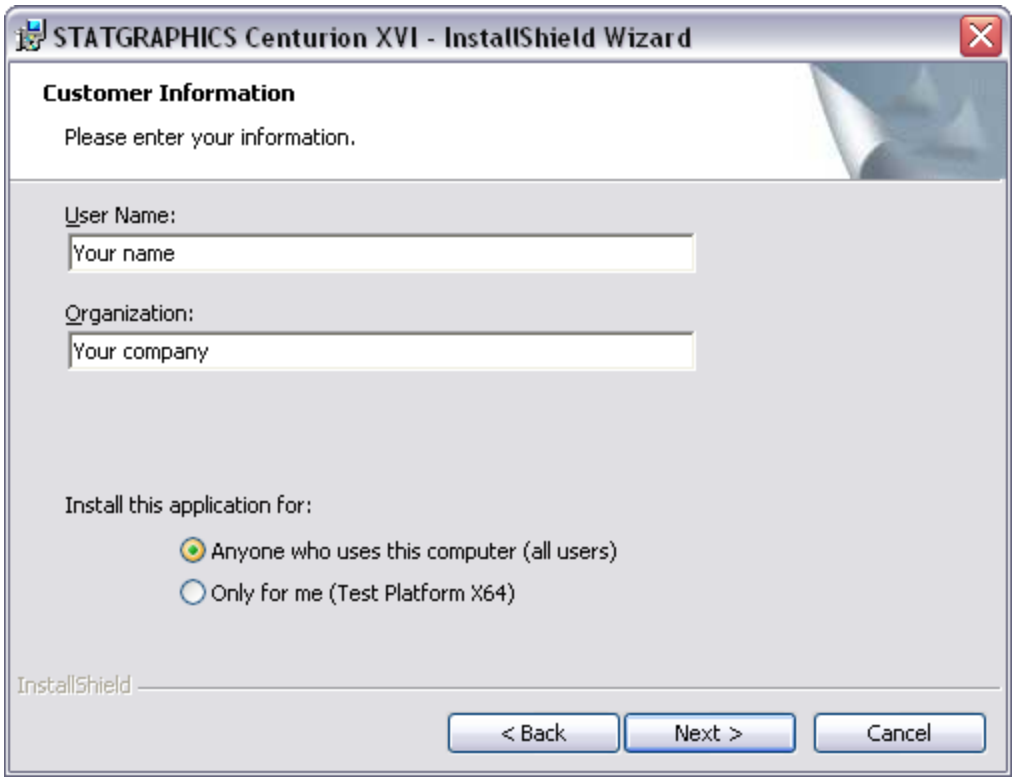


Figure 1-4. Customer Information Dialog Box

Enter the requested information. If you want to allow anyone who uses the computer to have access to STATGRAPHICS Centurion XVI, select the appropriate radio button.

Step 6: The next dialog box indicates the directory in which the program will be installed:

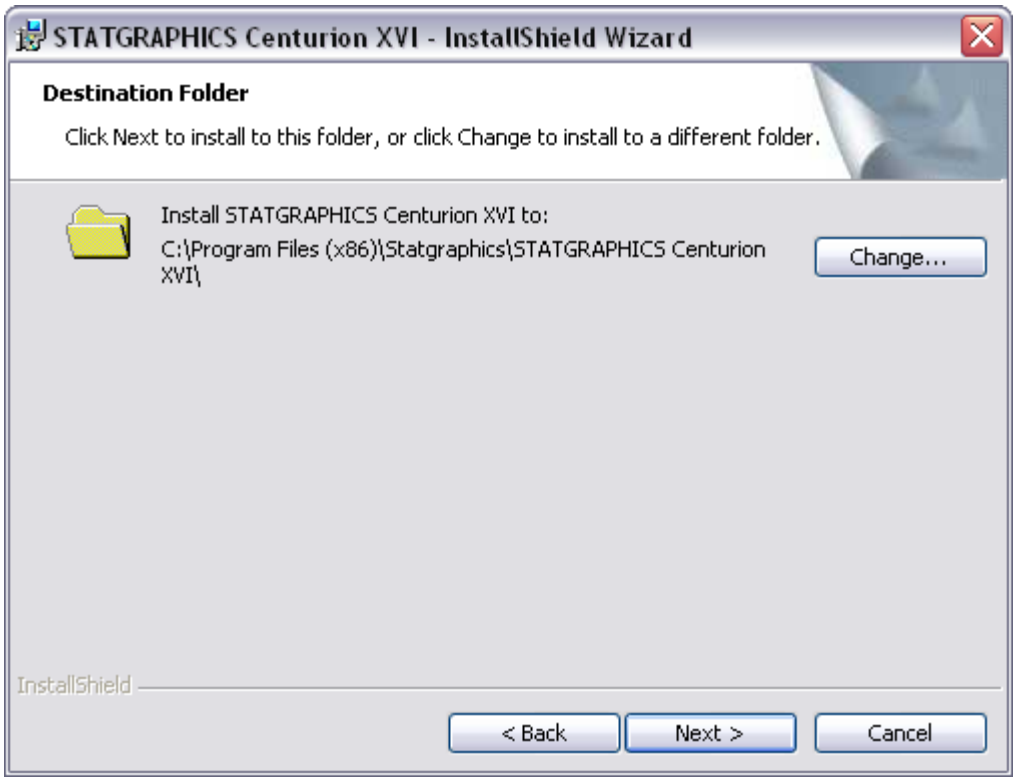


Figure 1-5. Destination Folder Dialog Box

By default, STATGRAPHICS Centurion XVI is installed in a subdirectory of *Program Files* named *STATGRAPHICS Centurion XVI*. If you are installing the program on a network server, install it in any location where all potential users have read access. Write access by users is not required. Consult the *Readme.txt* file from the STATGRAPHICS Centurion XVI CD or downloaded file for details on network installation.

Step 7: The next dialog box allows you to specify the type of installation to be performed:

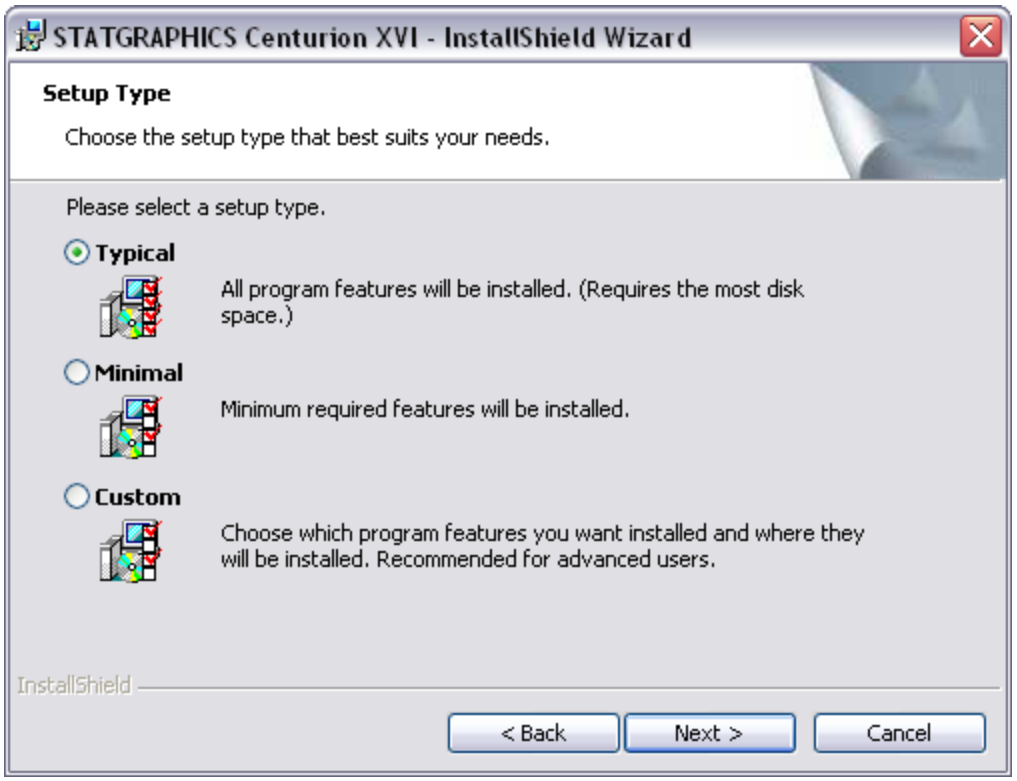


Figure 1-6. Setup Type Dialog Box

Select one of the following:

Typical – installs the program, help files, documentation, and sample data files. This requires a little more than 60MB of space on your hard disk.

Minimal – installs only the program and help files. This requires about 30MB of space on your hard disk.

Custom – installs only the components you select.

You can save hard disk space by performing a minimal install, but you won't have access to the on-line documentation and accompanying sample data files.

Step 8: Follow the remaining instructions to complete the installation. When the installation is complete, a final dialog box will be displayed:

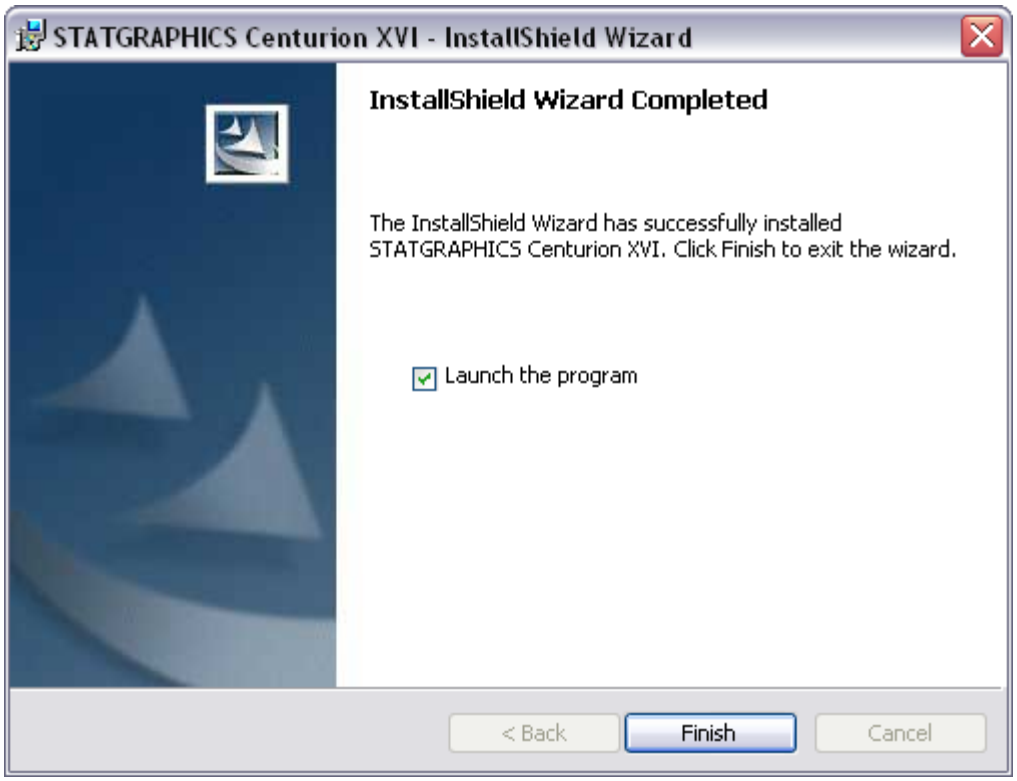


Figure 1-7. Final Installation Dialog Box

Click on *Finish* to complete the installation. Check the *Launch the program* button if you wish to start STATGRAPHICS Centurion XVI immediately, or follow the instructions below.

1.2 Running the Program

As part of the installation process, a shortcut to STATGRAPHICS Centurion XVI will be added to the Windows *Start* menu and also to your desktop. To launch the program:

Step 1: Click on the shortcut that was added to your desktop, or press the Windows *Start* button in the bottom left corner of your screen and click on the *Statgraphics* icon. You may also select *Programs Files – Statgraphics - STATGRAPHICS Centurion XVI* using Windows Explorer and click on the *sgwin* application icon to execute the program.

Step 2: When STATGRAPHICS Centurion XVI loads, it will open up a new window. The first time you launch the program, the *Welcome* dialog box will be displayed:

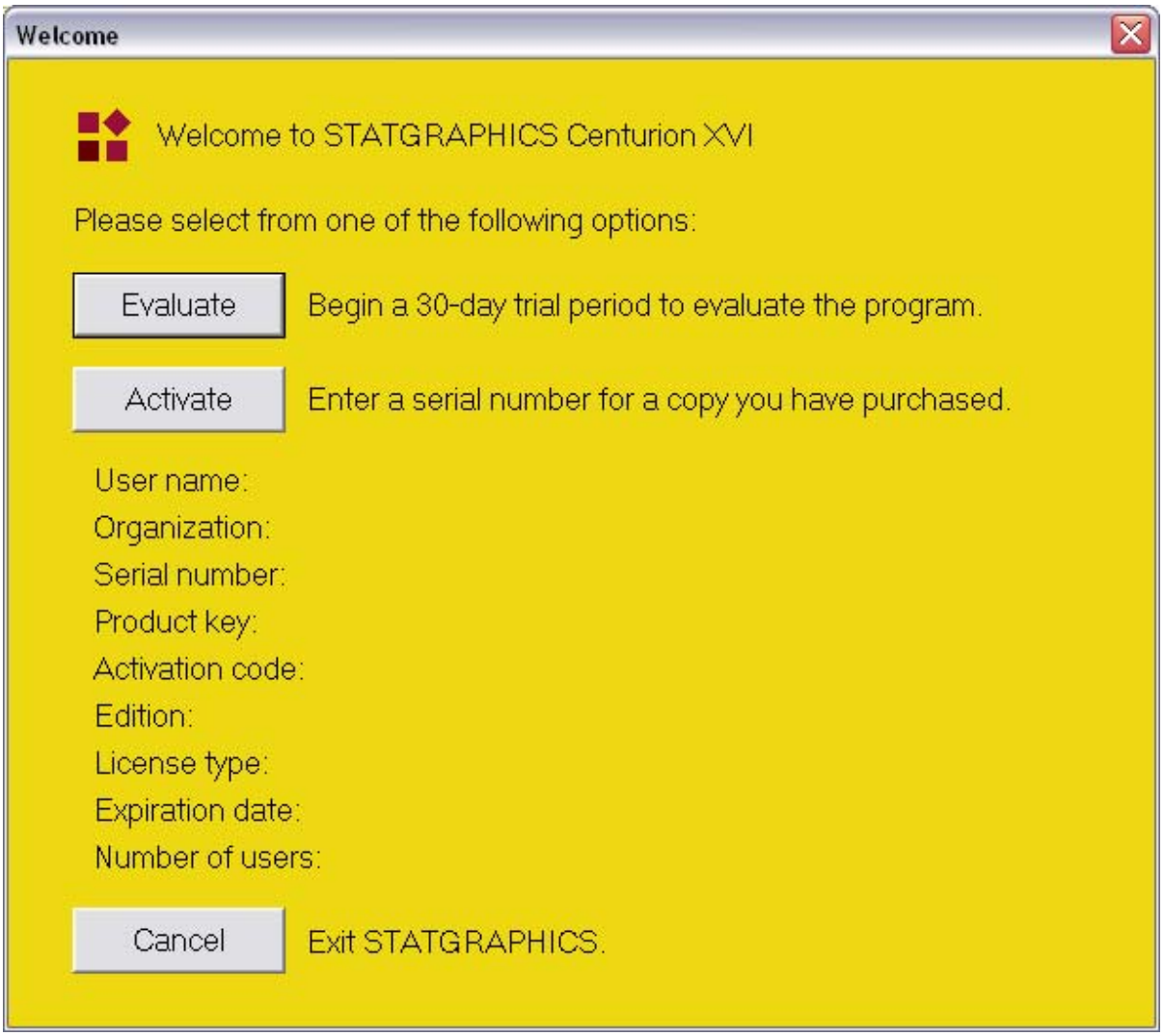


Figure 1-8. Welcome Dialog Box

You have two choices:

1. To begin a 30-day trial before purchasing the program, push the *Evaluate* button.

2. If you have already purchased the program and have received a serial number, press the *Activate* button.

If you push the *Evaluate* button, the following dialog box will be displayed:

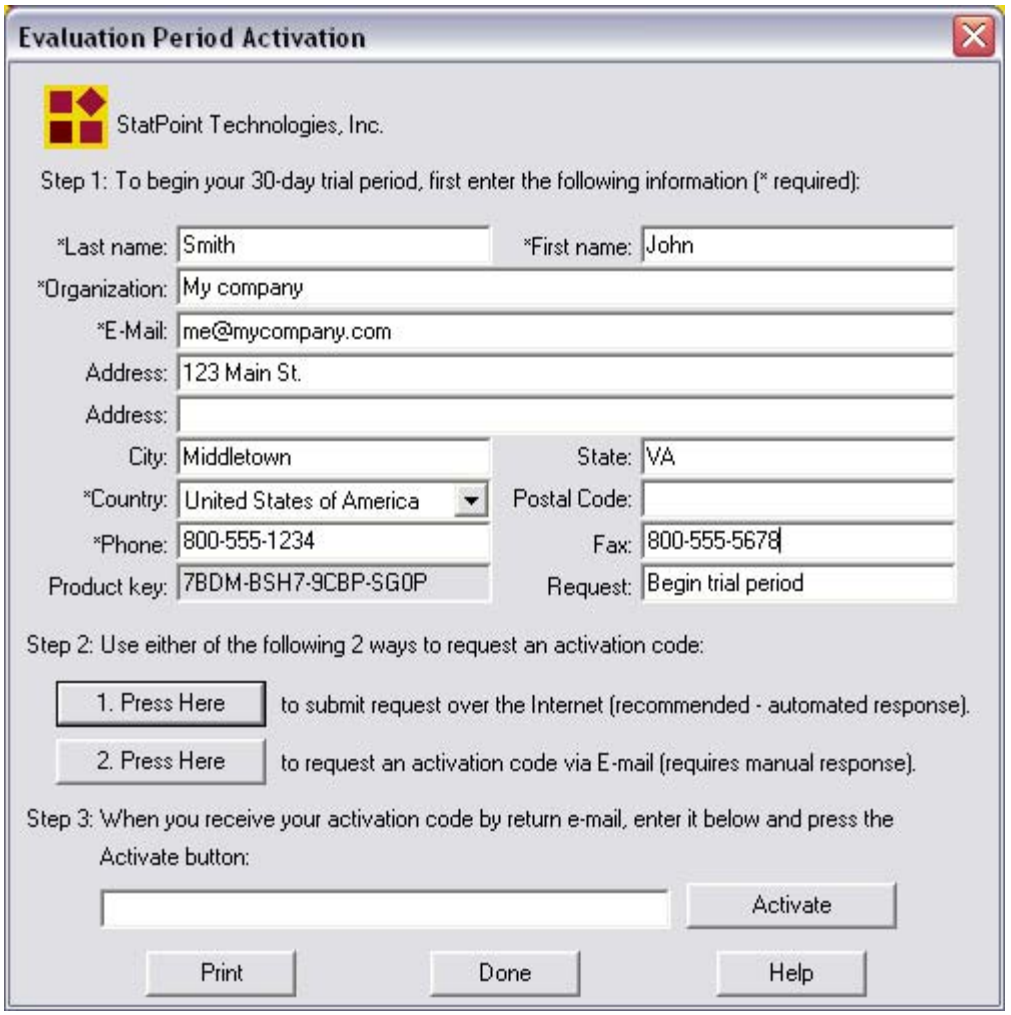


Figure 1-9. Evaluation Period Activation Dialog Box

The dialog box displays a 16-character *Product Key* that is unique to your computer. To begin your evaluation period, you must enter a matching *Activation Code*. To receive an Activation Code, press either of the two buttons under Step 2:

1. The button labeled *1. Press Here* automatically sends a message to StatPoint Technologies over the Internet requesting an Activation Code. A web service will respond to that request immediately, sending the Activation Code to the e-mail address that you supply.

2. The button labeled *2. Press Here* accesses your default e-mail program, placing the information in a new e-mail that you can send to StatPoint. E-mail requests will be processed during normal business hours.

To avoid delay, use the first method whenever possible.

Step 3: Once your request is processed, an e-mail will be sent to you containing the Activation Code. Enter the code in the field provided under *Step 3* and press the *Activate* button. If the code matches your product key, you will see the following message:

Figure 1-10. Activation Message

Press *OK* to enter the main section of the program.

NOTE #1: If you are running Microsoft Vista when you attempt to double-click on the STATGRAPHICS icon on your desktop to start the program you may not be successful. If this is the case, you must right-click on your mouse and select *Run as Administrator* from the list of options that appear.

NOTE #2: If you later install STATGRAPHICS Centurion XVI on a different computer you will have to repeat the process of obtaining an activation code, as the product key will be unique to each machine.

Step 4: The first time you run the program, you will be asked which menu system you wish to use:

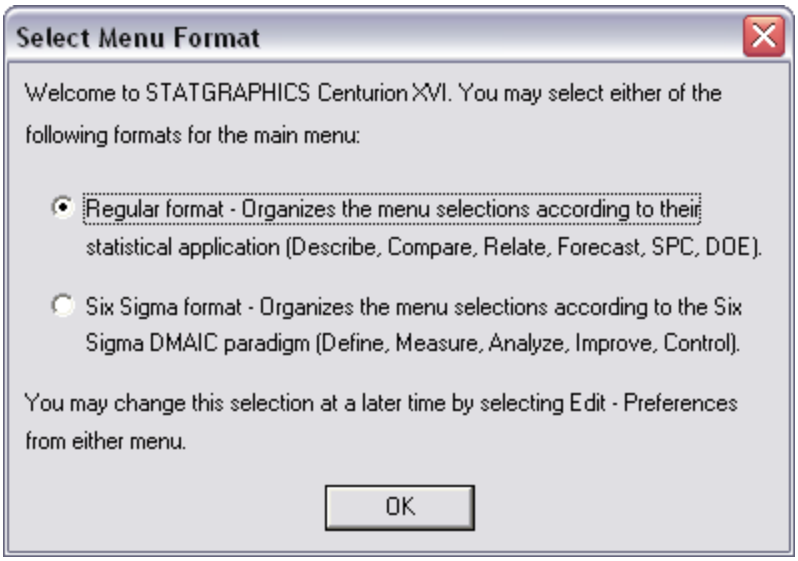


Figure 1-11. Menu Selection Dialog Box

You have a choice of the regular STATGRAPHICS Centurion XVI menu, which organizes the statistical procedures into the headings *Plot*, *Describe*, *Compare*, *Relate*, *Forecast*, *SPC*, and *DOE*, or the Six Sigma menu, which organizes the procedures into the headings *Define*, *Measure*, *Analyze*, *Improve*, *Control* and *Forecast*. Both menus include the same procedures. Only the organization is different. You may change your initial choice at a later time by selecting *Preferences* from the *Edit* menu within the program.

Step 5: The main STATGRAPHICS Centurion XVI window will then be created:

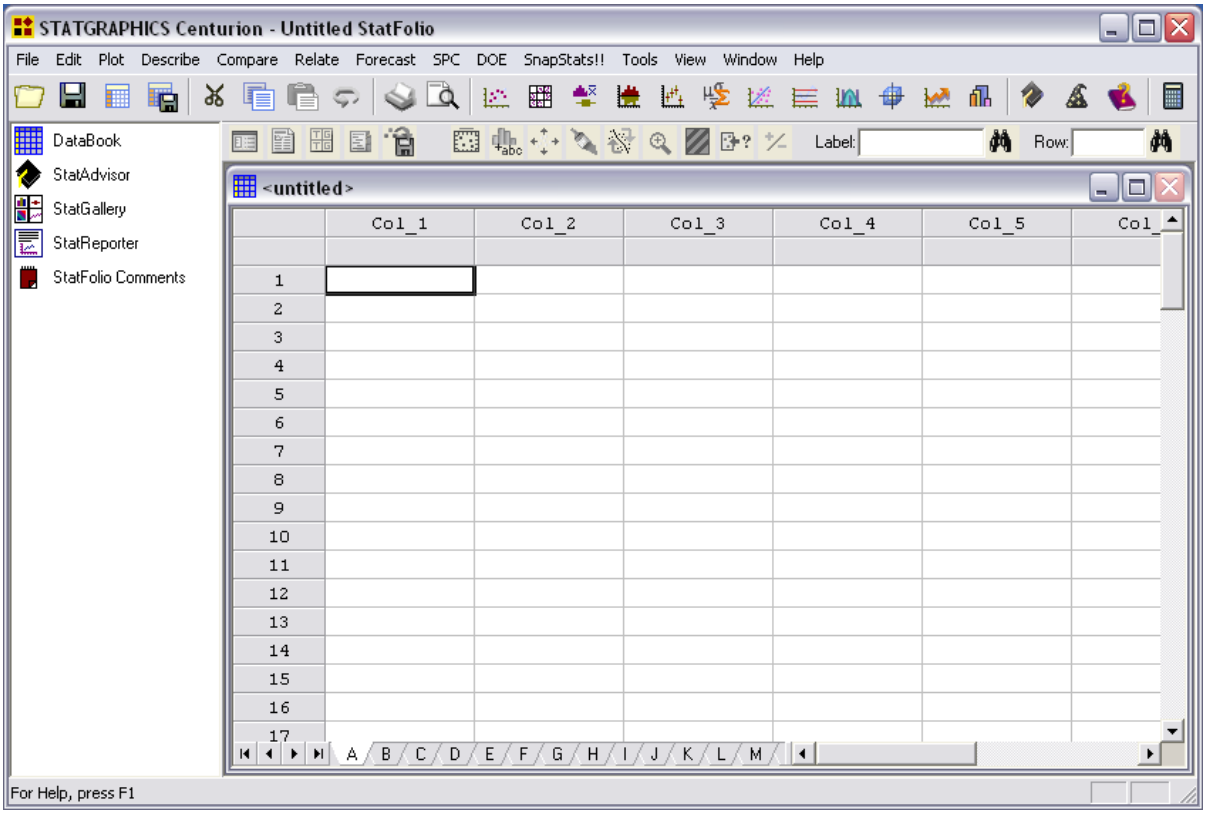


Figure 1-12. Main STATGRAPHICS Window

The sections that follow illustrate how to create a data file containing data from the 2000 United States Census.

1.3 Entering Data

In order to analyze data in STATGRAPHICS Centurion XVI, it must be placed into the STATGRAPHICS *DataBook*. The DataBook consists of 26 datasheets, indicated by the letters A through Z, each containing a rectangular array of rows and columns:

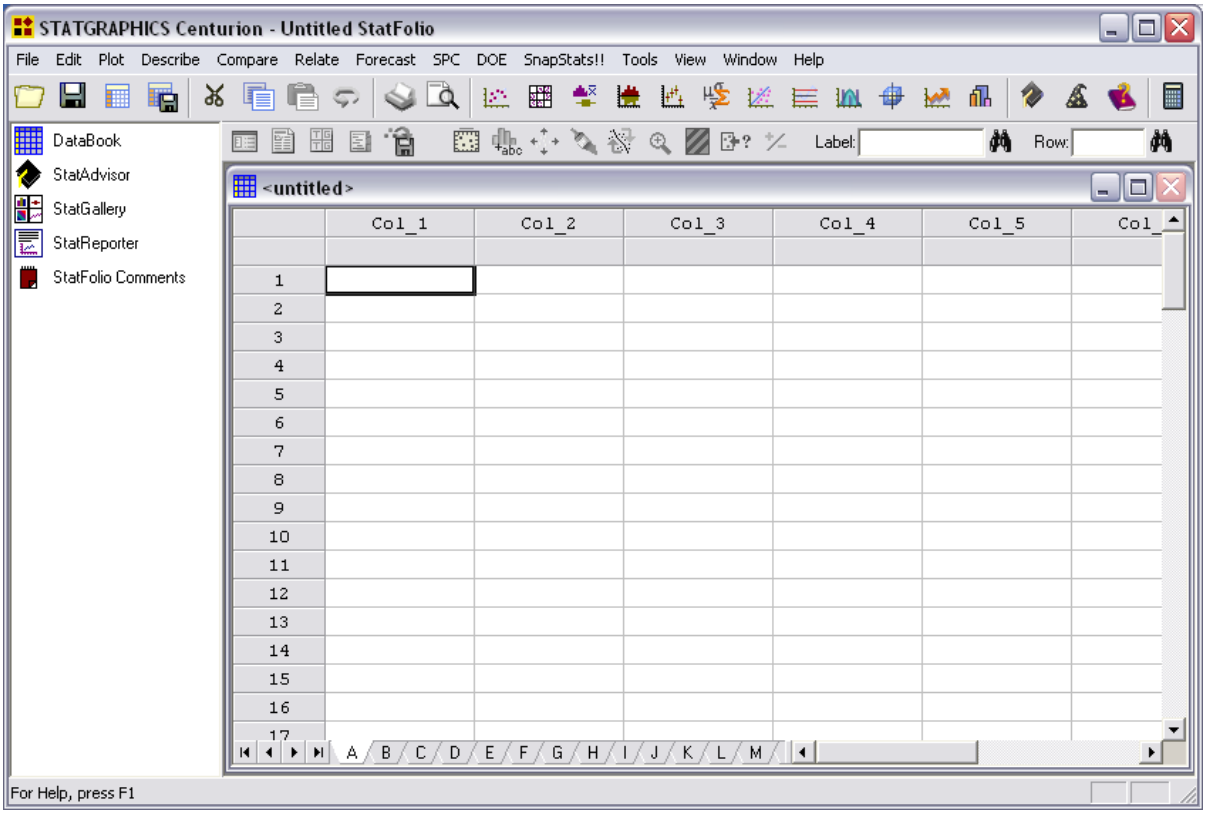


Figure 1-13. The STATGRAPHICS DataBook

In a typical datasheet, each row contains information about an individual sample, case or observation, while each column represents a variable.

For example, suppose you wished to use STATGRAPHICS Centurion XVI to analyze data from the 2000 United States Census. A small section of the results of that census is shown below:

State	Population	Median Age	% Female	Per Capita Income
Alabama	4,447,100	35.8	51.7	$18,819
Alaska	626,932	32.4	48.3	$22,660
Arizona	5,130,632	34.2	50.1	$20,275
Arkansas	2,673,400	36.0	51.2	$16,904
California	33,871,648	33.3	50.2	$22,711
Colorado	4,301,261	34.3	49.6	$24,049

Figure 1-14. Data from the 2000 U.S. Census

When entering this data into a STATGRAPHICS Centurion XVI datasheet, the information about each state would be placed into a different row. Five columns would be created to hold the names of the states and the census data.

To enter data such as that shown above into STATGRAPHICS Centurion XVI, you have two choices:

1. Type the data directly into the STATGRAPHICS Centurion XVI DataBook.

2. Enter the data into another program such as Excel and then read or copy it into STATGRAPHICS Centurion XVI.

In this section, we'll take the first approach. To begin, double-click on the header of the first column where the column name *Col_1* appears. This will display a dialog box that you can use to change important properties of that column:

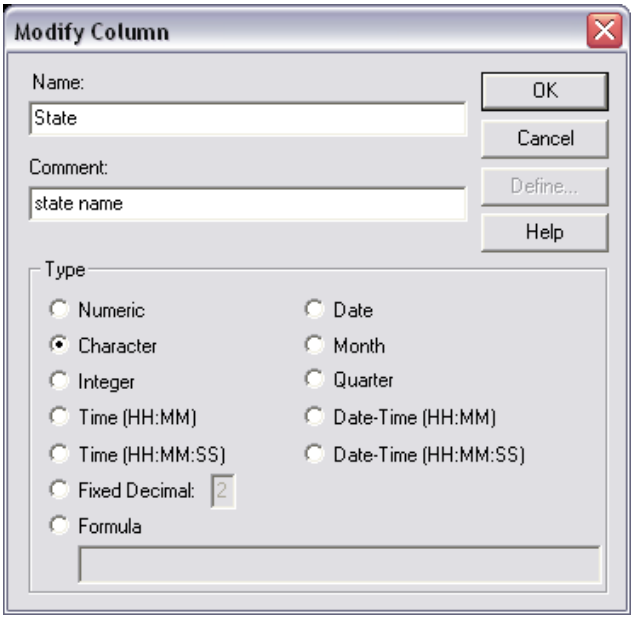


Figure 1-15. Dialog Box Used to Define Columns

Each column in a STATGRAPHICS Centurion XVI datasheet has a name, comment, and type associated with it:

- *Name*– Give each column a unique name containing from 1 to 32 characters. These names are used by the program to identify the variables to be analyzed when a statistical procedure is selected. They also serve as default labels on most graphs. Names may contain any characters and are not case sensitive. Spaces are permitted. The program will display an error message if you try to use the same name for more than one column in a datasheet, although columns in different datasheets may have identical names.

- *Comment* – Enter a comment identifying the data in the column. Comments may have up to 64 characters and are optional. If entered, they appear in the second line of the column header.

- *Type* – Specify the type of data to be entered in the column. In this case, the first column containing state names must be set to *Character*. The other columns may be left as *Numeric* or set to *Integer* or *Fixed Decimal* if you want to restrict the type of data that may be entered. For detailed information on column types, see Chapter 2.

After defining each column, press *OK*. Create five columns as shown below:

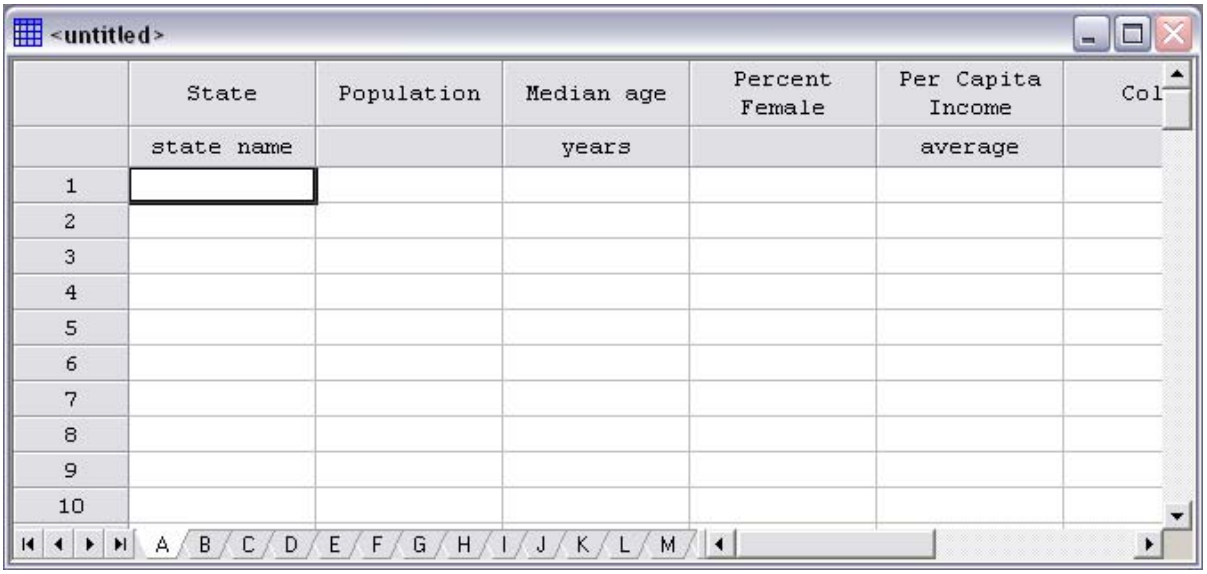


Figure 1-16. STATGRAPHICS Centurion XVI Data Sheet with Column Names

Now enter the data as you would in any spreadsheet, using the arrow keys to move from cell to cell. **DO NOT** enter commas when entering large numbers. When done, the datasheet should have the following appearance:

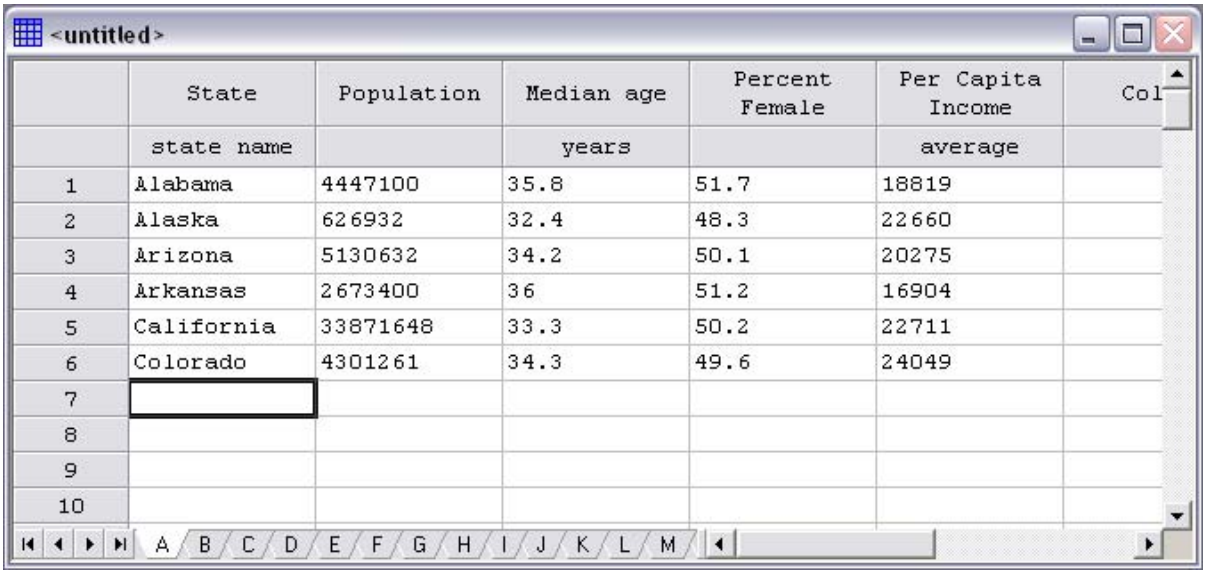


Figure 1-17. STATGRAPHICS Centurion XVI Data Sheet after Entering 6 Rows of Data

Finally, you need to save the data file. Choose *File* – *Save* – *Save Data File* from the main menu. Select a file name in which to save the data:

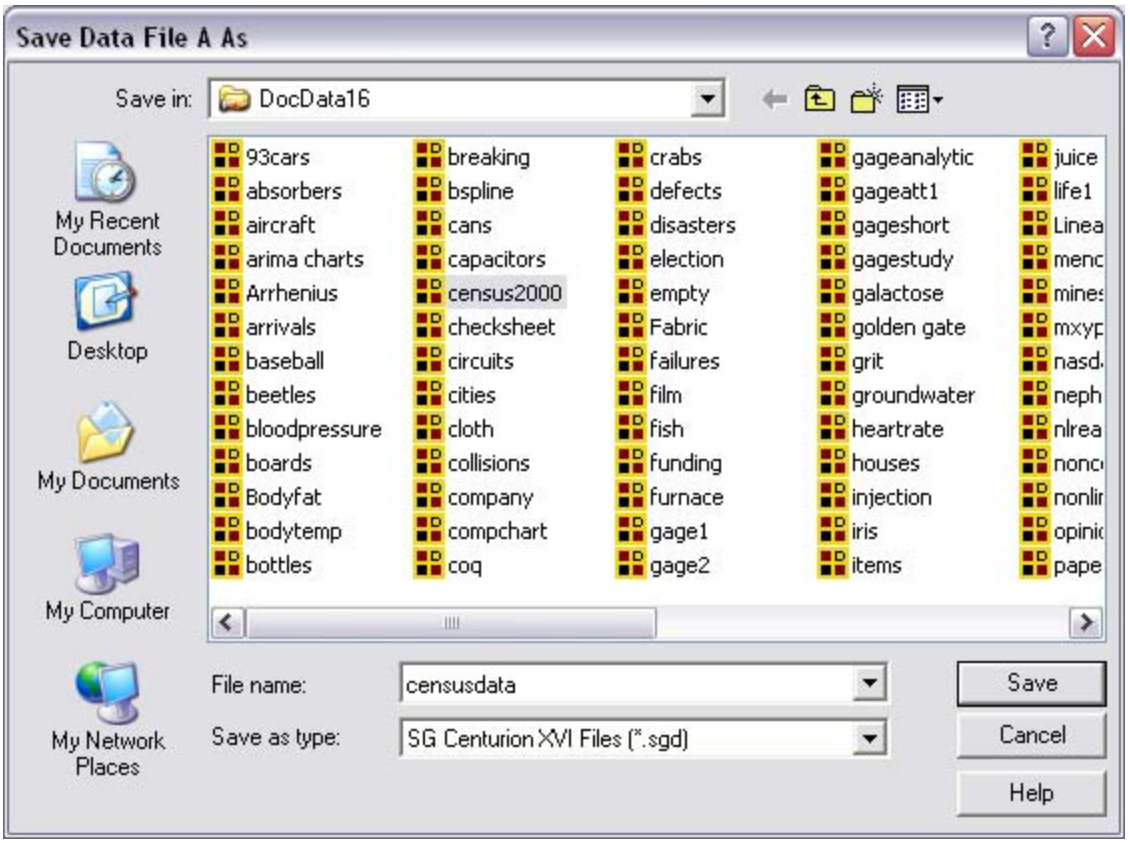


Figure 1-18. Save Data File Selection Dialog Box

Data files in STATGRAPHICS Centurion XVI are saved on disk by default with an extension of *.sgd*, which stores the data in XML format. When saving the file, you may change the setting in the *Save as type* field to a different file format if desired.

1.4 Reading a Saved Data File

Once the data have been entered into the datasheet, it is ready for analysis. To make the example more interesting, let's retrieve the census data for all 50 states and the District of Columbia, which is provided with STATGRAPHICS Centurion XVI in a file named *census2000.sgd*. To open that data file, select *File* – *Open* – *Open Data Source* from the top menu. You will first be asked to specify the location of the data you wish to access:

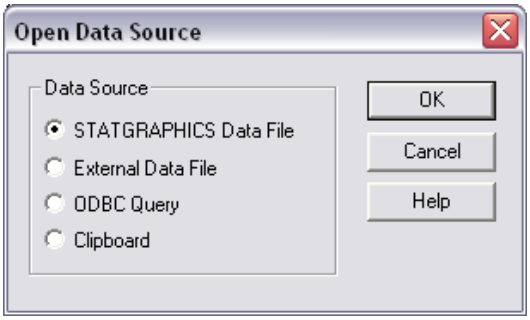


Figure 1-19. Open Data Source Dialog Box

The default selection is correct in this case. Next, select the name of the file containing the data:

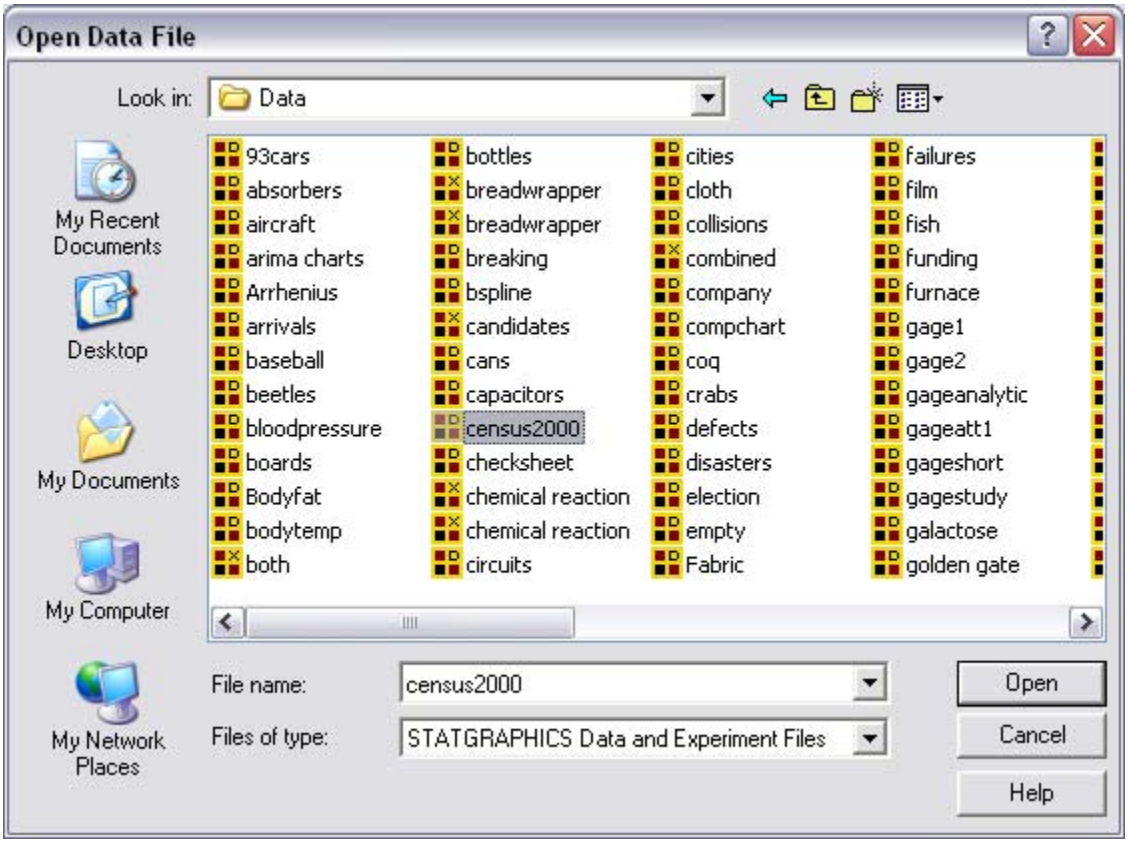


Figure 1-20. Open Data File Dialog Box

The sample file is located in the default data directory (usually *c:\Program Files\Statgraphics\STATGRAPHICS Centurion XVI\Data*). Opening the file loads the full 51 rows of data into the datasheet:

C:\Program Files (x86)\Statgraphics\STATGRAPHICS Centurion XVI\Data\census20...

	State	Population	Median Age	Percent Female	Per Capita Income
1	Alabama	4447100	35.8	51.7	18819
2	Alaska	626932	32.4	48.3	22660
3	Arizona	5130632	34.2	50.1	20275
4	Arkansas	2673400	36	51.2	16904
5	California	33871648	33.3	50.2	22711
6	Colorado	4301261	34.3	49.6	24049
7	Connecticut	3405565	37.4	51.6	28766
8	Delaware	783600	36	51.4	23305
9	D.C.	572059	34.6	52.9	28659
10	Florida	15982378	38.7	51.2	21557
11	Georgia	8186453	33.4	50.8	21154
12	Hawaii	1211537	36.2	49.8	21525

census2000 B C D E F G H I J

Figure 1-21. Datasheet Showing Contents of Census2000.sgd File

1.5 Analyzing the Data

Once the data have been loaded into the STATGRAPHICS Centurion XVI DataBook, any of the more than 160 statistical procedures may be accessed any of several ways:

1. By selecting the desired procedure from the main menu.

2. By pressing one of the shortcut buttons on the toolbar.

3. By invoking the StatWizard by pressing the button on the toolbar displaying a wizard's cap.

Let's begin by summarizing the variability in per capita income amongst the states. The best procedure for summarizing a single column of numeric data is the *One-Variable Analysis* procedure. This procedure calculates summary statistics such as the sample mean and standard deviation. It also creates several plots, including a histogram and box-and-whisker plot.

The location of the *One-Variable Analysis* procedure depends on the menu you are using:

1. **Classic menu**: Select *Describe – Numeric Data – One-Variable Analysis*.

2. **Six-Sigma menu**: Select *Analyze – Variable Data – One-Variable Analysis*.

Like all statistical procedures, the *One-Variable Analysis* begins by displaying a data input dialog box:

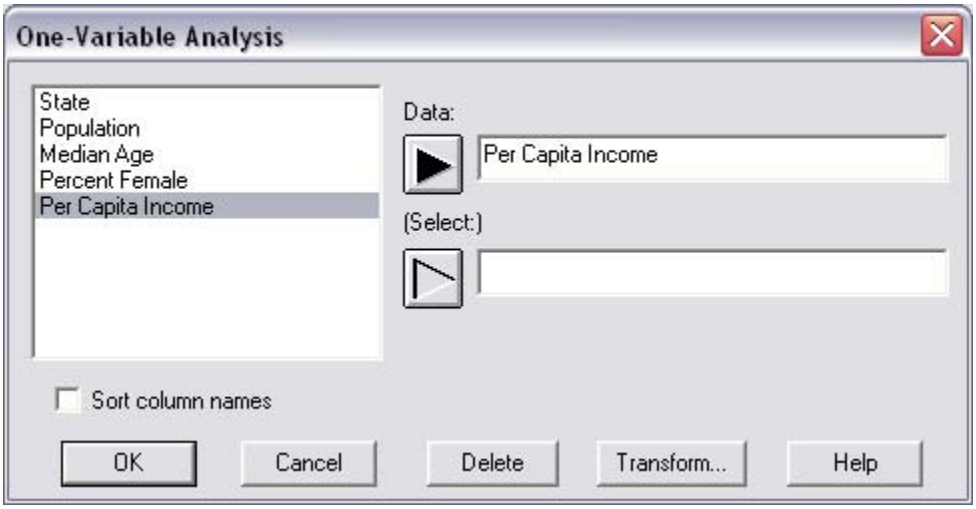


Figure 1-22. One-Variable Analysis Data Input Dialog Box

The list box at the left displays the names of all columns in the datasheets that contain data. To analyze the data in the *Per Capita Income* column, click on its name and then click on the button with the black arrow alongside the *Data* field. This places the name of the column containing the income data into the *Data* field. Leave the *Select* field blank (it is used only when you want to analyze a subset of the rows in the datasheet instead of all the rows).

When *OK* is pressed, the *Tables and Graphs* dialog box appears. This dialog box shows the tables and graphs that are available for the *One Variable Analysis* procedure. For now, the default settings will be acceptable:

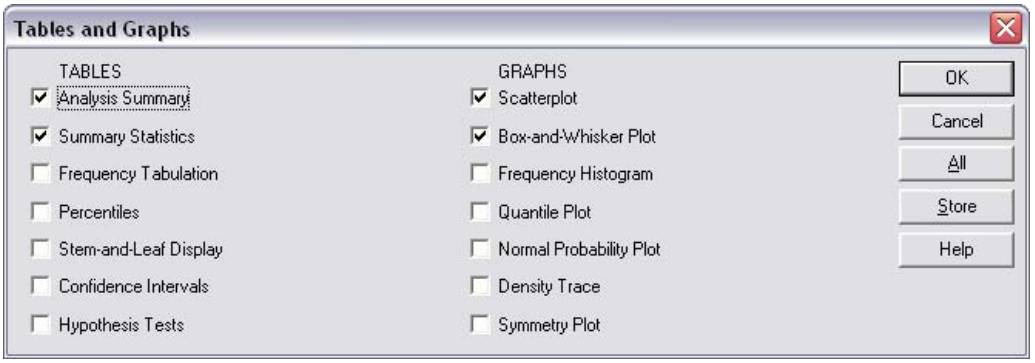


Figure 1-23. Tables and Graphs Dialog Box

When OK is pressed again, a new analysis window will be created:

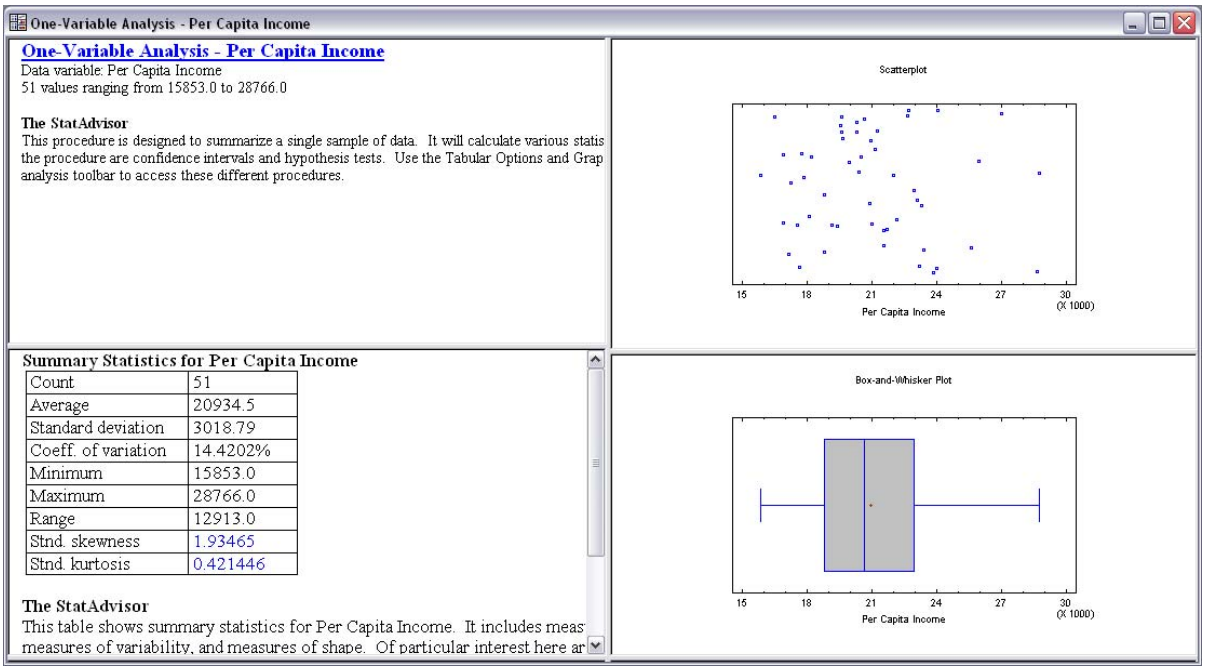


Figure 1-24. One-Variable Analysis Window

The window contains 4 "panes", divided by movable splitter bars. The two panes on the left display tabular output, while the two panes on the right display graphical output. If you double-click in the bottom left pane, the table of summary statistics will be maximized:

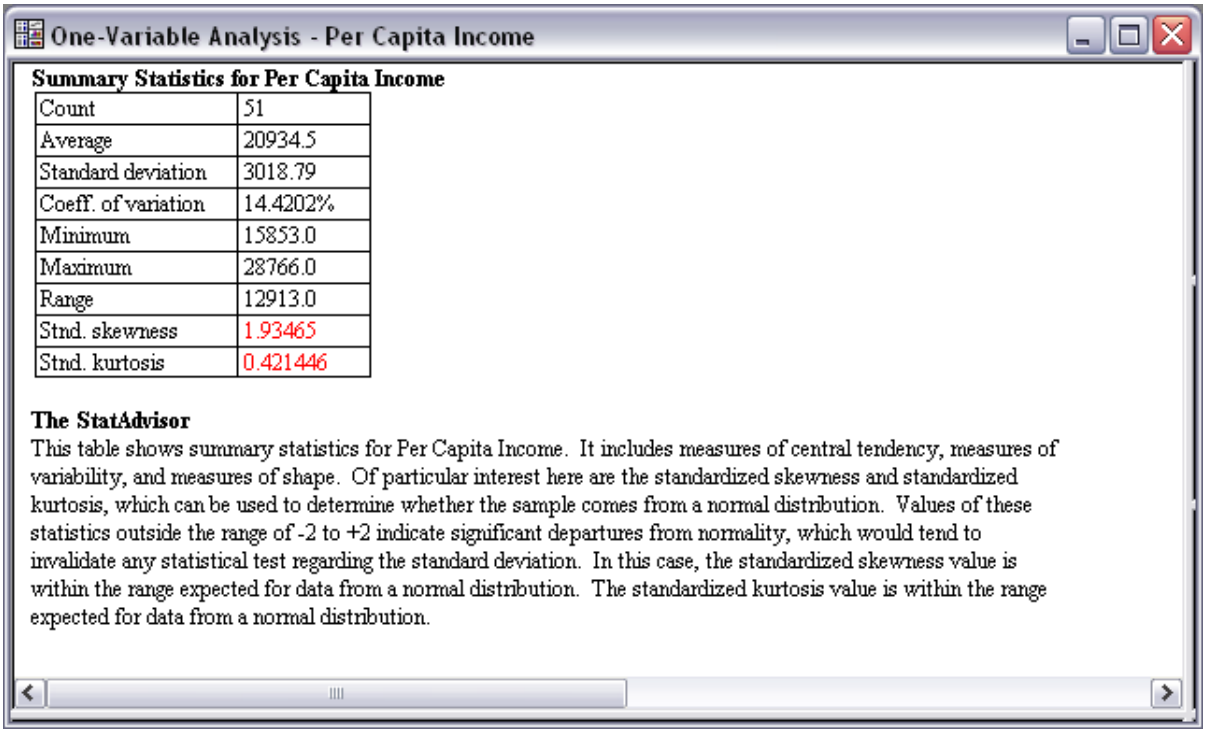


Figure 1-25. Maximized Summary Statistics Pane

Several interesting statistics are given in the table. Of the $n = 51$ states plus D.C., per capita income ranges between $15,853 and $28,766. The average per capita income is $20,934.50.

Beneath the table is the output of the StatAdvisor, which gives a short interpretation of the results. In this case, the StatAdvisor concentrates on the two highlighted statistics, which measure the skewness and kurtosis in the data. As explained by the StatAdvisor, data that come from a normal or Gaussian distribution should yield standardized skewness and standardized kurtosis values between –2 and +2. In this case, both statistics are within that range, indicating that a bell-shaped normal curve is a reasonable model for the observations, although the skewness is very close to being statistically significant.

Double-clicking on the summary statistics table again will restore the original split display. Double clicking on the bottom right pane then maximizes the box-and-whisker plot:

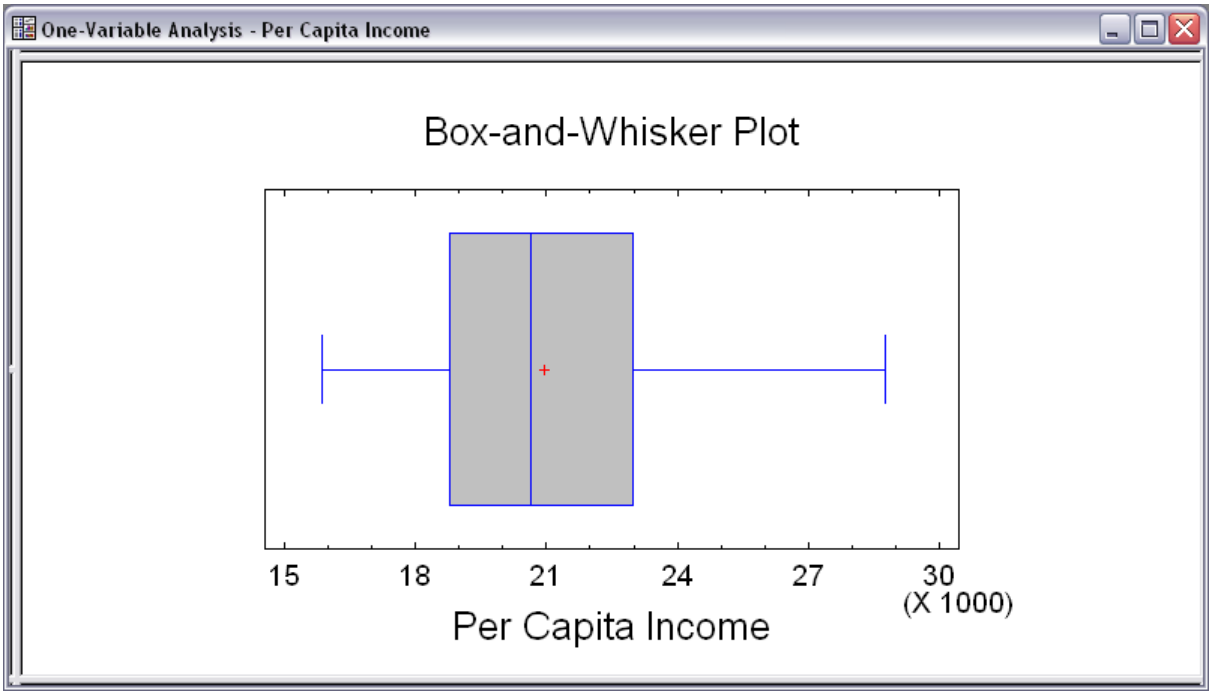


Figure 1-26. Maximized Box-and-Whisker Plot Pane

The box-and-whisker plot, invented by John Tukey, provides a 5-number summary of a data sample. The central box covers the middle half of the data, extending from the lower quartile to the upper quartile. The lines extending above and below the box (the whiskers) show the location of the smallest and largest data values. The median of the data is indicated by the vertical line within the box, while the plus sign (+) shows the location of the sample mean. The fact that the upper whisker is slightly longer than the lower, while the mean is somewhat greater than the median, is indicative of positive skewness in the data.

1.6 Using the Analysis Toolbar

When an analysis window such as the *One-Variable Analysis* is first displayed, only some of the available tables and graphs are included. To display additional output, you must push the appropriate button on the *Analysis Toolbar*, which is displayed immediately above the analysis title:

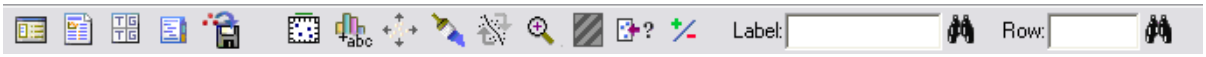


Figure 1-27. The Analysis Toolbar

The buttons on the analysis toolbar are very important. The actions of the six leftmost buttons are summarized below:

	Name	*Function*
	Input dialog	Displays the data input dialog box so that the selected data column(s) may be changed.
	Analysis options	Selects options that apply to all tables and graphs in the current analysis.
	Tables and graphs	Displays a list of other tables and graphs that may be created.
	Pane options	Selects options that apply only to the currently maximized table or graph.
	Save results	Allows calculated statistics to be saved to columns of a datasheet.
	Graphics options	Allows you to change the titles, scaling, and other features of the currently maximized graph.

Figure 1-28. Important Buttons on the Analysis Toolbar

Additional buttons to the right allow other actions when a graph is maximized, as explained in Chapter 5.

For example, if the *Tables and Graphs* button is pressed, a dialog box will be displayed listing other graphs available in the *One-Variable Analysis* procedure:

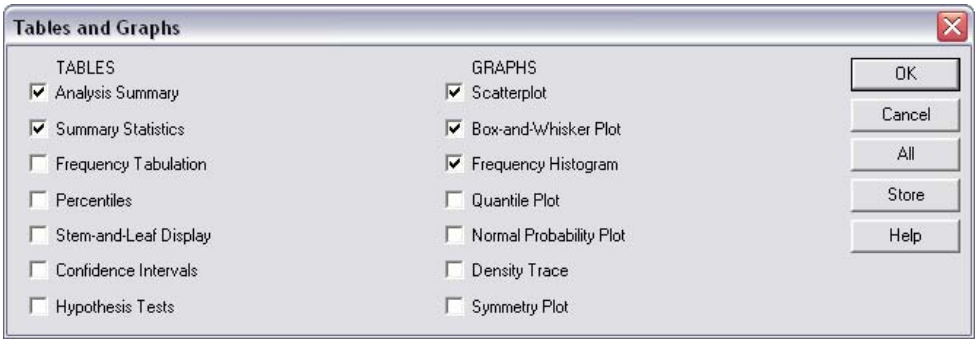


Figure 1-29. List of Available Tables and Graphs

Checking the box next to *Frequency Histogram* and pressing *OK* adds a third pane to the right-hand side of the analysis window:

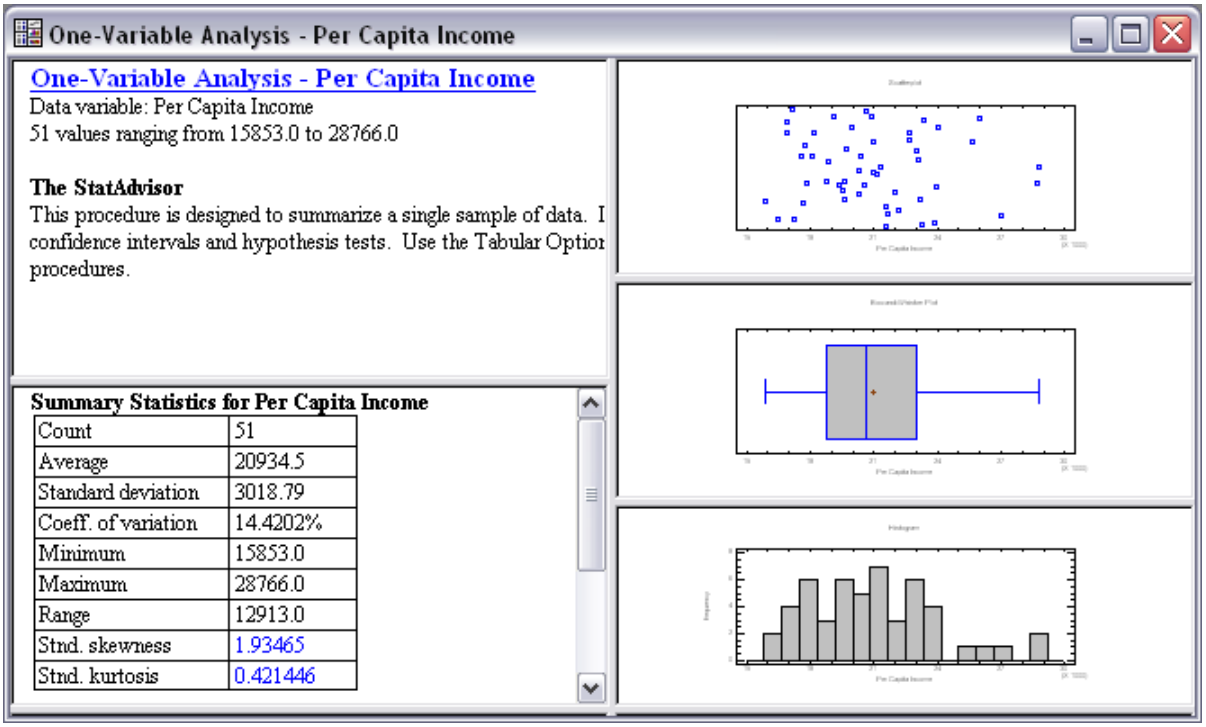


Figure 1-30. One-Variable Analysis Window with Added Frequency Histogram

If you double-click on the histogram to maximize it and then press the *Pane options* button, a dialog box is displayed with options specific to the histogram:

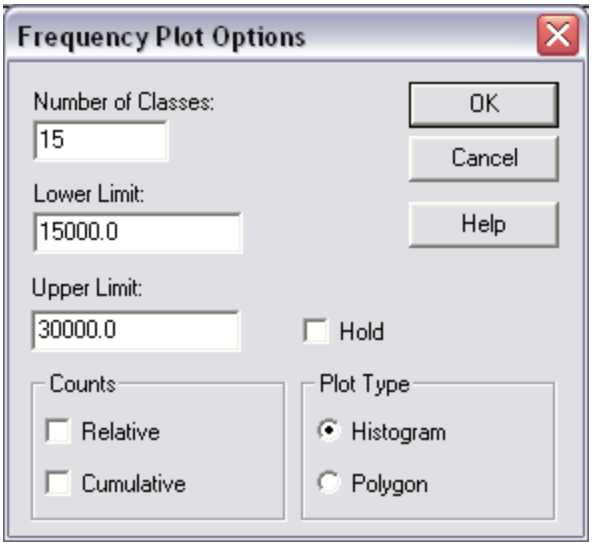


Figure 1-31. Frequency Histogram Pane Options Dialog Box

Using this box, the number of bars in the histogram can be changed, as well as the range that they cover. If *Number of Classes* is set to 15 and the *OK* button is pressed, the histogram will change to reflect the new selection:

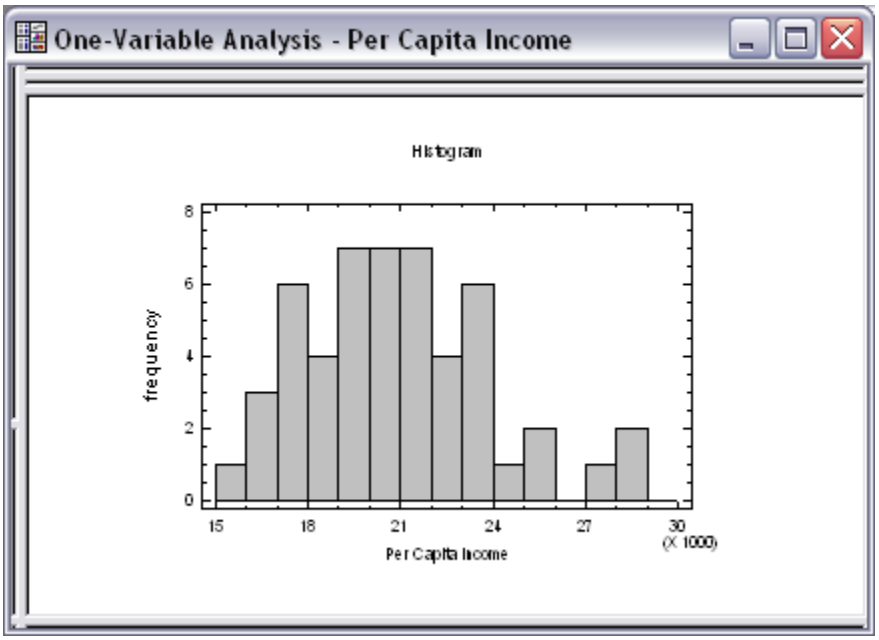


Figure 1-32. Frequency Histogram After Changing the Number of Classes

You may also change the fill pattern and/or color of the bars in the histogram by pressing the *Graphics options* button. This displays a tabbed dialog box that allows you to change most features of the graph. If you click on the *Fill* tab, the following will be displayed:

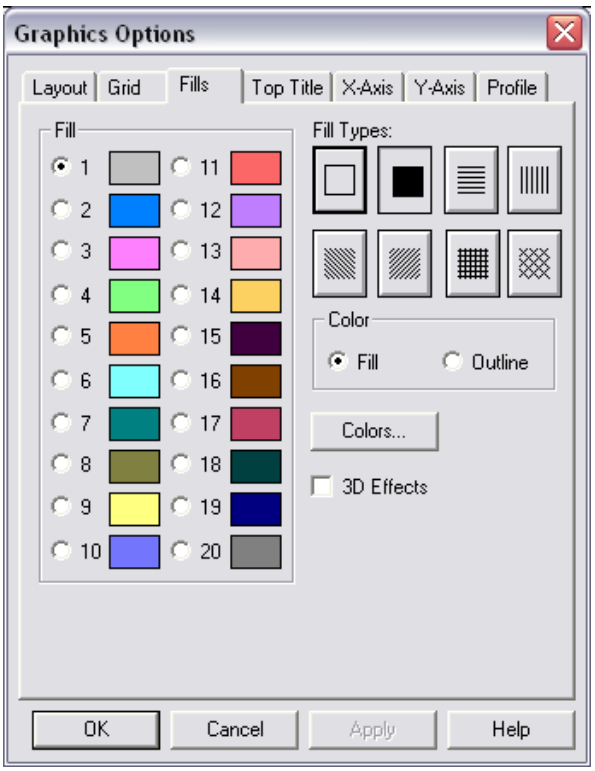


Figure 1-33. Graphics Options Tabbed Dialog Box

Clicking on radio button #1 and then selecting a new *Fill Type* or *Color* will change the bars in the histogram.

NOTE: The operations of many of the buttons on the analysis toolbar can also be accessed by clicking the alternate mouse button in the pane containing a table or graph. This displays a popup menu listing the available operations.

1.7 Disseminating the Results

Once an analysis has been performed, the results can be disseminated in various ways. These include:

Action	*Method*
Print the output.	Press the printer button on the main toolbar to print all tables and graphs, or click on a single pane with the alternate mouse button and select *Print* from the popup menu to print a single table or graph.
Publish the output for viewing in a web browser.	Select *StatPublish* from the *File* menu. A dialog box will be displayed for you to specify the location of the HTML output.
Copy the output to another application.	Click on the table or graph to be copied and select *Copy* from the *Edit* menu. Then activate the other application and select *Edit – Paste*.
Save the analysis in a report.	Press the alternate mouse button and select *Copy Analysis to StatReporter*. The StatReporter, described in Chapter 7, can be saved as an RTF file for import into programs such as Microsoft Word.
Save a graph in an image file.	Maximize the graph to be saved. Then select *Save Graph* from the *File* menu.

Figure 1-34. Methods for Disseminating Analysis Results

Each of these operations is described in later chapters.

1.8 Saving Your Work

You can save the current STATGRAPHICS Centurion XVI session at any time by selecting *Save StatFolio* from the *File* menu and entering a file name:

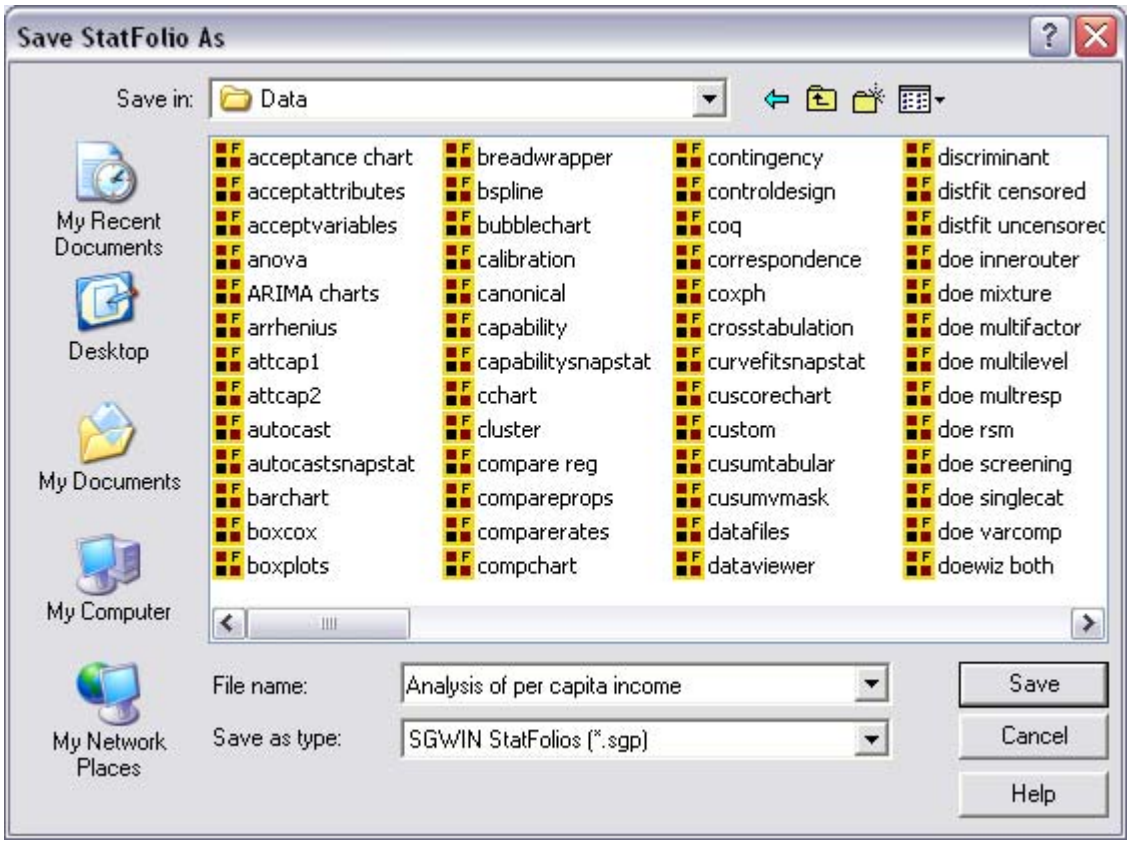


Figure 1-35. Dialog Box for Saving StatFolio

A StatFolio consists of instructions on how to create each of the analyses in your current session, with pointers to the files or databases containing your data. If you reload the StatFolio at a later date, it will automatically reread the data and recreate the analyses. Any options you have selected for the analyses will be retained.

NOTE #1: If the data in the data sources change between the time a StatFolio is saved and the time it is reloaded, the analyses will change to reflect the new values. This provides a simple method for rerunning analyses that need to be repeated on a periodic basis without having to recreate them.

NOTE #2: The data and the StatFolio are stored in different files. If you need to move a StatFolio from one computer to another, be sure to move the data file(s) as well.

Chapter

2

Data Management

Accessing data from files and databases, transforming data values, generating patterned data.

In order to analyze data in STATGRAPHICS Centurion XVI, it must first be placed in the STATGRAPHICS Centurion XVI *DataBook*. The DataBook is a tabbed window, consisting of 26 datasheets. A datasheet is a rectangular array of rows and columns. Each column in a datasheet represents a variable. Each row represents a case or observation. For example, the datasheet below contains information on a number of different makes and models of automobiles.

C:\Program Files (x86)\Statgraphics\STATGRAPHICS Centurion XVI\Data\93cars.sgd

	Make	Model	Type	Min Price	Mid Price
				price for basic version in $1,000	average of min and max prices in $1,000
1	Acura	Integra	Small	12.9	15.9
2	Acura	Legend	Midsize	29.2	33.9
3	Audi	90	Compact	25.9	29.1
4	Audi	100	Midsize	30.8	37.7
5	BMW	535i	Midsize	23.7	30
6	Buick	Century	Midsize	14.2	15.7
7	Buick	LeSabre	Large	19.9	20.8
8	Buick	Roadmaster	Large	22.6	23.7
9	Buick	Riviera	Midsize	26.3	26.3

93cars B C D E F G H I J K L

Figure 2-1. Sample Datasheet

This chapter describes everything you need to know about data and STATGRAPHICS Centurion XVI, including how to access it, how to manipulate it, and how to use it in statistical analyses.

2.1 The DataBook

Each column in the STATGRAPHICS Centurion XVI datasheet represents a different variable. Variables are usually attributes or measurements associated with the items that define the rows of the datasheet. For example, in the *93cars* datasheet, there is a column identifying the make of each automobile, a column identifying its type, columns containing the recorded miles per gallon in city and highway driving, columns containing the automobile's length, height and weight, and similar information. Each column has a *name* and *type* associated with it. The *name* is used to identify the data to use in a statistical analysis. The *type* affects how it will be analyzed. Also associated with each column is an optional *comment*, which is used to provide additional information about the contents of a column. NOTE: the data were obtained from the Journal of Statistical Education Data Archive (www.amstat.org/publications/jse/jse_data_archive.html) and are used by permission.

To display or change the properties of any column in a datasheet, double-click on the column name to display the *Modify Column* dialog box:

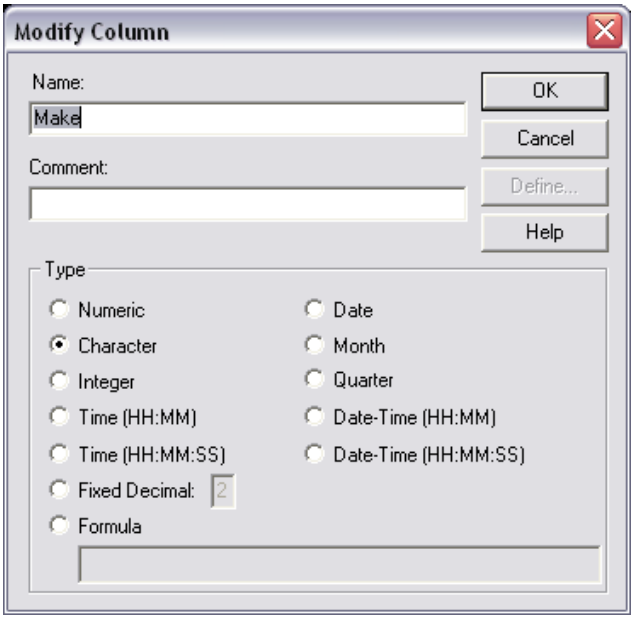


Figure 2-2. Dialog Box Used to Modify Column Properties

You may specify:

1. **Name**: from 1 to 32 characters. When performing statistical analyses, columns are identified using these names. Each column in a datasheet must have a unique name, though columns in different datasheets may have the same name. Names may include any character, including spaces. Variable names are **not** case sensitive.

2. **Comment**: from 0 to 64 characters, providing additional information about the contents of the column.

3. **Type**: the type of data permitted in the column. The following types may be specified:

Type	*Contents*	*Example*
Numeric	Any valid number.	3.14
Character	An alphanumeric string	Chevrolet
Integer	An integer number	105
Date	Month, day and year	4/30/05
Month	Month and year	4/05
Quarter	Quarter and year	Q2/05
Time (HH:MM)	Hour and minute	3:15
Time (HH:MM:SS)	Hour, minute and second	3:15:53
Date-Time (HH:MM)	Month, day, year, hour and minute	4/30/05 3:15
Date-Time (HH:MM:SS)	Month, day, year, hour, minutes and second	4/30/05 3:15:53
Fixed Decimal	Number with 1 to 9 places	34.10
Formula	Calculated from other columns	MPG City/MPG Highway

Figure 2-3. Column Types

When entering data into a datasheet, the data must conform to the type of column in which it is entered. For example, attempting to type a name into a numeric column will result in it being rejected. When entering data, the format of the data must also match your current Windows settings. In particular, STATGRAPHICS Centurion XVI honors the current Windows settings for:

1. Decimal separator for numeric values
2. Time format and time separator for times
3. Short date format and date separator for dates

To check the settings of your computer, access the Windows *Control Panel*.
When entering a date, you must use the format specified on the *Edit - Preferences* dialog box, either 4-digits years (as in 4/30/2005) or 2-digit years (as in 4/30/05). If a 2-digit year is used, it is assumed to fall within the years 1950 through 2049.

More information about formula columns may be found in a later section of this chapter titled *Manipulating Data*.

2.2 Accessing Data

Chapter 1 showed how data can be entered into a datasheet by hand. More often, users will access data that already exists in another file or application. There are 3 basic ways of putting existing data into a STATGRAPHICS Centurion XVI datasheet:

1. **Read an existing data file**: If the data have previously been entered into a file, you can read it into the datasheet by selecting *File – Open – Open Data Source*. This allows you to read data stored in various file formats, including Excel files, delimited ASCII text files, XML files, STATGRAPHICS files, and files from other statistical packages.

2. **Copy and paste using the Windows clipboard**: If you have the data loaded into a program such as Excel, you can easily copy it to the Windows clipboard and then paste it into STATGRAPHICS Centurion XVI by selecting *Edit – Paste*.

3. **Issue a SQL query to retrieve it from a database:** If the data resides in an ODBC-compatible database, such as Oracle or Microsoft Access, it can be retrieved by selecting *File – Open – Open Data Source* and then selecting *ODBC Query*.

2.2.1 Reading Data from a STATGRAPHICS Centurion Data File

To read data that have already been saved in a STATGRAPHICS Centurion data file, select any of the 26 datasheets in the DataBook by clicking on its tab. Then select *File – Open – Open Data Source* and specify *STATGRAPHICS Data File* on the dialog box shown below:

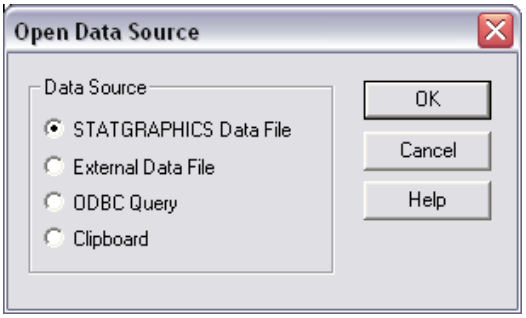


Figure 2-4. Open Data Source Dialog Box

After pressing OK, select the desired STATGRAPHICS file:

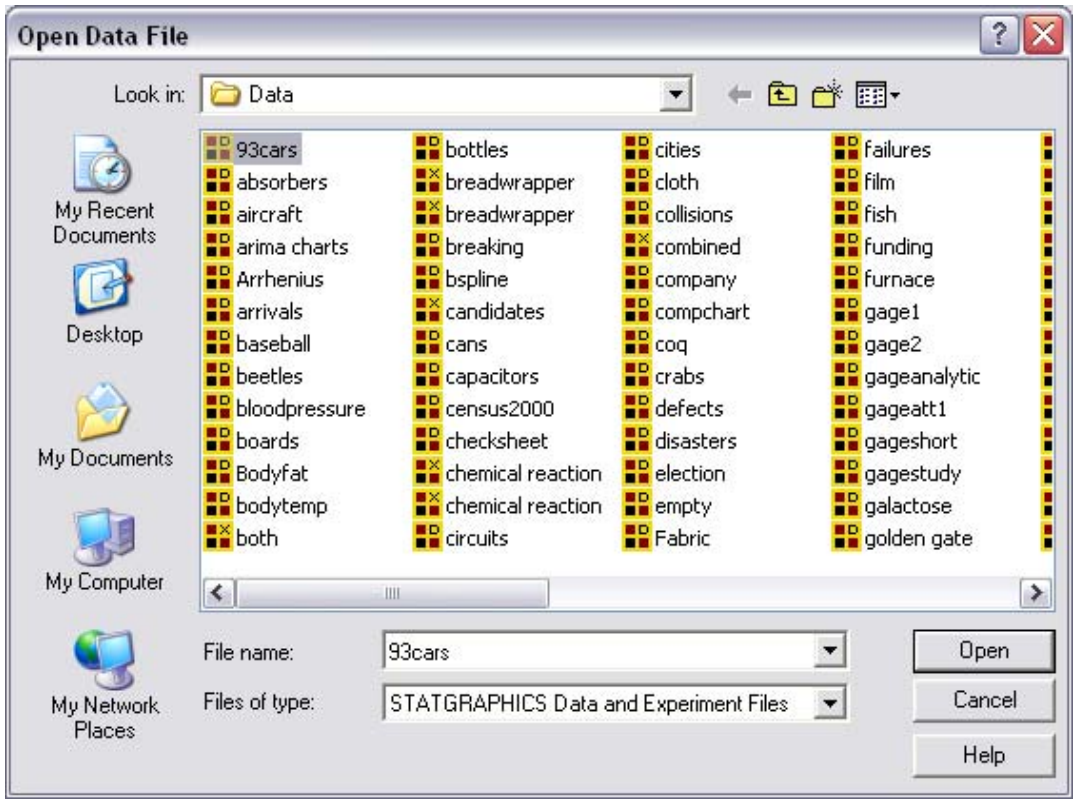


Figure 2-5. Selecting a STATGRAPHICS Data File

You can read data files from STATGRAPHICS Centurion XVI or any previous version of STATGRAPHICS, including STATGRAPHICS *Plus*. The data in the file will replace the contents of the currently selected datasheet.

2.2.2 Reading Data from an Excel, ASCII, XML, or Other External Data File

To read data that have been saved in a data file created by another application, select any of the 26 datasheets in the DataBook by clicking on its tab. Then select *File* – *Open* – *Open Data Source* and specify *External Data File* on the dialog box shown below:

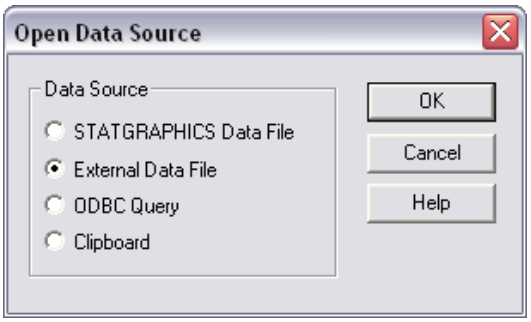


Figure 2-6. Open Data Source Dialog Box

After pressing OK, a dialog box will be displayed on which to specify the file to be imported and other relevant information:

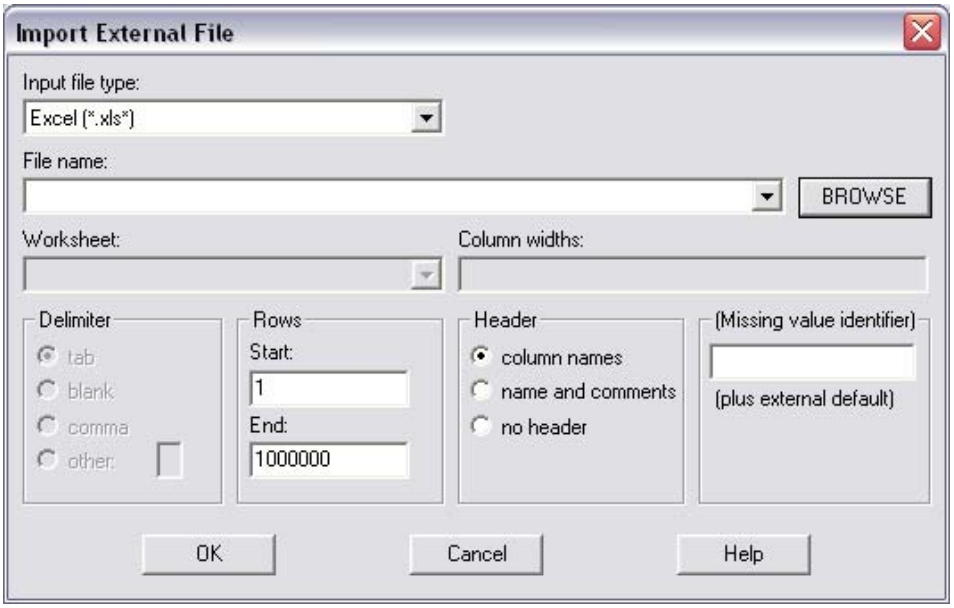


Figure 2-7. Selecting an External Data File

The fields on this dialog box include:

1. **Input file type** – type of file to be imported. STATGRAPHICS Centurion XVI can import data from many other applications, including Excel, Matlab, Minitab, JMP, SPSS, SAS, and many other statistical packages.

2. **File name** – name of the file to be imported. Press the *BROWSE* button to select the desired file.

3. **Worksheet** – name of the worksheet to import (if relevant). Only one sheet may be read at a time.

4. **Column widths** – width of each column, separated by commas (for formatted ASCII files only).

5. **Delimiter** – column delimiter (for delimited ASCII files only).

6. **Rows** - the range of rows within the worksheet that will be read. This range includes the variable names and comments, if present.

7. **Header** - information contained in the first 2 rows of the specified range (for spreadsheet programs such as Excel). The two rows immediately above the data to be read may contain column names and/or comments. If names are not contained in the file, then default names will be generated.

8. **Missing value identifier** - any special symbol used in the external file to indicate missing data, such as *NA*. Cells containing the specified value will be converted to empty cells when placed in the STATGRAPHICS Centurion XVI datasheet.

When *OK* is pressed, the data from the external file will be read into STATGRAPHICS Centurion XVI . Each column will be scanned and an appropriate column type assigned to it. The data are then ready to be analyzed.

2.2.3 Transferring Data Using Copy and Paste

The easiest way to transfer data from another application to STATGRAPHICS Centurion XVI is often via the Windows clipboard. For example, if data reside in an Excel file, Excel may be started and the data copied to the clipboard by selecting the desired data within Excel and then choosing *Copy* from the Excel *Edit* menu. Upon returning to STATGRAPHICS, the data may be pasted directly into a STATGRAPHICS Centurion XVI datasheet by selecting *Paste* from the STATGRAPHICS *Edit* menu. When data is pasted into a column of a datasheet,

STATGRAPHICS Centurion XVI automatically scans the data and selects an appropriate type for the column.

When copying and pasting data, column names and comments may also be transferred. Include the column names and comments in Excel when copying the data to the clipboard. On the STATGRAPHICS Centurion XVI side, click in the header row of the STATGRAPHICS Centurion XVI datasheet before selecting *Paste*. The information at the top of the clipboard will then be pasted into the header row(s).

2.2.4 Querying an ODBC Database

STATGRAPHICS Centurion XVI also allows you to read data from an Oracle, Access, or other database using ODBC. To access data from a database, first select *File – Open – Open Data Source*. Then select *Query Database* from the initial dialog box:

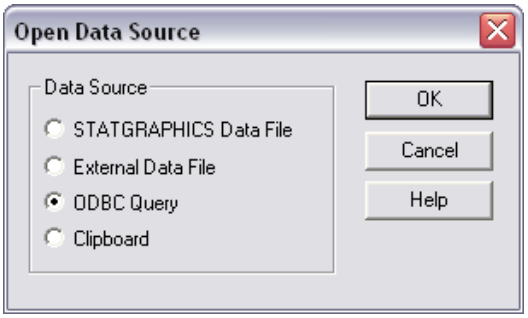


Figure 2-8. Open Data Source Dialog Box

A sequence of additional dialog boxes will be displayed on which you:

1. Select the name of the database to be read.
2. Select the fields to be transferred.
3. Specify a filter to limit the records that are retrieved.
4. Specify a sort order for the results.

A SQL query is then constructed and the results placed in the active STATGRAPHICS Centurion XVI datasheet. Detailed information on constructing ODBC queries may be found in the PDF document titled *Data Files and StatLink*.

2.3 Manipulating Data

Once data have been placed into a STATGRAPHICS Centurion XVI datasheet, it can be manipulated in several important ways:

1. The data may be copied and pasted into other locations.

2. Additional columns may be created from existing columns.

3. Data may be transformed using an algebraic expression or mathematical function.

4. The datasheet may be sorted according to one or more columns.

5. Data values may be recoded to form groups or for other reasons.

6. Data extending over multiple columns can be rearranged into a single column if required by a statistical procedure.

These important operations are described below.

2.3.1 Copying and Pasting Data

The STATGRAPHICS Centurion XVI datasheet supports many typical spreadsheet operations, including *cut*, *copy*, *paste*, *insert*, and *delete*. The one important fact to remember when using these operations is that every column has a specified type. If you inadvertently paste character data into a numeric column, STATGRAPHICS Centurion XVI will change the type of that column to accommodate the new data. If you ever have any doubt about a column's type, click on the column header to display the *Modify Column* dialog box. You can change the type of the column using that dialog box.

2.3.2 Creating New Variables from Existing Columns

STATGRAPHICS Centurion XVI has a wide array of operators to assist in performing mathematical calculations. One of the most important uses of these operators in data analysis is to create new variables based on existing columns. In STATGRAPHICS Centurion XVI, new variables may be created:

1. "On-the-fly" directly within the data fields on data input dialog boxes, without saving the variable in the datasheet.

2. By creating a new column in any of the 26 datasheets in the DataBook.

For example, suppose information was desired about the ratio of miles per gallon in city driving versus miles per gallon in highway driving for each automobile in the *93cars* data file. That file contains 2 separate columns, one named *MPG City* and one named *MPG Highway*. To summarize the distribution of the ratios, you could select the *One-Variable Analysis* procedure and specify the ratio directly in the *Data* field of the data input dialog box:

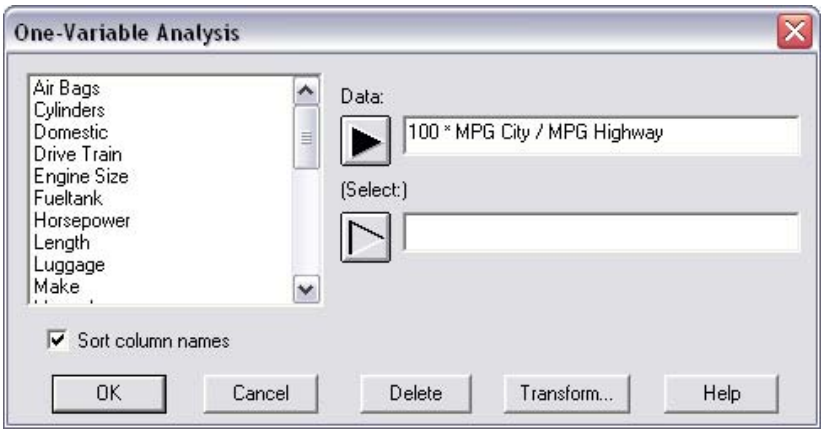


Figure 2-9. Creating a Transformation "On-The-Fly"

When OK is pressed, an analysis will be generated for 100 times the ratio, without ever changing the data in the datasheet:

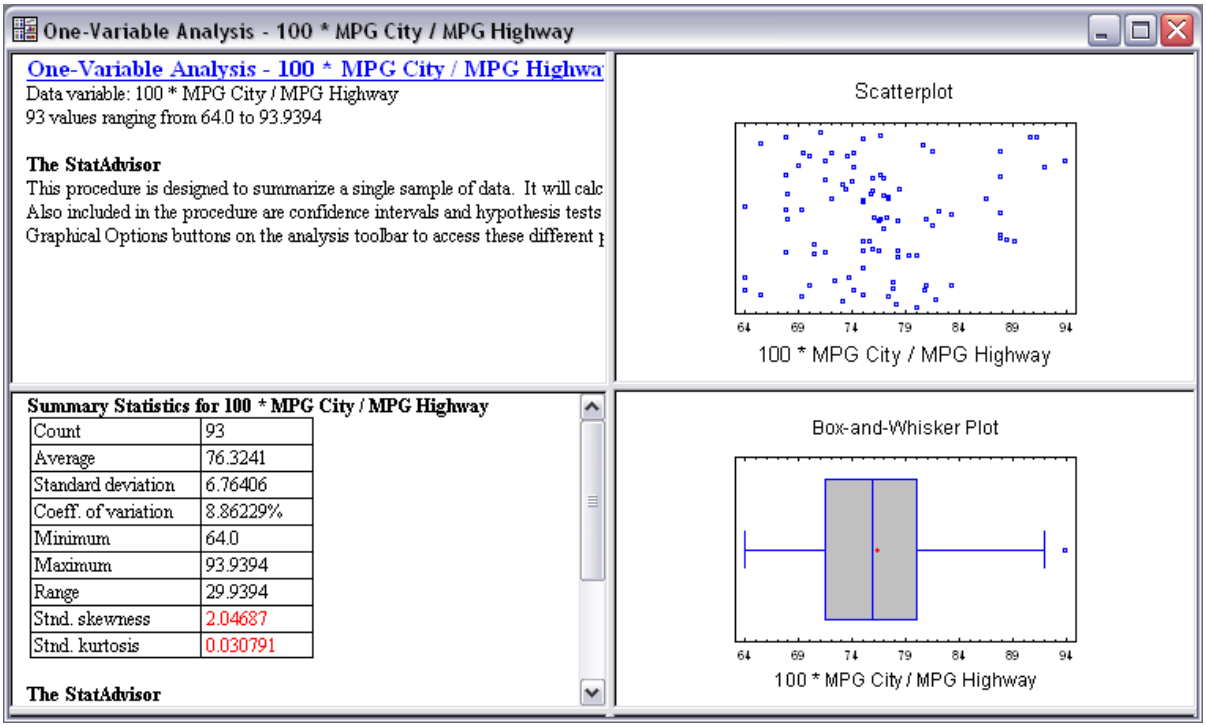


Figure 2-10. One-Variable Analysis of Transformed Data

The average ratio is approximately 76.3%, ranging from a low of 64.0% to a high of 93.9%. The ability to do analyses without modifying the datasheets is very important in facilitating the exploration of data.

If desired, a new column could be created in a datasheet containing the transformed values. For example, you could return to the window containing the *93cars* data and double-click on the column header labeled *Col_27*. The *Modify Column* dialog box could then be used to define a new variable of type *formula* with the desired transformation:

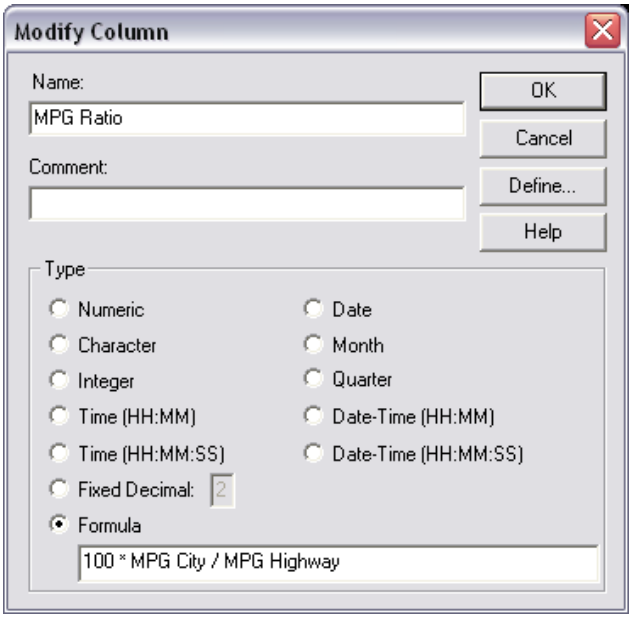
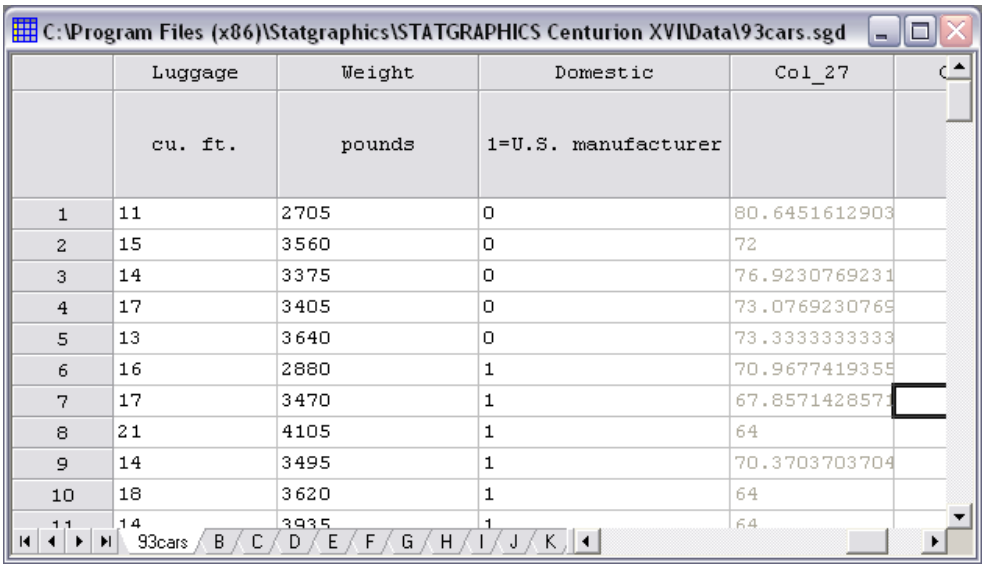


Figure 2-11. Creating a Formula Column

This will create a new column whose values are calculated from the original two columns containing the miles per gallon data. *Formula* columns are displayed in the datasheet using a gray scale, since they are automatically calculated from other columns:



	Luggage	Weight	Domestic	Col_27
	cu. ft.	pounds	1=U.S. manufacturer	
1	11	2705	0	80.6451612903
2	15	3560	0	72
3	14	3375	0	76.9230769231
4	17	3405	0	73.076923076
5	13	3640	0	73.3333333333
6	16	2880	1	70.9677419355
7	17	3470	1	67.857142857
8	21	4105	1	64
9	14	3495	1	70.3703703704
10	18	3620	1	64
11	14	3935	1	64

Figure 2-12. Appearance of a Formula Column in a Datasheet

If the values in the *MPG City* or *MPG Highway* columns change, *MPG Ratio* will be automatically recalculated to reflect those changes.

> NOTE: Recalculation of formula columns does not normally occur until the data in those columns is needed for a calculation or is saved or printed. You can specify a recalculation to occur immediately by selecting *Update Formulas* from the *Edit* menu.

2.3.3 Transforming Data

STATGRAPHICS Centurion XVI also contains a large number of mathematical functions that may be used to transform existing data. As when creating new variables, transformations may be done either directly within fields of a data input dialog box or by creating new columns in a datasheet.

For example, suppose it was desired to plot the miles per gallon that an automobile obtained versus the natural logarithm of vehicle weight. Selecting the *X-Y Plot* procedure from the main menu displays the following data input dialog box:

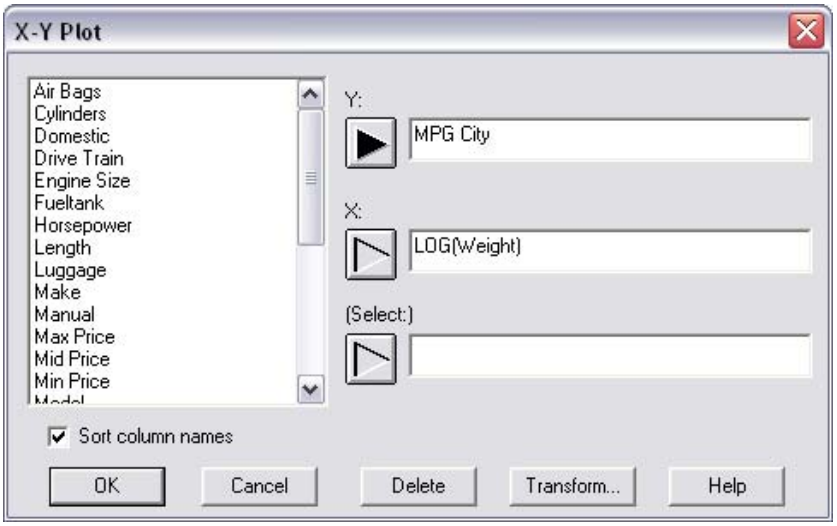


Figure 2-13. Transforming Data on a Data Input Dialog Box

Instead of typing the name of a column in a data field, you may type a STATGRAPHICS Centurion expression. STATGRAPHICS Centurion expressions are formulas that operate on data using algebraic symbols and special operators. A wide variety of operators are available, as

described in the PDF document titled *STATGRAPHICS Operators*. The table below shows commonly used operators:

Operator	*Use*	*Example*
+	Addition	X+100
-	Subtraction	X-100
/	Division	X/100
*	Multiplication	X*100
^	Exponentiation	X^2
ABS	Absolute value	ABS(X)
AVG	Average	AVG(X)
DIFF	Backward differencing	DIFF(X)
EXP	Exponential function	EXP(10)
LAG	Lag by k periods	LAG(X,k)
LOG	Natural logarithm	LOG(X)
LOG10	Log base 10	LOG10(X)
MAX	Maximum	MAX(X)
MIN	Minimum	MIN(X)
SD	Standard deviation	SD(X)
SQRT	Square root	SQRT(X)
STANDARDIZE	Conversion to Z-scores	STANDARDIZE(X)

Figure 2-14. Commonly Used STATGRAPHICS Operators

When constructing a STATGRAPHICS Centurion expression, multiple operators may be combined using normal algebraic precedence rules. For example, the following expression converts each value in the column named *Weight* to a fraction equal to the distance between the minimum and maximum values amongst all of the automobiles:

(*Weight* – MIN(*Weight*)) / (MAX(*Weight*) - MIN(*Weight*))

The parentheses are necessary to insure that the subtractions are done before the division. Expressions are not case sensitive, nor is the inclusion of blank spaces relevant.

Every data input dialog box includes a button labeled *Transform*, as in *Figure 2-13*. This button may be used to help create STATGRAPHICS Centurion expressions, if you do not remember which operators to use. If you place the cursor in a data field and then press *Transform*, a dialog box similar to that shown below will be displayed:

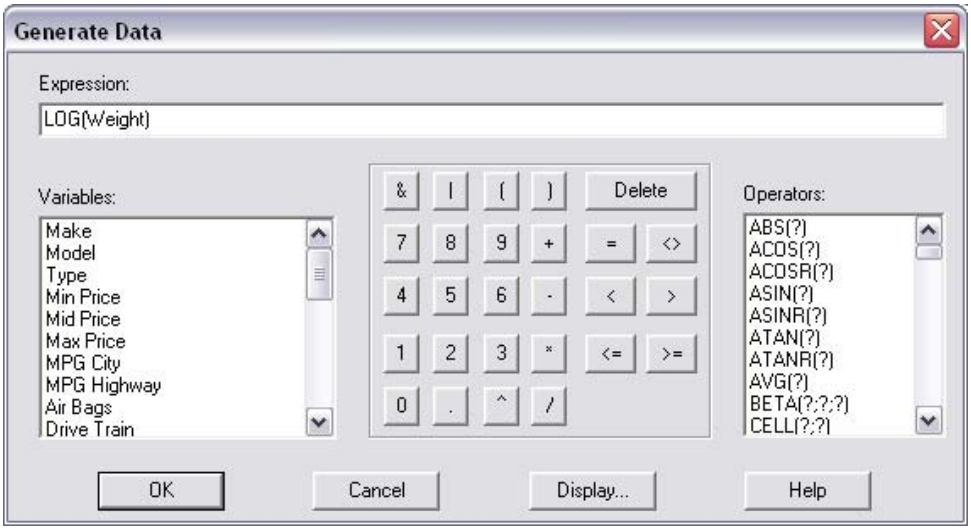


Figure 2-15. Dialog Box Displayed by the Transform Button

Along the right is a list of all STATGRAPHICS Centurion operators, with an indication of the number of arguments that must be supplied. Clicking on an operator name places it in the *Expression* field. After you replace the question marks with column names or numbers, you may press the *Display* button to see the first several values generated by the expression, or press the *OK* button to have the expression entered into the data input dialog box.

> NOTE: You do not need to use the *Transform* button if you would rather type the expression yourself on the data input dialog box.

Once a transformation has been specified on the data input dialog box, as in *Figure 2-13*, that transformation will be used when the procedure is run:

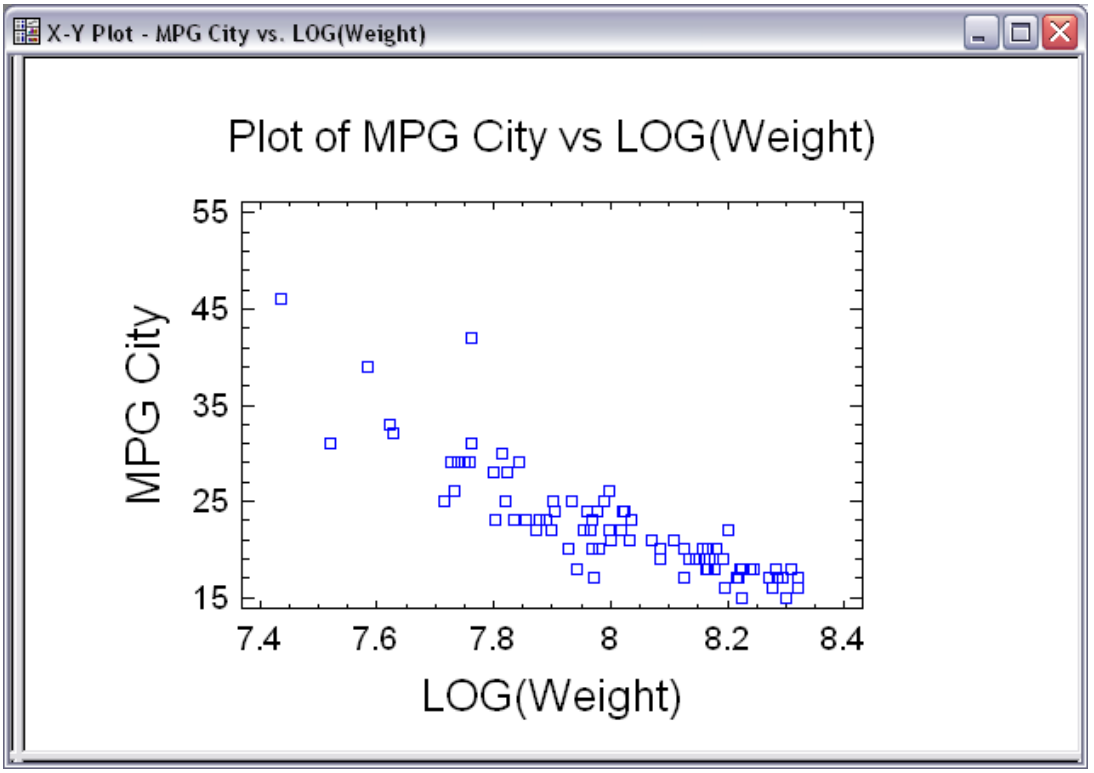


Figure 2-16. X-Y Plot Procedure Using Transformed values of Weight

STATGRAPHICS Centurion operators may also be used when creating *formula* columns, similar to the illustration in the preceding section.

2.3.4 Sorting Data

The contents of a datasheet may be sorted by highlighting the column or columns to be used to define the sort order and then selecting *Sort Data* from the *Edit* menu. For example, to sort the data in the *93cars* file according to miles per gallon, highlight the columns named *MPG City* and *MPG Highway* and then select *Sort Data*. The following dialog box will be displayed:

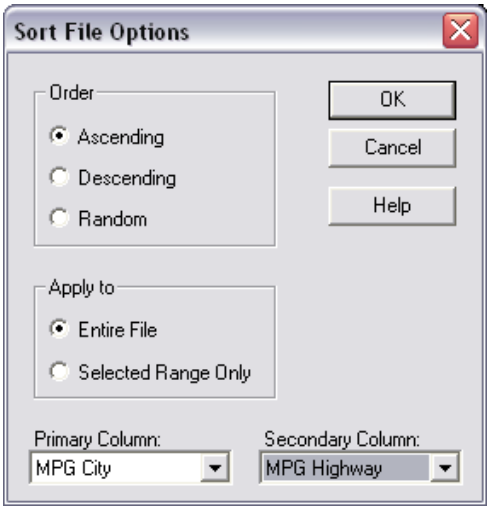


Figure 2-17. Sort Options Dialog Box

You may specify either one or two columns on which to base the sort, as well the sort order. Sorting by *MPG City* and then *MPG Highway* sorts first by miles per gallon in city driving and then, for automobiles with the same value of *MPG City*, by miles per gallon in highway driving:

C:\Program Files (x86)\Statgraphics\STATGRAPHICS Centurion XVI\Data\93cars.sgd

	Mid Price	Max Price	MPG City	MPG Highway	
	average of min and max prices in $1,000	price for a premium version in $1,000	miles per gallon in city driving	miles per gallon in highway driving	1
1	16.6	18.6	15	20	0
2	19.9	25.3	15	20	1
3	23.7	24.9	16	25	1
4	34.7	36.3	16	25	1
5	40.1	42.7	16	25	2
6	19	24.4	17	21	1
7	19.7	22.7	17	21	0
8	47.9	50.4	17	22	1
9	19.1	21.5	17	23	0
10	38	41.5	17	25	1
11	32.5	32.5	17	25	1
12	18.8	19.6	17	26	1
13	34.3	35.3	17	26	2
14	22.7	26.6	18	22	1

93cars B C D E F G H I J K

Figure 2-18. 93cars.sgd File after Sorting

NOTE: The statistical procedures do not require you to sort the data before using them, since they will automatically sort the data if necessary. Also, the data file on disk is not changed when you perform a sort unless you resave the data. Sorting only affects the order in which the rows are displayed in the datasheet.

2.3.5 Recoding Data

It is sometimes convenient to recode data, either by grouping it into similar groups or by assigning new labels. To recode a column of data, first click on the header of the column to be recoded. Then select *Recode Data* from the *Edit* menu. The following dialog box will be displayed:

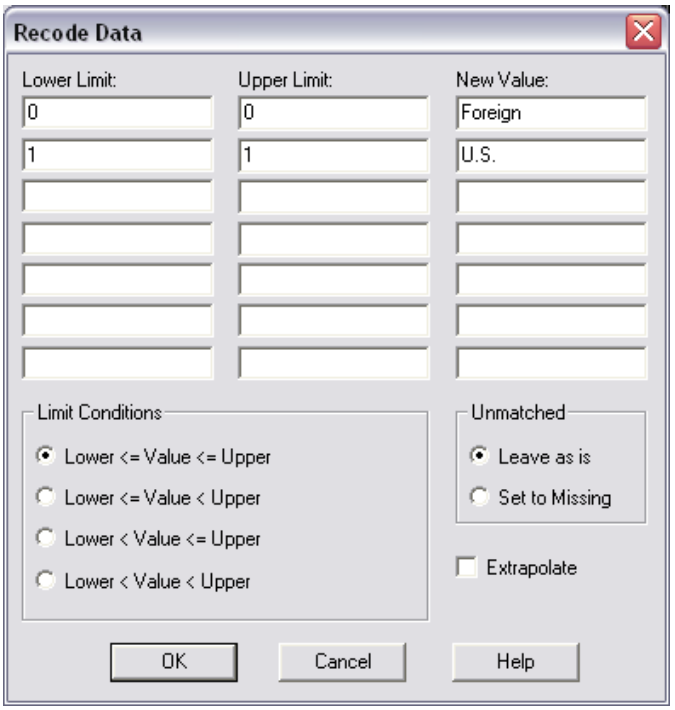


Figure 2-19. Dialog Box for Recoding Data

For example, the column named *Domestic* in the *93cars* file contains a 1 for each car made by a U.S. automaker and a 0 for all other cars. To change all 0's in the column to "Foreign" and all 1's to "U.S.", the dialog box above could be used. Up to 7 ranges of values may be specified at one time for recoding.

The PDF document titled *Edit Menu* has a detailed discussion of two recoding examples.

2.3.6 Combining Multiple Columns

Many statistical procedures in STATGRAPHICS Centurion XVI expect the data to be analyzed to be in a single column. Sometimes data is not arranged in such a format. As a simple example, suppose you have a sample of 12 observations, arranged into 4 columns as follows:

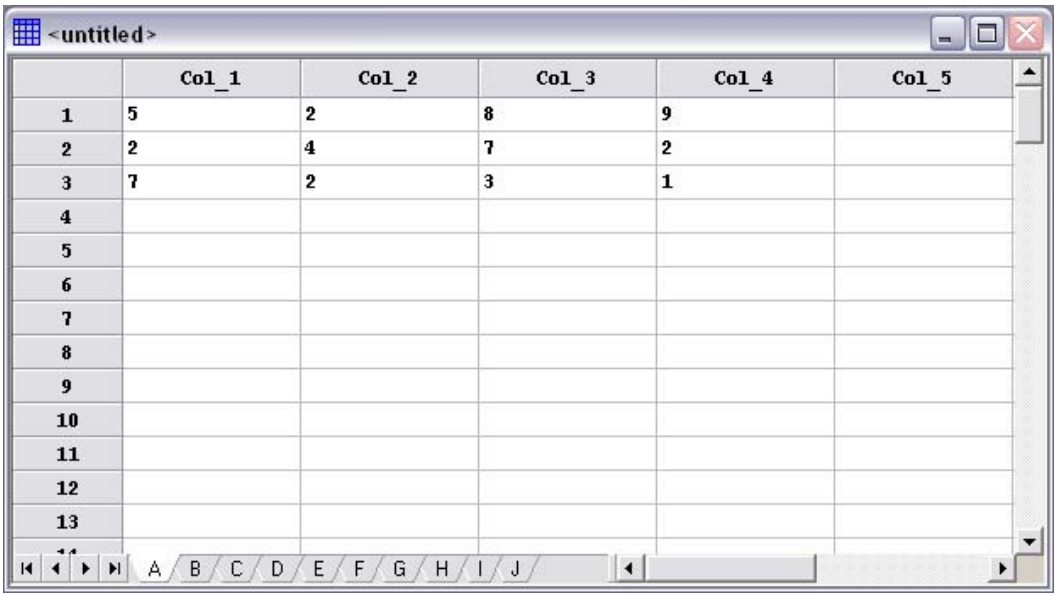


Figure 2-20. Sample Data in Multiple Columns

To place this data in a single column, multiple copy and paste operations could be performed. A simpler solution is to use the *Combine columns* procedure, found under *Edit* on the main menu. This procedure first presents a data input dialog box requesting the names of the columns containing the data:

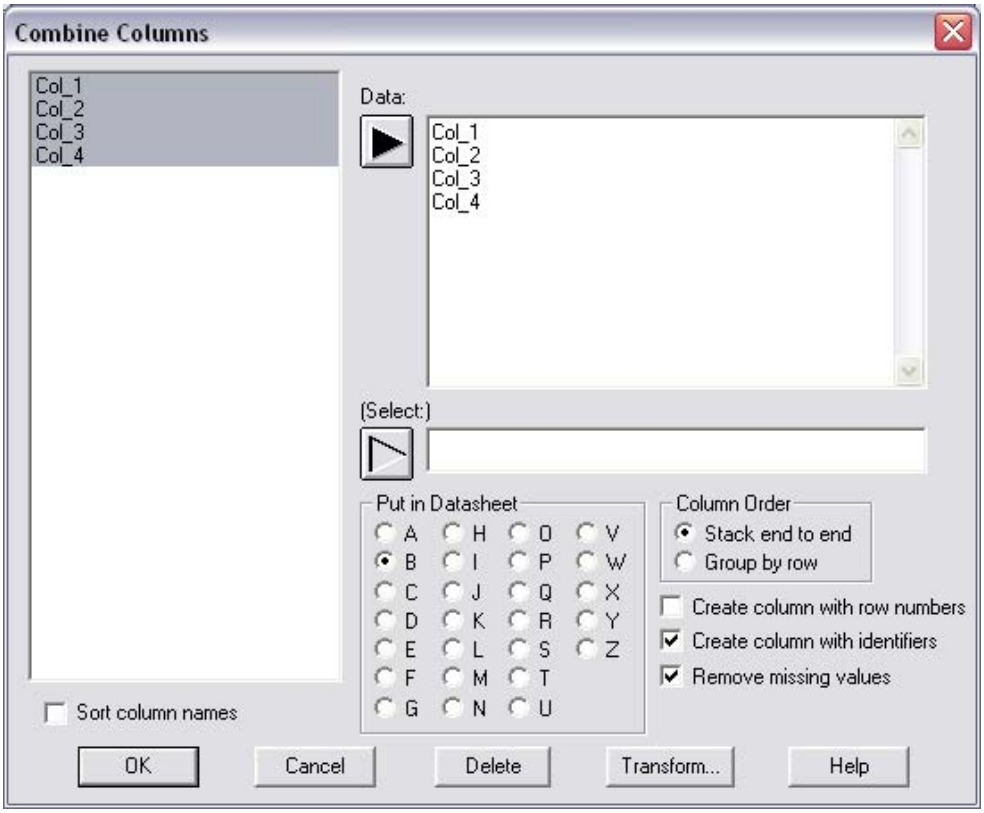


Figure 2-21. Data Input Dialog Box for Combine Columns

It contains the following fields:

1. **Data** – columns to be combined.

2. **Select** – standard subset selection field to choose a subset of the rows.

3. **Put in Datasheet** – target data sheet for the combined data.

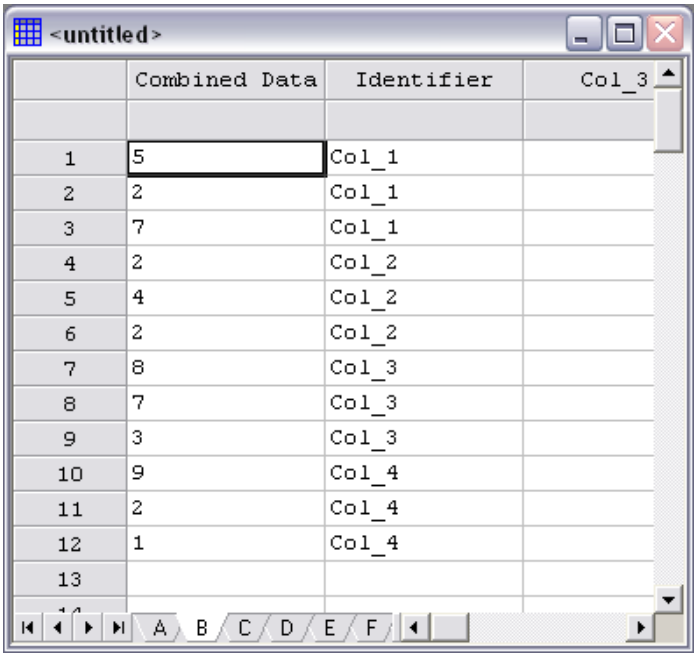
4. **Column Order** – whether to stack the data by column (one column after the other) or by row.

5. **Create column with row numbers** – whether to create a second column identifying the original row containing each data value.

6. **Create column with identifiers** – whether to create a second column identifying the original column containing each data value.

7. **Remove missing values** – whether to skip all blank cells rather than leaving a placeholder.

When OK is pressed, the data are combined into a single column as shown below:



	Combined Data	Identifier	Col_3
1	5	Col_1	
2	2	Col_1	
3	7	Col_1	
4	2	Col_2	
5	4	Col_2	
6	2	Col_2	
7	8	Col_3	
8	7	Col_3	
9	3	Col_3	
10	9	Col_4	
11	2	Col_4	
12	1	Col_4	
13			

A B C D E F

Figure 2-22. Data Combined into Single Column

2.4 Generating Data

STATGRAPHICS Centurion XVI has the ability to generate data and place it in columns of a datasheet. This section describes two important examples:

1. Generating data with simple patterns.

2. Generating random numbers.

2.4.1 Generating Patterned Data

Several procedures in STATGRAPHICS Centurion XVI, particularly those that perform an analysis of variance, expect the data to be analyzed to be placed into a single column of the datasheet, together with one or more code columns identifying the explanatory factors. For example, consider the data in the following two-way table:

Blend	*Treatment 1*	*Treatment 2*	*Treatment 3*
1	75	82	91
2	78	85	93
3	77	84	92
4	75	85	96

To analyze this data using the *Multifactor ANOVA* procedure, it needs to be placed into a datasheet in the following format:

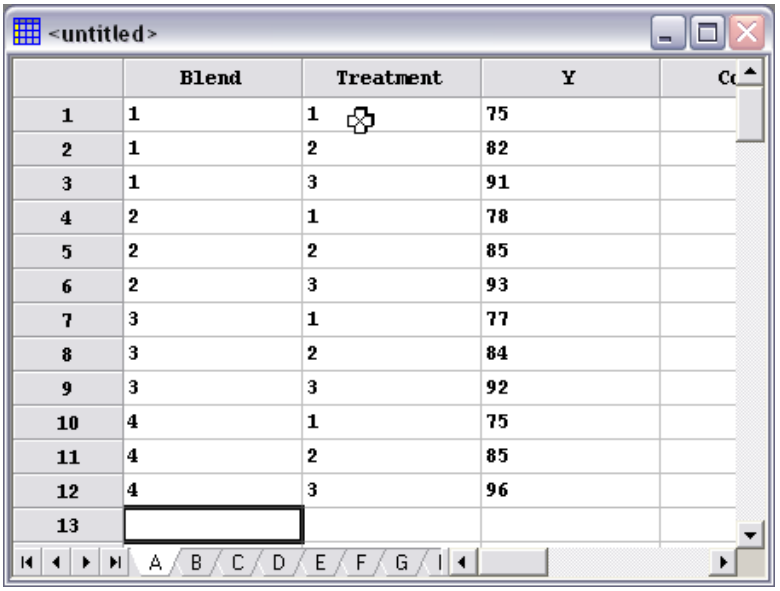


Figure 2-23. Desired Data Structure

The first two columns indicate the levels of the factors corresponding to each data value. The third column contains all of the observations.

To create such a file, the easiest solution is often to type in the first two columns. However, since the columns follow simple patterns, you could generate them instead using special STATGRAPHICS Centurion operators. For example, the blend numbers can be generated by

clicking on the column #1 header and then selecting *Generate Data* from the *Edit* menu. This displays the following dialog box, into which an expression has been entered:

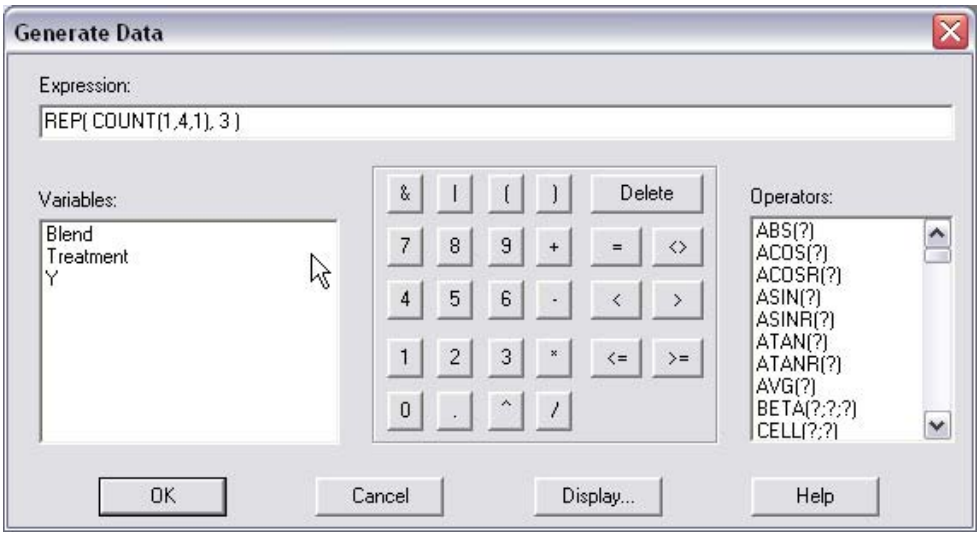


Figure 2-24. Generating Blend Numbers

The *Generate Data* option evaluates a STATGRAPHICS Centurion expression and places the result into the selected column. In the expression shown above, two important operators are used:

> *COUNT(from, to, by)* – generates values beginning at *from* and ending at *to*, at intervals equal to *by*. COUNT(1,4,1) thus generates the integers 1, 2, 3, and 4.
>
> *REP(X, repetitions)* – repeats each value in X *repetitions* times, in groups. In this case, each integer between 1 and 4 is repeated 3 times.

The treatment numbers can be generated in a similar manner by clicking on the column #2 header, selecting *Generate Data* from the *Edit* menu, and entering the following:

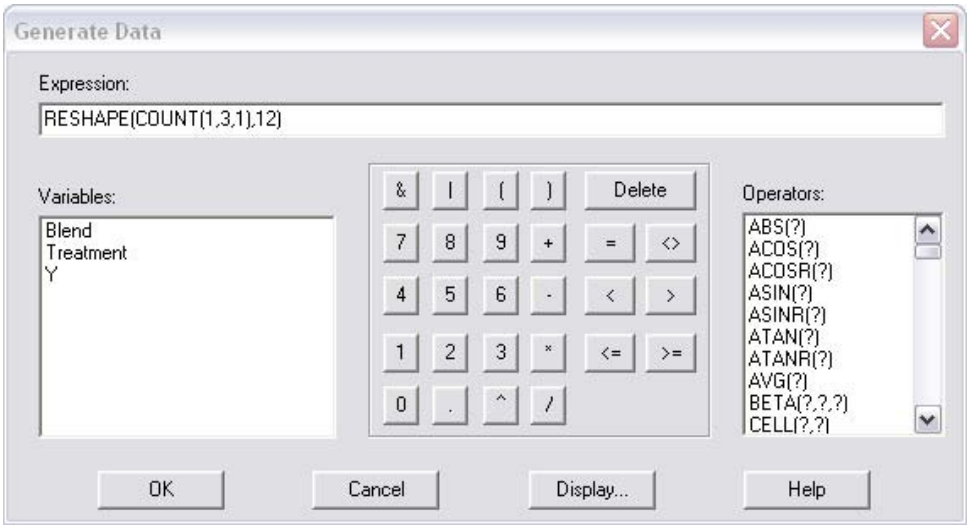


Figure 2-25. Generating Treatment Numbers

This expression uses an additional operator:

> *RESHAPE(X, size)* – repeats the values in X in a circular fashion until *size* values have been generated. In this case, the sequence 1, 2, 3 is repeated 4 times.

These pattern generators can be helpful when the data file to be created is large.

2.4.2 Generating Random Numbers

Random numbers may be generated in STATGRAPHICS Centurion XVI in two ways:

1. If the numbers come from an exponential, gamma, lognormal, normal, uniform, or Weibull distribution, they may be generated within a datasheet by clicking on a column header, selecting *Generate Data* from the *Edit* menu, and entering the appropriate STATGRAPHICS Centurion expression.

2. For other distributions, the random numbers must be generated from within the *Probability Distributions* procedure.

As an example, suppose 100 random numbers are desired from a normal distribution with a mean of 20 and a standard deviation equal to 2. Click on the header of an empty column in any datasheet to select that column. Then select *Generate Data* from the *Edit* menu and complete the dialog box as shown below:

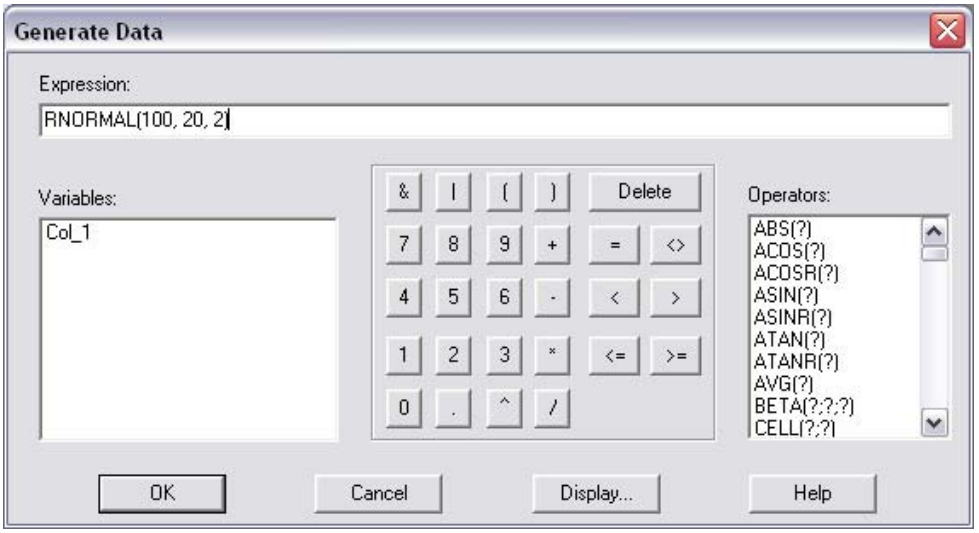


Figure 2-26. Generating Random Numbers from a Normal Distribution

The syntax of the RNORMAL operator is:

> *RNORMAL*(*n*, *mu*, *sigma*) – generates *n* pseudo-random numbers from a normal distribution with mean *mu* and standard deviation *sigma*.

Press *OK* to generate the random numbers and place them into the selected column.

The syntax of the other random number generators is contained in the PDF document titled *STATGRAPHICS Centurion Operators*.

2.5 DataBook Properties

This chapter has described many important aspects of data handling within STATGRAPHICS Centurion XVI. In particular, it has shown how to read data from files and databases and how to manipulate that data once it has been placed in a STATGRAPHICS Centurion XVI datasheet. At any given time, the status of the datasheets may be displayed by activating the DataBook window and selecting *DataBook Properties* from the *Edit* menu or by selecting *StatLink* from the *File* menu:

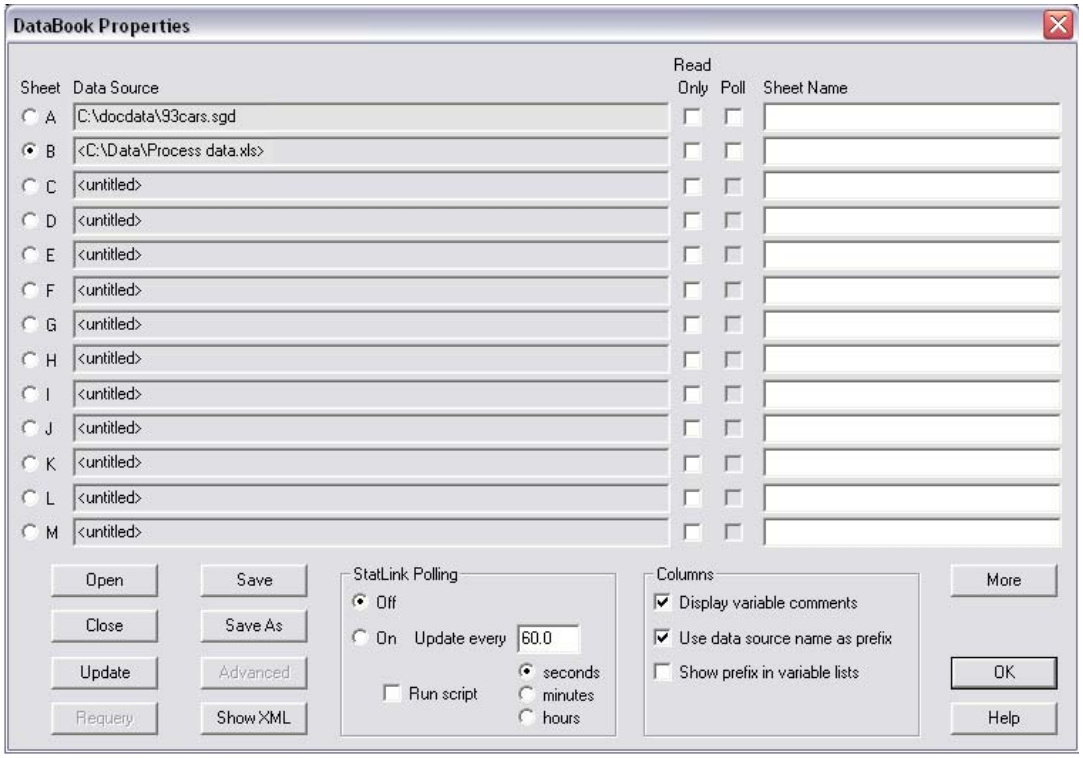


Figure 2-27. DataBook Properties Dialog Box

This dialog box shows the current source of the data within each datasheet. If desired, datasheets may be made read-only so that data in them cannot be changed inadvertently. It is also possible to poll the data source (reread it) at regular intervals and have the statistical procedures update automatically. These important features are described in Chapter 5.

2.6 Data Viewer

A new procedure has been added to view data files in STATGRAPHICS Centurion XVI. The procedure, accessed by selecting *Data Viewer* from the *Tools* menu, produces a summary of the number of nonmissing and unique values and the minimum and maximum values of any variables selected:

DataViewer

Number of columns: 26
Number of rows: 93
Number of complete cases: 82

Column			*Nonmissing*	*Unique*		
Name	*Comment*	*Type*	*Values*	*Values*	*Minimum*	*Maximum*
Air Bags	0=none, 1=driver only, 2=driver and passenger	Numeric	93	3	0	2
Cylinders		Numeric	92	5	3.0	8.0
Domestic	1=U.S. manufacturer	Numeric	93	2	0	1
Drive Train		Character	93	3		
Engine Size	liters	Numeric	93	25	1.0	5.7
Fueltank	gallons	Numeric	93	38	9.2	27.0
Horsepower	maximum	Numeric	93	57	55.0	300.0
Length	inches	Numeric	93	51	141.0	219.0
Luggage	cu. ft.	Numeric	82	16	6.0	22.0
Make		Character	93	32		
Manual	0=no, 1=yes	Numeric	93	2	0	1
Max Price	price for a premium version in $1,000	Numeric	93	79	7.9	80.0
Mid Price	average of min and max prices in $1,000	Numeric	93	81	7.4	61.9
Min Price	price for basic version in $1,000	Numeric	93	79	6.7	45.4
Model		Character	93	93		
MPG City	miles per gallon in city driving	Numeric	93	21	15.0	46.0
MPG Highway	miles per gallon in highway driving	Numeric	93	22	20.0	50.0
Passengers	persons	Numeric	93	6	2.0	8.0
Rear seat	inches	Numeric	91	24	19.0	36.0
Revs per Mile	revs per mile in highest gear	Numeric	93	78	1320.0	3755.0
RPM	revs per minute at maximum horsepower	Numeric	93	24	3800.0	6500.0
Type		Character	93	6		
U Turn Space	feet	Numeric	93	14	32.0	45.0
Weight	pounds	Numeric	93	81	1695.0	4205.0
Wheelbase	inches	Numeric	93	27	90.0	229.0
Width	inches	Numeric	93	16	60.0	78.0

Figure 2-28. DataBook Properties Dialog Box

Chapter

3

Running Statistical Analyses

Generating an analysis, selecting additional tables and graphs, selecting options, changing the input data, and saving the results.

There are over 160 statistical procedures on the main STATGRAPHICS Centurion XVI menu. Each selection accesses a different statistical procedure. All procedures, however, work in the same basic way:

1. When an analysis is selected from the menu, a *data input dialog box* is displayed. The fields on this dialog box are used to specify the variables to be analyzed.

2. If the selected procedure has options that affect all tables and graphs within the procedure, an *Analysis Options* dialog box is displayed to select the desired settings.

3. If the selected procedure has more than just a single table and a single graph, a *Tables and Graphs* dialog box is displayed on which the desired output can be selected.

4. The specified data is then read and analyzed, and a new *analysis window* is created.

5. The selected options can be changed using the *Analysis Options* button on the analysis toolbar, in response to which all tables and graphs in the analysis window will be updated.

6. If desired, additional tables and graphs may be requested by pressing the *Tables and Graphs* button on the analysis toolbar.

7. Individual tables and graphs can be modified by maximizing the corresponding pane and selecting *Pane Options* from the analysis toolbar.

8. For graphs, the default title, scaling, point types, fonts, etc. may be changed by double-clicking on the graph to maximize it and then selecting *Graphics Options* from the analysis toolbar.

9. Tables and graphs may be printed, published as HTML files, copied to other applications such as Microsoft PowerPoint, or saved in the StatReporter.

10. Numerical results may be saved to columns of any datasheet using the *Save Results* button on the analysis toolbar.

11. The entire analysis may be saved to disk as a *StatFolio* for later retrieval.

In this chapter, a typical analysis is described in detail. The goal of the analysis is to construct a statistical model relating the miles per gallon achieved in city driving for the $n = 93$ automobiles in the *93cars.sgd* data file to their weight. A scatterplot of the data is shown below:

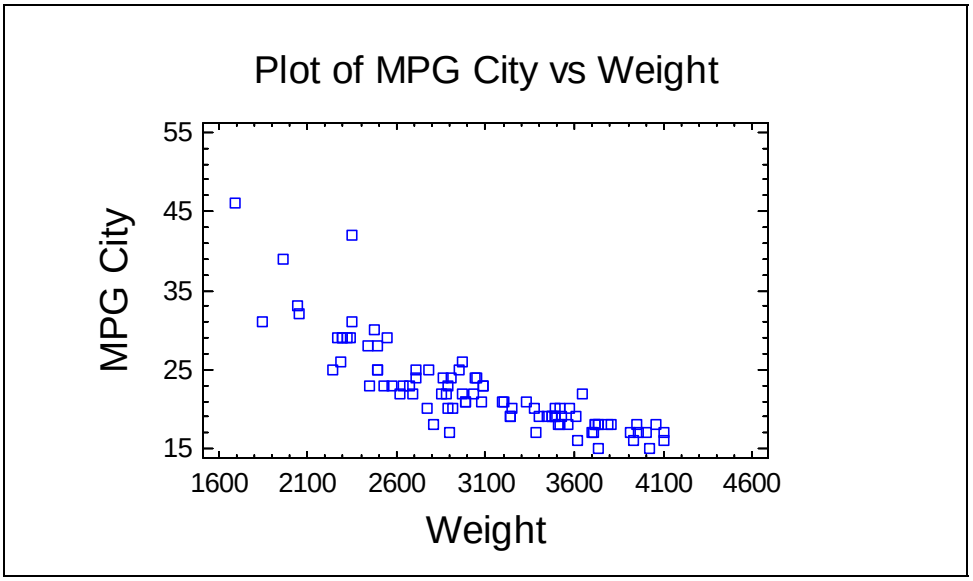


Figure 3-1. X-Y Plot of Miles per Gallon in City Driving versus Weight in Pounds

As might be expected, miles per gallon is negatively correlated with vehicle weight. Some non-linearity is evident in the relationship, and at least one point appears to be a potential outlier.

The primary procedure in STATGRAPHICS Centurion XVI for fitting a statistical model relating two variables is the *Simple Regression* procedure. That procedure fits both linear and nonlinear models. The simplest model relating one dependent variable Y to one independent variable X is a straight line of the form

$$Y = a + b X$$

where *b* equals the slope of the line and *a* equals the Y-intercept. Curvilinear models such as the exponential model

$$Y = exp(a + b X)$$

may be used if the relationship is not linear.

3.1 Data Input Dialog Boxes

The *Simple Regression* procedure is located on the main menu:

1. If using the Classic menu, under *Relate – One Factor*.
2. If using the Six Sigma menu, under *Improve – Regression Analysis – One Factor*.

It begins by displaying a typical data input dialog box:

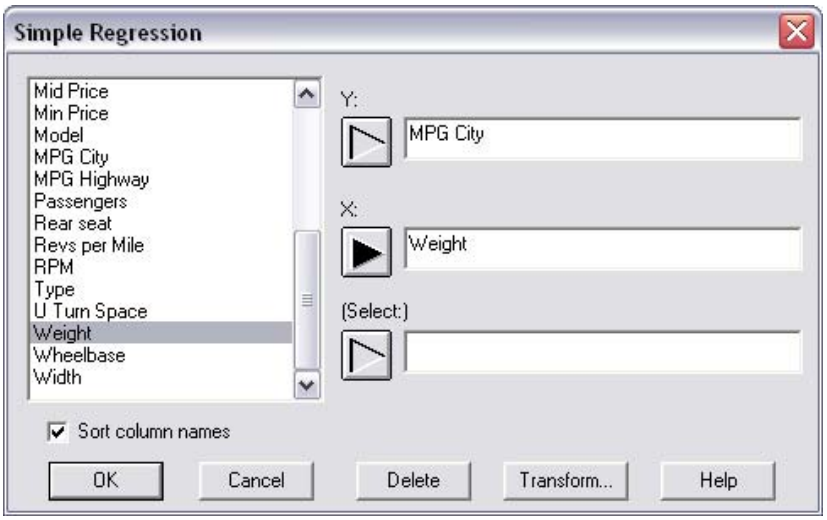


Figure 3-2. Simple Regression Data Input Dialog Box

The first two input fields are required:

Y: The dependent or response variable.

X: The independent or predictor variable.

In data entry fields, you can enter either the name of a column such as *MPG City* or a STATGRAPHICS Centurion expression such as *LOG(MPG City)*. If more than one datasheet contains a column with the indicated name, you must precede the name with an indication of the desired datasheet. For example, if both datasheets A and B contained a column named *Weight* and you wanted to use the column in datasheet A, you would have to enter the name as *A.Weight*.

The *Select* field may be used to select a subset of the rows in the datasheet. For example, if you enter a statement such as *FIRST(50)* in that field, only the first 50 rows in the datasheet will be used. Typical entries in the *Select* field are:

Entry	*Use*	*Example*
FIRST(k)	Selects the first *k* rows.	FIRST(50)
LAST(k)	Selects the last *k* rows.	LAST(50)
ROWS(start,end)	Selects rows between *start* and *end*, inclusive.	ROWS(21,70)
RANDOM(k)	Selects a random set of *k* rows.	RANDOM(50)
column < value	Selects only rows for which *column* is less than *value*.	Passengers < 5
column <= value	Selects only rows for which *column* is less than or equal to *value*.	Passengers <= 5
column > value	Selects only rows for which *column* is greater than *value*.	Passengers > 5
column >= value	Selects only rows for which *column* is greater than or equal to *value*.	Passengers >= 5
column = value	Selects only rows for which *column* equals *value*.	Cylinders = 6
column <> value	Selects only rows for which *column* does not equal *value*.	Cylinders <> 4
condition1 & condition2	Selects only rows that meet both conditions.	Cylinders = 6 & Make = "Ford"
condition1 | condition2	Selects only rows that meet at least one of the conditions.	Cylinders = 6 | Make = "Ford"
binarycolumn	Selects only rows for which the value in *binarycolumn* does not equal 0.	Domestic

Figure 3-3. Allowable Entries for the Select field

When specifying a condition involving a non-numeric variable, *value* must be enclosed in double quotes and *is* case-sensitive. Multiple conditions may be combined using the conditional AND (&) and OR (|) symbols.

Each of the allowable entries in the *Select* field actually generates a sequence of Boolean 0's and 1's, where 0 represents *FALSE* and 1 represents *TRUE*. When used in the *Select* field of a data input dialog box, the result is the selection of all rows for which the condition is *TRUE* and the exclusion of all rows for which the condition is *FALSE*.

3.2 Analysis Windows

Once the data have been specified, a new *analysis window* is created:

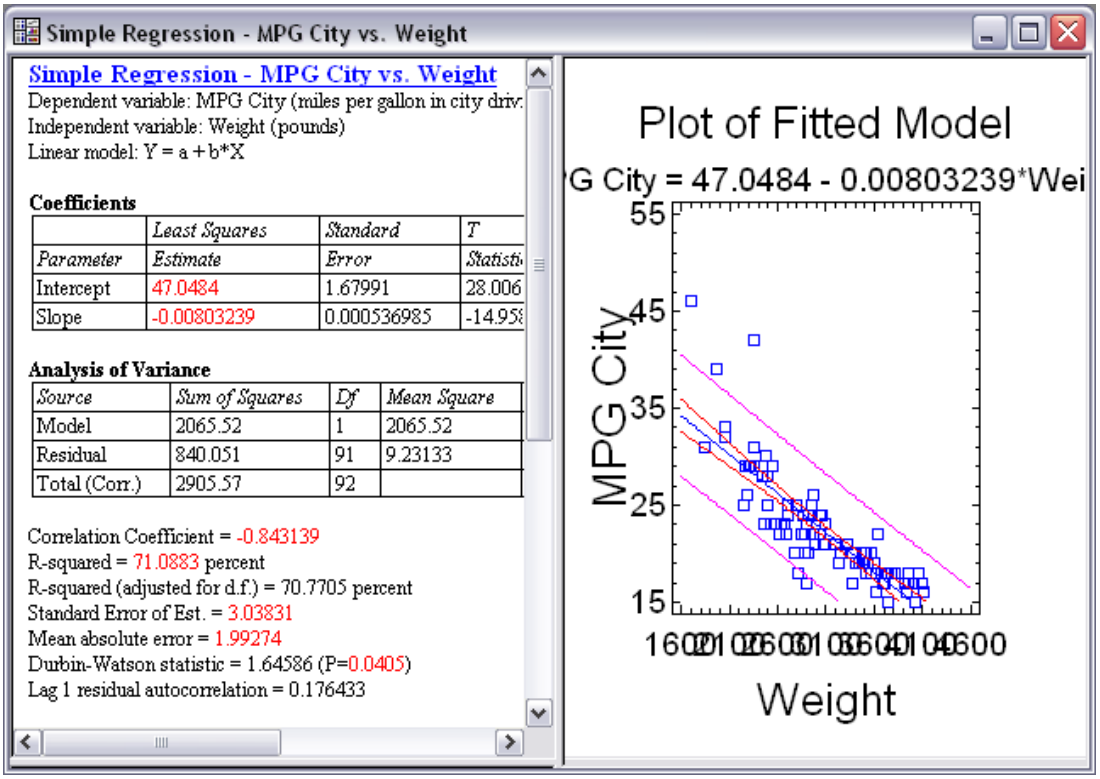


Figure 3-4. Simple Regression Analysis Window

The window is a "splitter window", with multiple *panes* divided by movable splitter bars. Tables are located along the left side of the window, while graphs are located along the right.

You can maximize the table or graph in any pane by double-clicking on it, in which case it will fill the window:

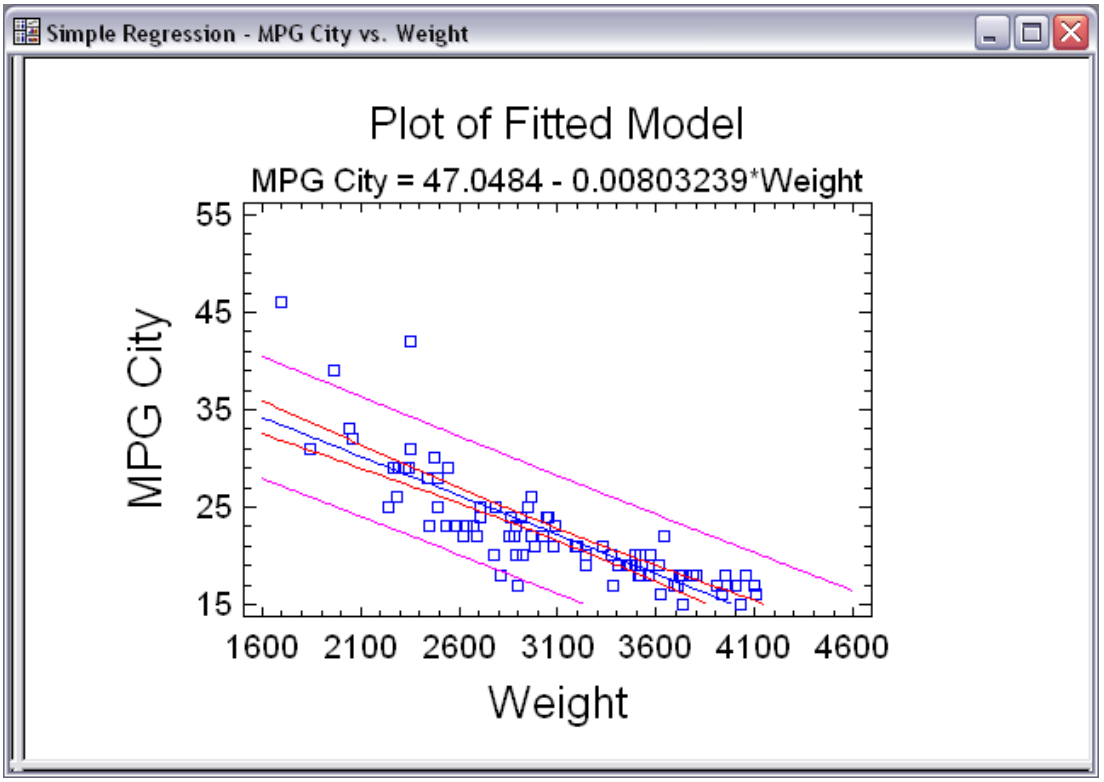


Figure 3-5. Simple Regression Analysis Window with Maximized Pane

Double-clicking on the pane a second time restores the multiple pane display.

When an analysis window has the focus, a second toolbar is activated directly beneath the main STATGRAPHICS Centurion XVI toolbar. The *analysis toolbar* appears as shown below:

Each of the buttons on this toolbar performs an important operation.

3.2.1 Input Dialog Button

 When pressed, this button displays the data input dialog box originally used to specify the data variables, as shown in *Figure 3-2*. If you change the data variables and press *OK*, the analysis will change to reflect the new selections. This enables you to try different combinations of data without having to start a new analysis.

3.2.2 Analysis Options Button

 Most analyses have multiple options. When first run, default values are selected for these options, which are often sufficient. However, pressing the *Analysis Options* button within any procedure will allow these basic settings to be changed. For *Simple Regression,* the *Analysis Options* dialog box specifies the type of model to be fit and the method for estimating the unknown model coefficients:

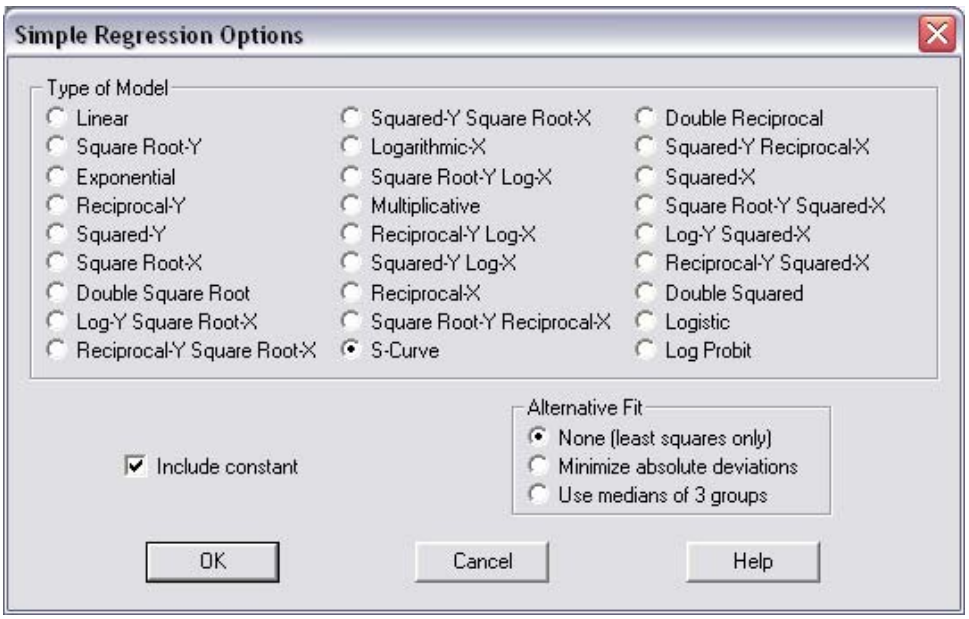


Figure 3-6. Simple Regression Analysis Options Dialog Box

If you examine the output in *Figure 3-9* below, it may be noted from the table of alternative models that several curvilinear models give a higher R-squared value than the linear model. At the top of the list is the *S-Curve* model. If this model is selected on the *Analysis Options* dialog box and the *OK* button is pressed, the entire analysis will change to reflect the new model. As may be seen by examining the plot of the fitted model, an S-Curve captures the curvature in the data quite well:

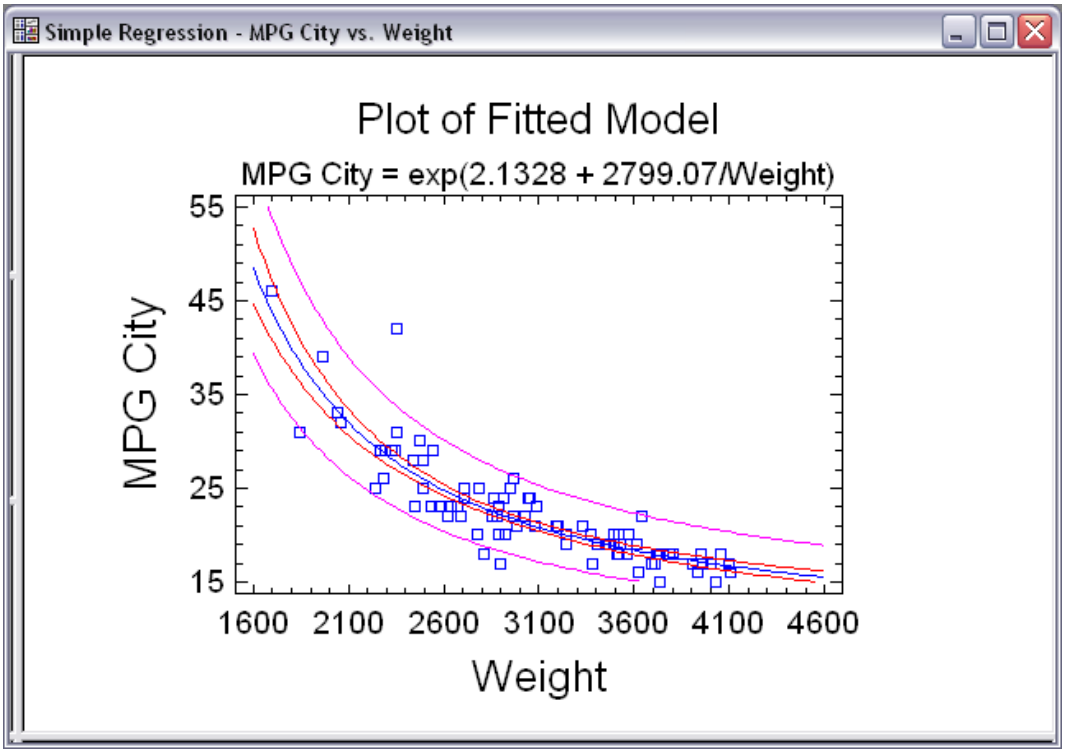


Figure 3-7. Fitted S-Curve Model

3.2.3 Tables and Graphs Button

This button displays a list of additional tables and graphs that may be added to the analysis window. For *Simple Regression*, the available tables and graphs are:

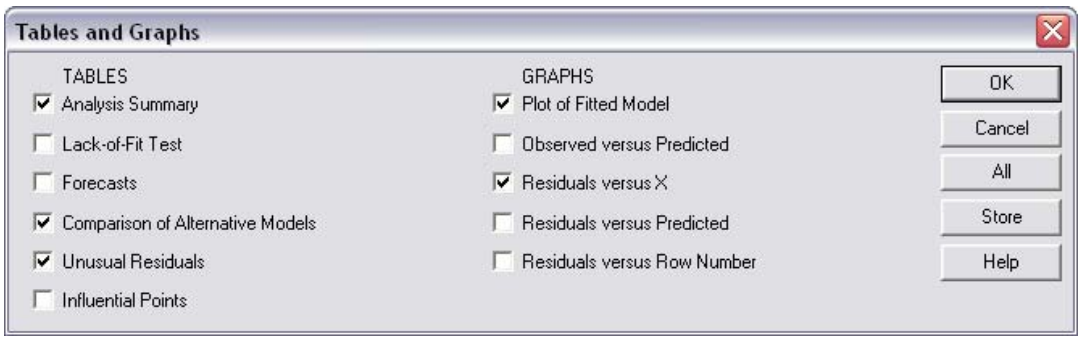


Figure 3-8. Simple Regression Tables and Graphs Dialog Box

For example, if you elect to add tables showing alternative models and unusual residuals, new text panes will be added to the analysis window:

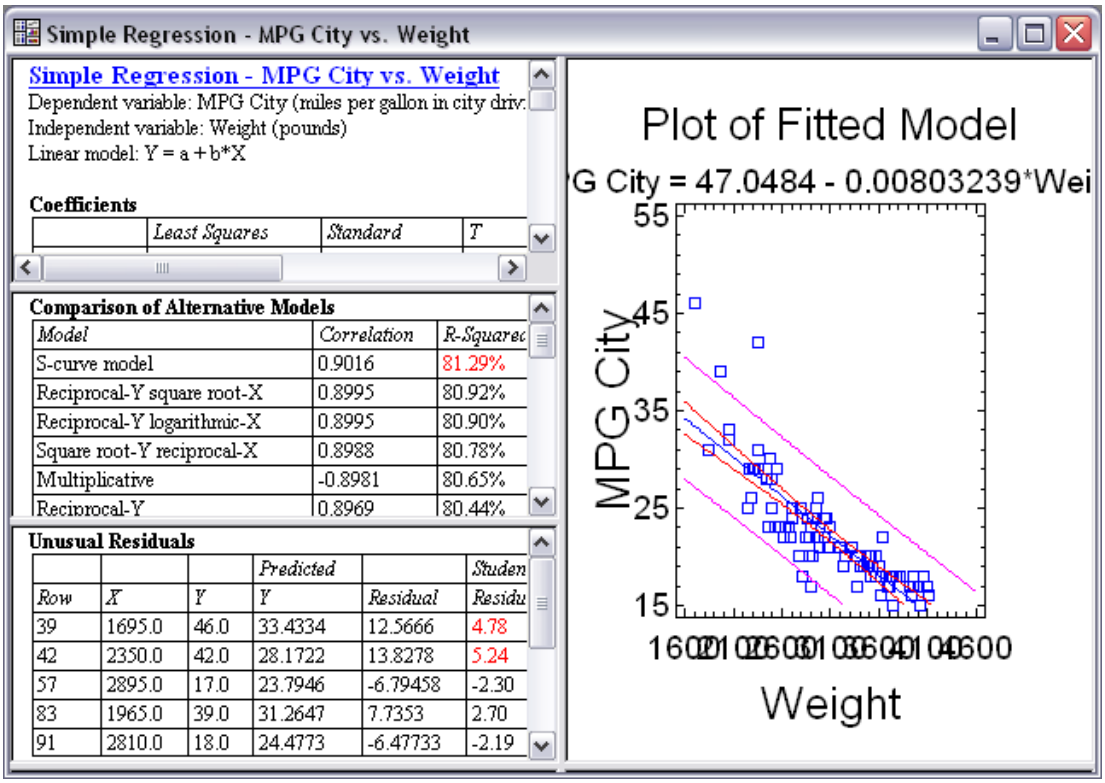


Figure 3-9. Simple Regression Analysis Window with Added Tables

Adding a residual plot places an additional graph in the analysis window:

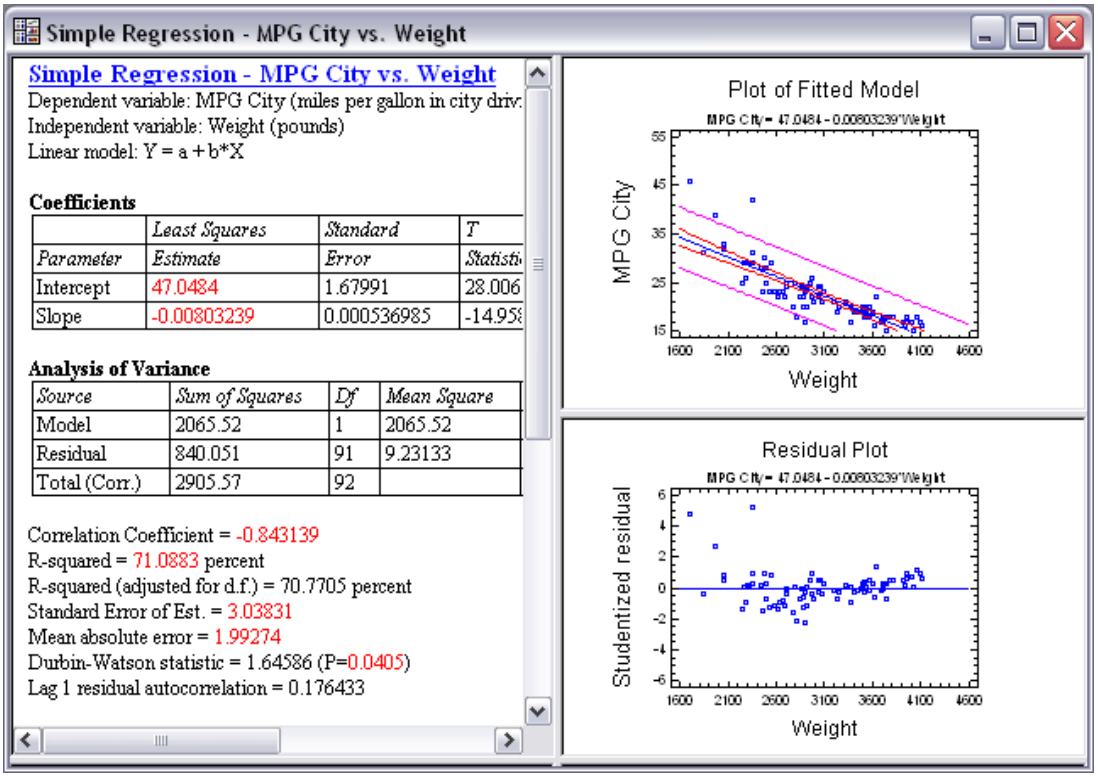


Figure 3-10. Simple Regression Analysis Window with Added Graph

3.2.4 Pane Options Button

 In addition to options that apply to the entire analysis window, many individual tables and graphs have options that apply only to them. These options are accessible by first maximizing the selected table or graph and then pressing *Pane Options*. For a *Fitted Model Plot*, the pane options are:

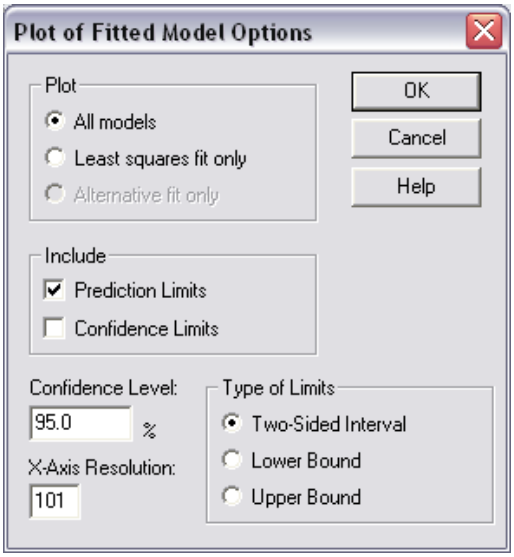


Figure 3-11. Pane Options Dialog Box for the Fitted Model Plot

For example, removing the check mark alongside *Confidence Limits* and pressing *OK* will replot the graph without the inner limits:

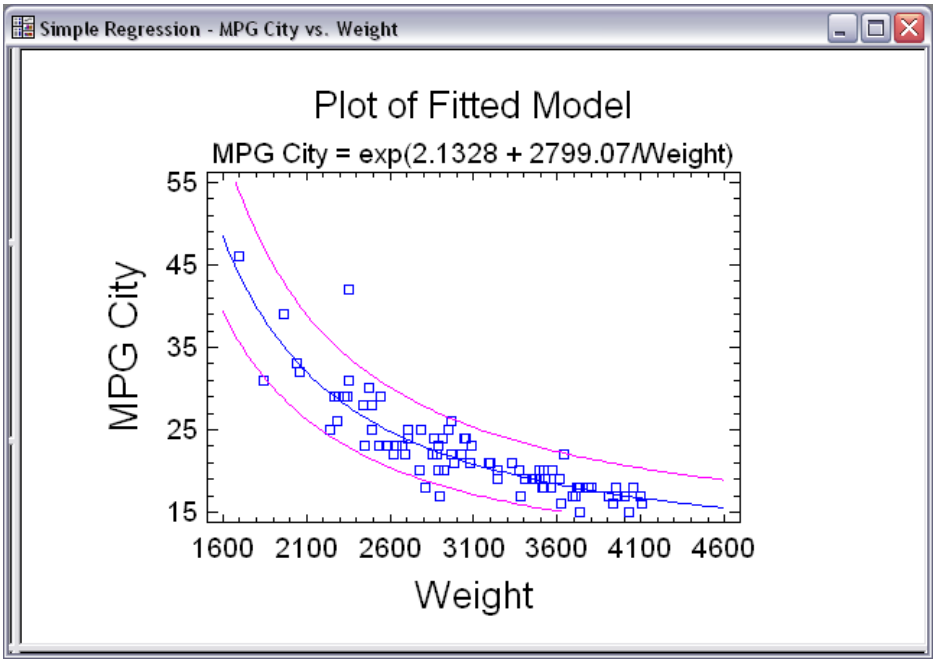


Figure 3-12. Fitted Model Plot Without Confidence Limits

3.2.5 Save Results Button

 This button allows you to save numerical results calculated by the statistical analysis back to columns of a datasheet. For *Simple Regression*, it displays the following choices:

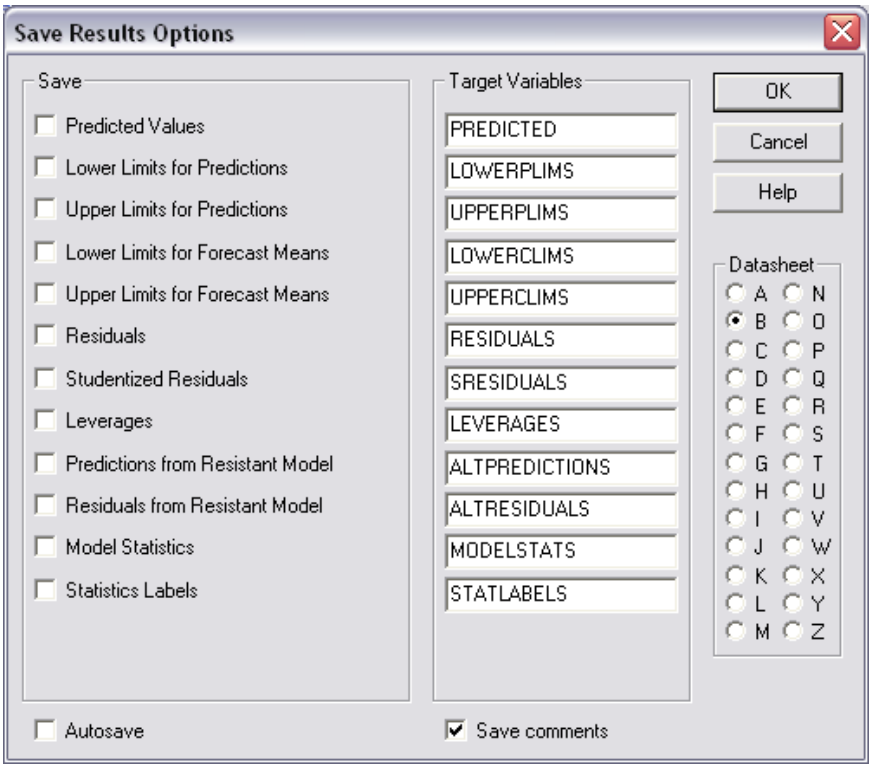


Figure 3-13. Simple Regression Save Results Dialog Box

To save information, check the items of interest in the *Save* field. For each item to be saved, assign a column name under *Target Variables* and indicate the desired *Datasheet*. If you wish to save a comment along with the data, check *Save comments*.

The *Autosave* box is used to automatically resave the selected item if and when the analysis is rerun. This is useful if you intend to save the analysis in a StatFolio, since analyses are rerun whenever StatFolios are loaded. By checking the *Autosave* box, you can set up a StatFolio to automatically calculate and save desired statistics. When combined with the scripting capability described in Chapter 5, this enables you to automate many tasks.

3.2.6 Graphics Buttons

Whenever a graph is maximized within the analysis window, several additional buttons are enabled. These buttons include:

 *Graphics options* – displays a dialog box used to change colors, labels, axis scaling, and other similar features.

 *Add text* – used to add additional text to the graph.

 *Jitter* – used to offset points randomly in the horizontal or vertical direction to prevent their overplotting each other.

 *Brush* – colors points on a scatterplot according to the value of a selected variable.

 *Smooth/Rotate* – smoothes a 2-dimensional plot, or rotates a 3-dimensional plot.

 *Pan or zoom* – stretches or zooms the graph in the X-, Y-, or Z-direction.

 *Explore* – dynamically explores response surface and contour plots.

 *Identify* – displays a label identifying a point when clicked on with the mouse.

 *Locate by name* – highlights in red any points with values equal to that entered in the *Locate* field (used in conjunction with the *Identify* button).

 *Locate by row* – highlights in red any points corresponding to the row number entered in the *Row* field.

Each of these buttons is described in detail in Chapter 4.

3.2.7 Exclude Button

Some statistical procedures allow you to interactively remove suspected outliers from an analysis by maximizing a graph, clicking on the suspect point, and pressing this button. For example, the plot in Figure 3-12 shows one point that is well outside the prediction limits. Clicking on that point and pressing the *Exclude* button causes the model to be refit without the point. The fitted model plot shows the new model, indicating which point (or points) have been removed with an X:

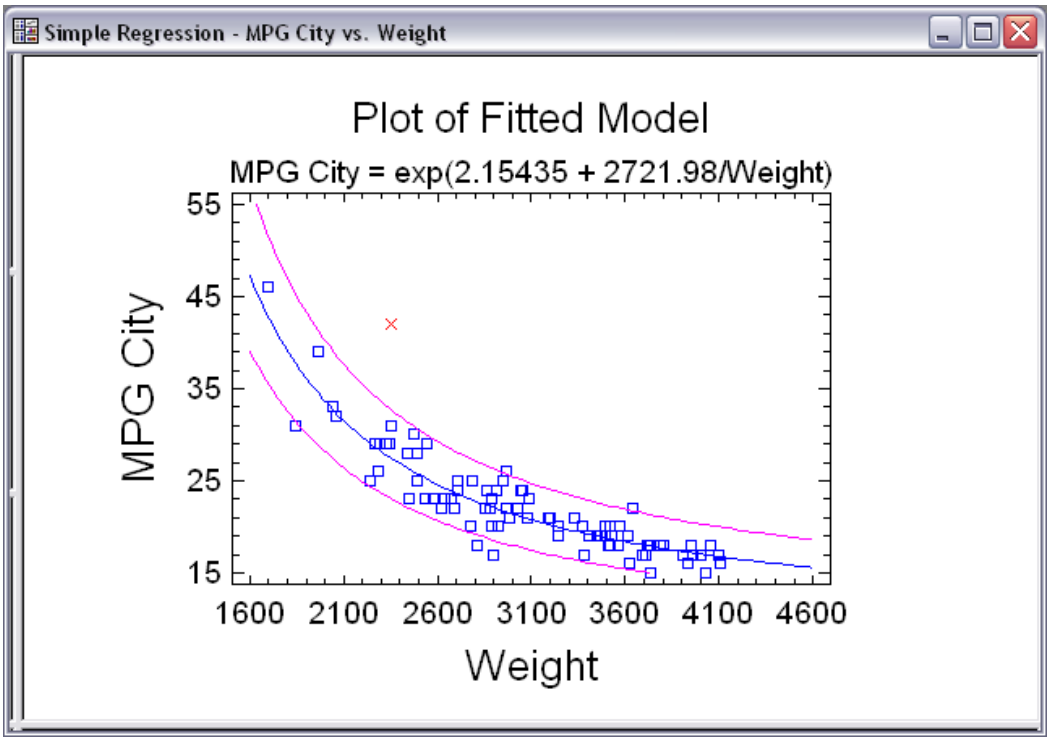


Figure 3-14. Fitted S-Curve Model after Excluding a Suspected Outlier

All of the other tables and graphs in the analysis window will also change to reflect the new model.

Multiple points may be excluded from a model by clicking on them one at a time and pressing the *Exclude* button. Clicking on a point that has been removed will put it back into the model.

3.3 Printing the Results

To print the results of a statistical analysis, two options are available:

1. To print all of the tables and graphs within the analysis window, press the *Print* button on the analysis toolbar or select *Print* from the *File* menu.

2. To print a single table or graph, click within its pane with the alternate mouse button and select *Print* from the popup menu that is displayed.

When printing the entire analysis, the following dialog box will be displayed:

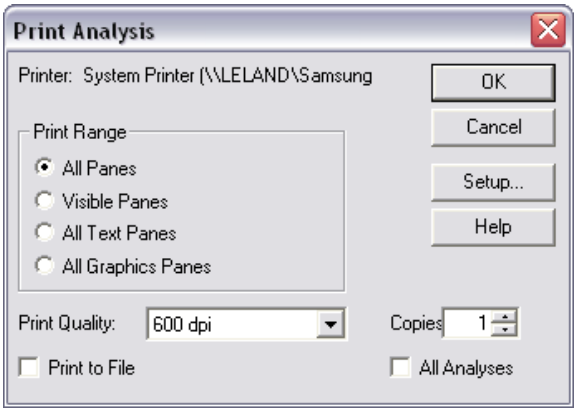


Figure 3-15. Dialog Box for Printing an Analysis

Under *Print Range*, specify the panes to be printed. You may simultaneously print the output in other analysis windows by checking *All Analyses*.

Additional options used when printing are contained on the dialog box accessible by selecting *Page Setup* from the *File* menu:

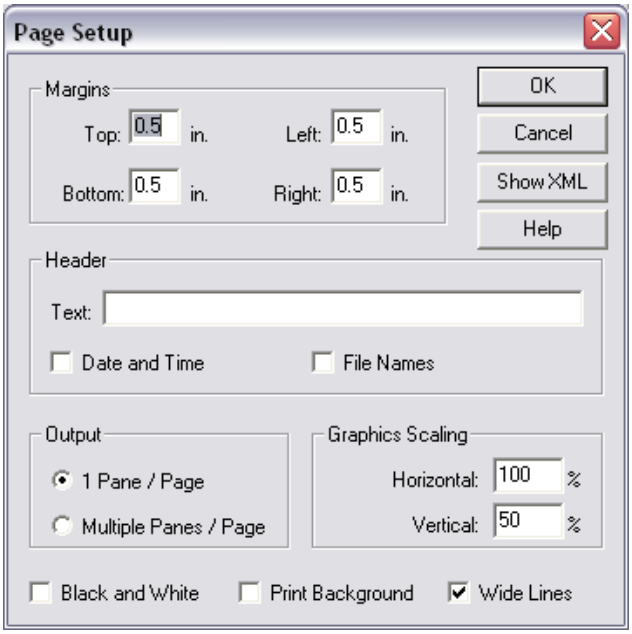


Figure 3-16. Page Setup Dialog Box

On this dialog box, you can:

1. Specify *margins* for the printed pages.

2. Indicate *header* information to be printed at the top of each page.

3. Indicate whether each pane (table or graph) should be displayed on a separate page, or whether *multiple panes* should be placed on a page if they will fit.

4. Specify the relative *size of graphs* as a percentage of the page dimensions.

5. Elect to plot the output in *black and white*, even if your printer has color capabilities.

6. Print the color *background* (if any) of your graphs.

7. Plot *wide lines* using 2 pixels instead of 1. This last option can make graphs appear much bolder on a high resolution printer.

Other options, such as whether to print the output in portrait or landscape mode, are set by selecting *Print Options* from the *File* menu, which accesses the dialog box supplied with your printer driver.

3.4 Publishing the Results

The output from a statistical analysis may be published in HTML format for viewing from within a web browser by selecting *StatPublish* from the *File* menu. This enables you to make the output available to everyone in your organization, whether or not they have STATGRAPHICS Centurion XVI on their computers. Publishing is described in Chapter 5.

You may also copy the analysis to the StatReporter, which allows you to annotate the output and save it in an RTF (rich text format) file, which may be read directly into programs such as Microsoft Word. Use of the StatReporter is described in Chapter 6.

Chapter

4

Graphics

Modifying graphs, saving graphics profiles, interacting with graphs, saving graphs in image files, and copying graphs to other applications.

Together, the 160 statistical procedures in STATGRAPHICS Centurion XVI create hundreds of different types of graphs. To facilitate the data analysis process, default titles, scaling, and other attributes are selected whenever a new graph is created. For analysis purposes, the defaults often suffice. But when it comes time to publish the final results, creating a publication-quality graph is important.

This chapter describes everything you need to know to work with graphs in STATGRAPHICS Centurion XVI. It shows you how to dress them up for publication. It shows you how to copy them to applications such as Microsoft Word and PowerPoint. It also shows you how to interact with graphs. For example, you might see an interesting point and wish to know more about it. Or you might want to spin a 3D plot around to get a sense of any relationship that might be present between the variables portrayed on the X, Y and Z axes.

As an example, we will consider again the data in the *93cars.sgd* file. To begin, the fitted model plot relating miles per gallon in city driving and vehicle weight will serve to illustrate some of the important graphics operations.

4.1 Modifying Graphs

The *Simple Regression* procedure is commonly used to fit curves relating a response variable Y and a second explanatory variable X. As illustrated in the last chapter, an S-Curve model provides a good fit to the relationship between the *MPG City* column and the *Weight* column in the *93cars.sf6* file.

When first created, a plot of the fitted S-Curve model is displayed as shown below:

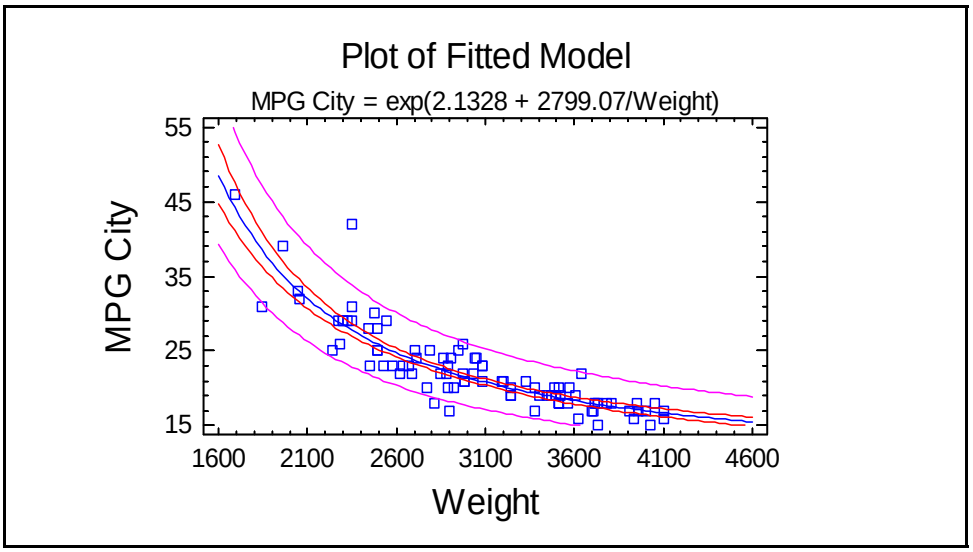


Figure 4-1. Fitted Model Plot with Default Titles and Scaling

The titles, scaling, point and line types, colors, and other graphics attributes were automatically generated.

4.1.1 Layout Options

To modify a graph once it has been created, first double-click on it so that it fills the analysis window. Then click on the *Graphics Options* button  located on the analysis toolbar. A tabbed dialog box will be displayed, with tabs corresponding to different graphics elements.

The *Layout* tab of the *Graphics Options* dialog box is used to change some of the basic features of the graph:

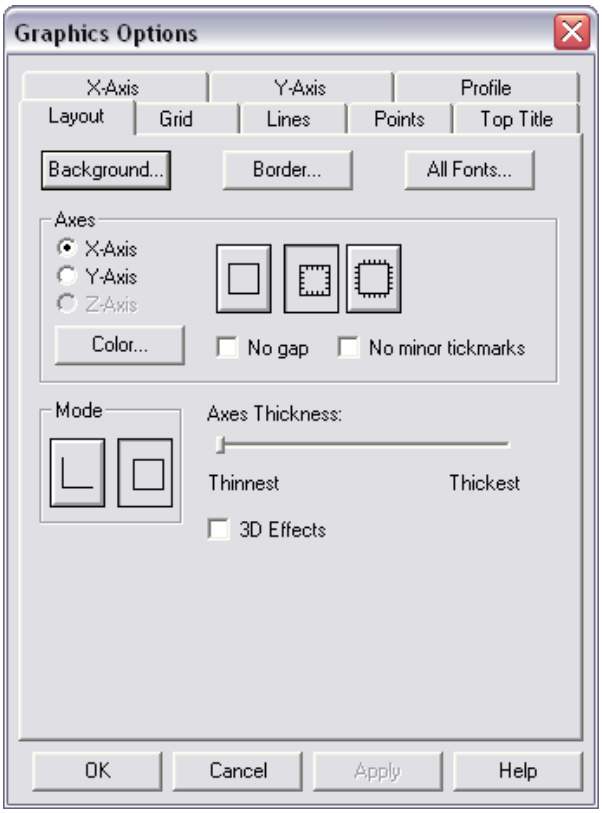


Figure 4-2. Layout Tab on Graphics Options Dialog Box

This includes the orientation of the axis tick marks, the thickness of the axes, and the color of the graph's background and border. For example, changing the *Background* color to yellow and adding *3D Effects* modifies the plot as shown below:

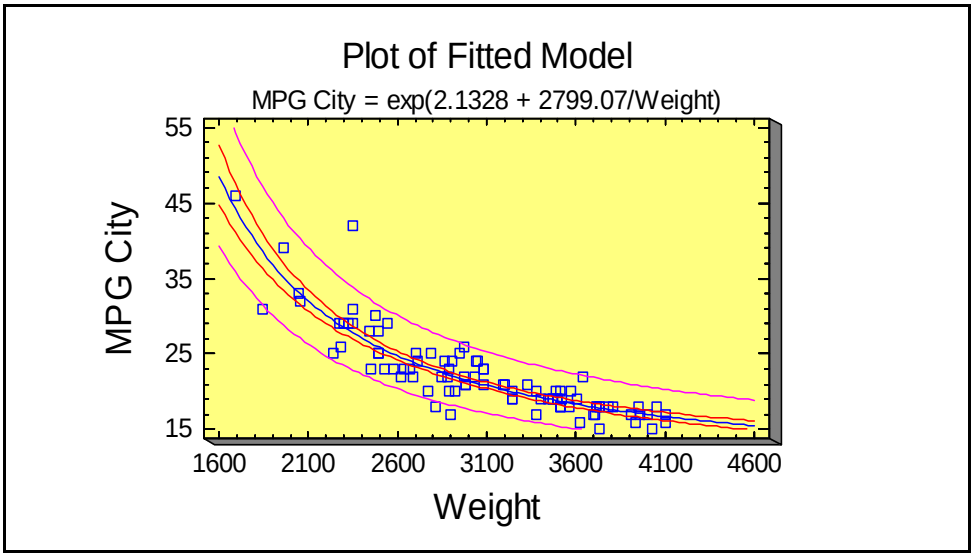


Figure 4-3. Plot after Modifying Background Color and Selecting 3D Effects

NOTE: This color change can be seen in the help documentation provided with your software by clicking on *Help – User Guide*.

4.1.2 Grid Options

The *Grid* tab is used to add a grid to the plot:

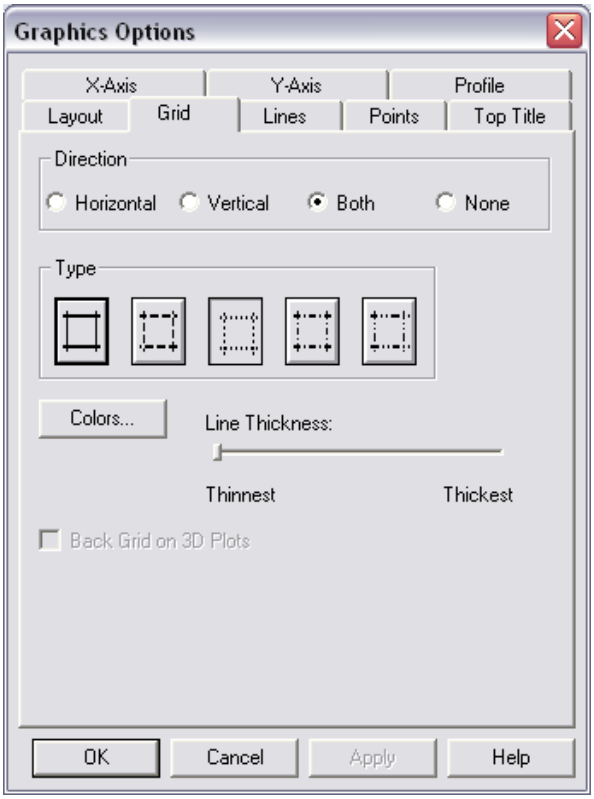


Figure 4-4. Grid Tab on Graphics Options Dialog Box

Adding a gray, dashed-line grid in *Both* directions produces the following graph:

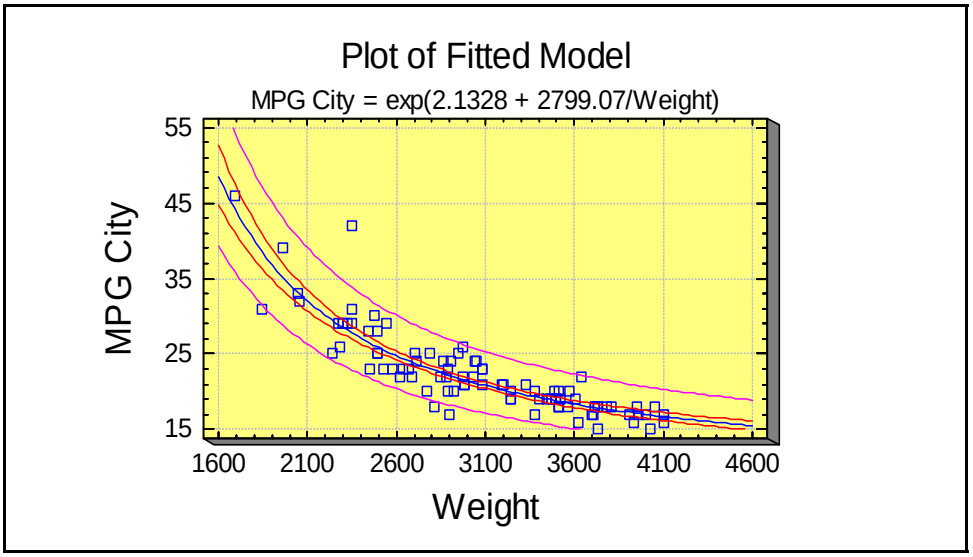


Figure 4-5. Plot after Adding a Grid

4.1.3 Lines Options

The *Lines* tab is used to specify the type, color and thickness of lines on a graph:

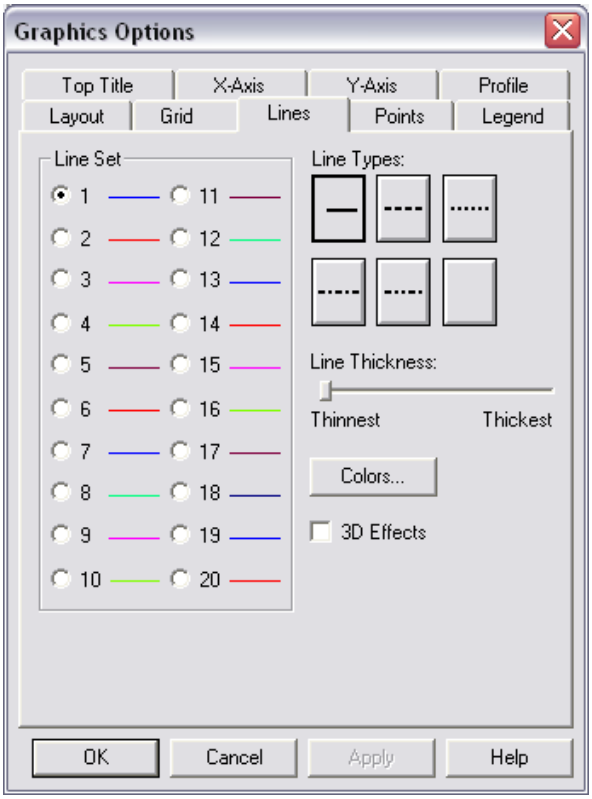


Figure 4-6. Lines Tab on Graphics Options Dialog Box

A plot such as that of the fitted model has three line sets: the line of best fit, the inner confidence limits, and the outer prediction limits. To change any of these types, click on radio button #1, #2 or #3 and then select the desired attributes. Increasing the thickness of the center line and changing the other line types results in:

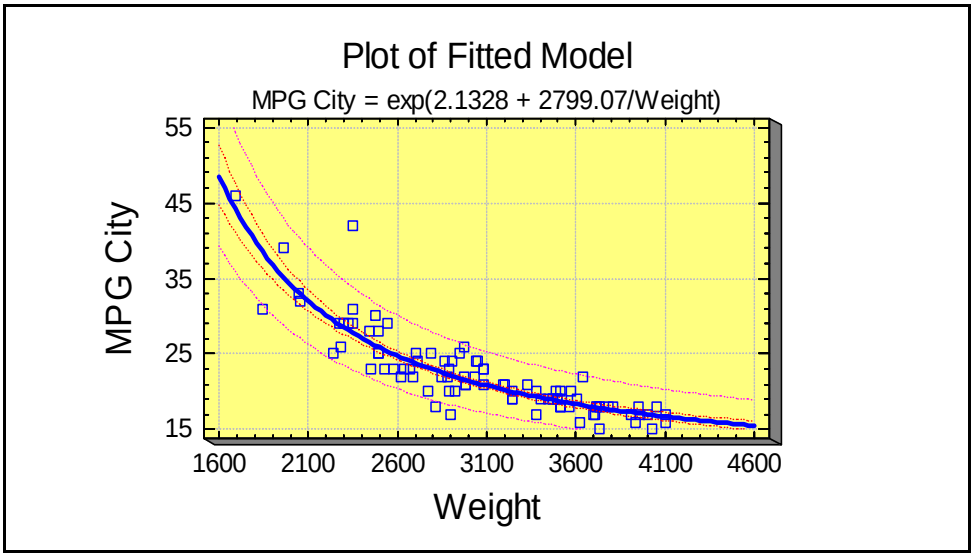


Figure 4-7. Plot after Modifying the Line Types

NOTE: you can only change the thickness of solid lines.

4.1.4 Points Options

The *Points* tab is used to specify the type, color and size of points on a graph:

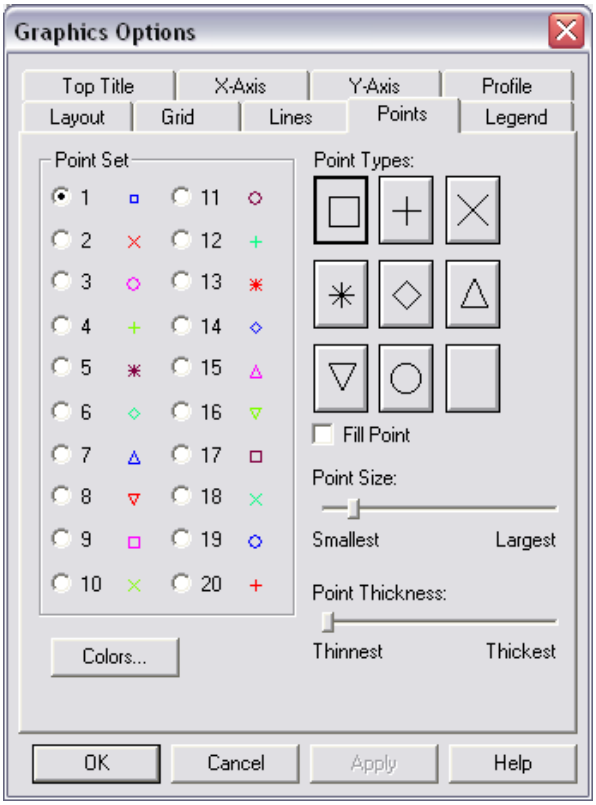


Figure 4-8. Points Tab on Graphics Options Dialog Box

Radio button #1 controls the attributes of the first set of points on a graph. In the current example, there is only one set. Changing the points to solid diamonds creates the following plot:

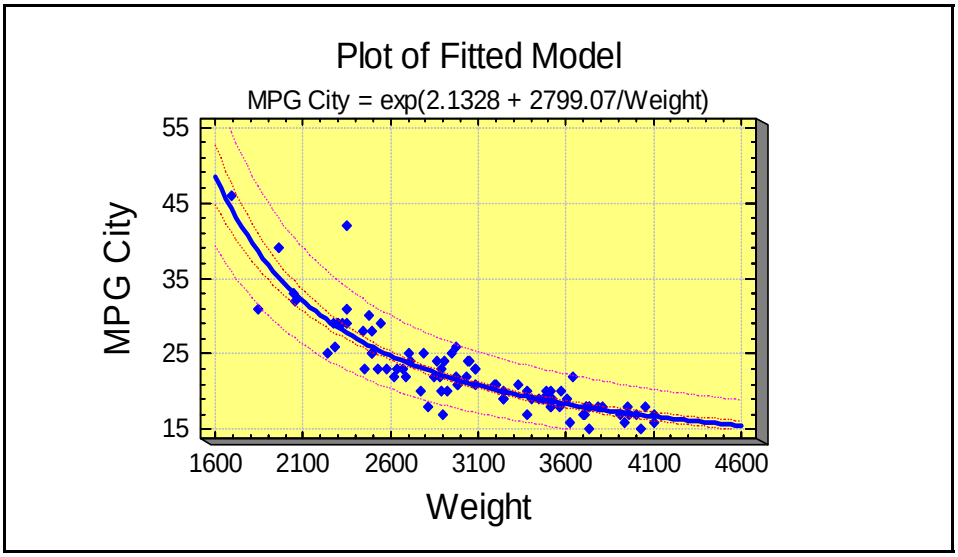


Figure 4-9. Plot after Modifying the Point Type

4.1.5 Top Title Options

The *Top Title* tab is used to specify the text and font type for the information displayed above a graph:

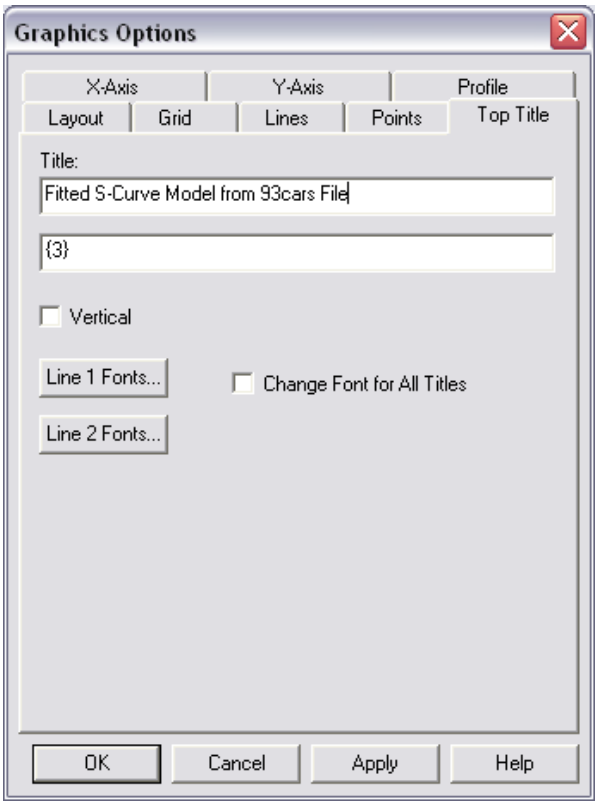


Figure 4-10. Top Title Tab on Graphics Options Dialog Box

Graphs have up to 2 title lines. An entry such as "{3}" in a title field indicates that the text is automatically generated by the analysis procedure, usually containing variable names or calculated statistics. You may change any title, including those that are automatically created. You may also drag the title to a new location with your mouse:

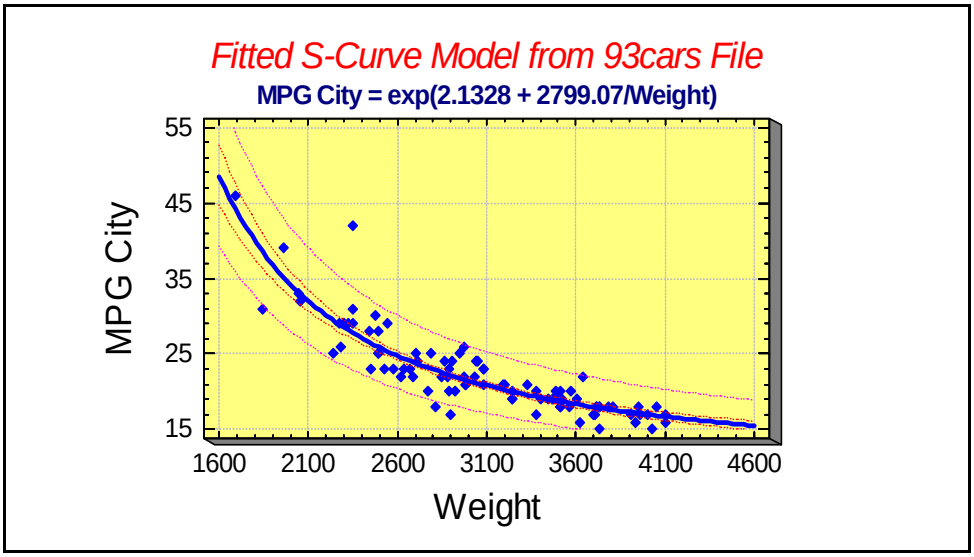


Figure 4-11. Plot after Modifying the Top Title

4.1.6 Axis Scaling Options

The *Graphics Options* dialog box also contains tabs that allow you to modify the axis titles and scaling:

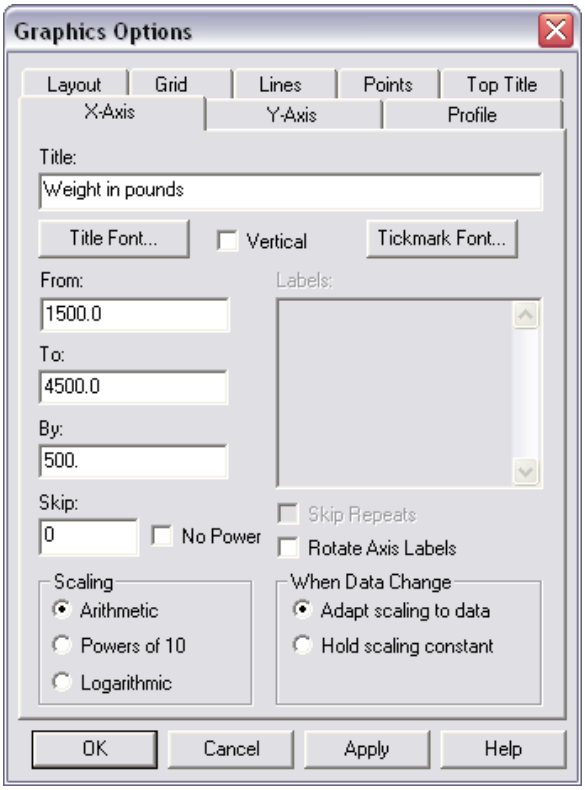


Figure 4-12. X-Axis Tab on Graphics Options Dialog Box

There are several important fields on this dialog box:

1. *Title*: title plotted along the axis.

2. *From*, *To*, *By*, and *Skip*: sets the tickmark scaling. The value in *Skip* is used to prevent displaying certain tickmarks if they run into each other. For example, a value of 1 in the *Skip* field would skip showing every other tickmark.

3. *Rotate Axis Labels:* changes the tickmark labels to vertical.

4. *No Power:* suppresses the display of large and small numbers using labels such as (X 1000).

5. *Scaling:* draws the axis using two different base 10 logarithmic scales.

6. *When Data Change:* specifies whether the scaling will stay constant or change when new data are plotted.

7. *Tickmark Font:* press these buttons to change the color, size, or style of the title and tickmarks.

The output generated from the above dialog box changes is shown below:

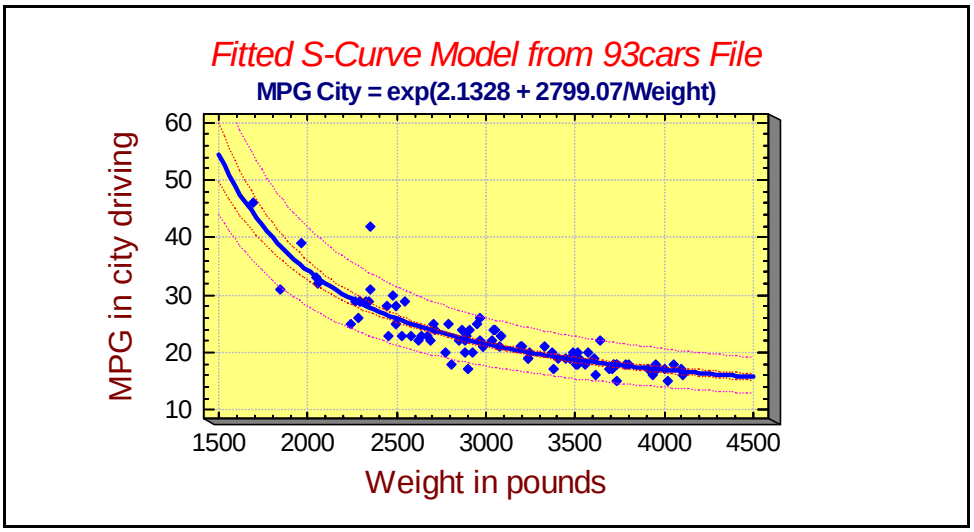


Figure 4-13. Plot after Modifying the Axis Titles and Scaling

4.1.7 Fill Options

Some plots, such as histograms, contain solid areas. The *Fills* tab on the *Graphics Options* dialog box controls the color and fill type of bars, polygons, and pie slices:

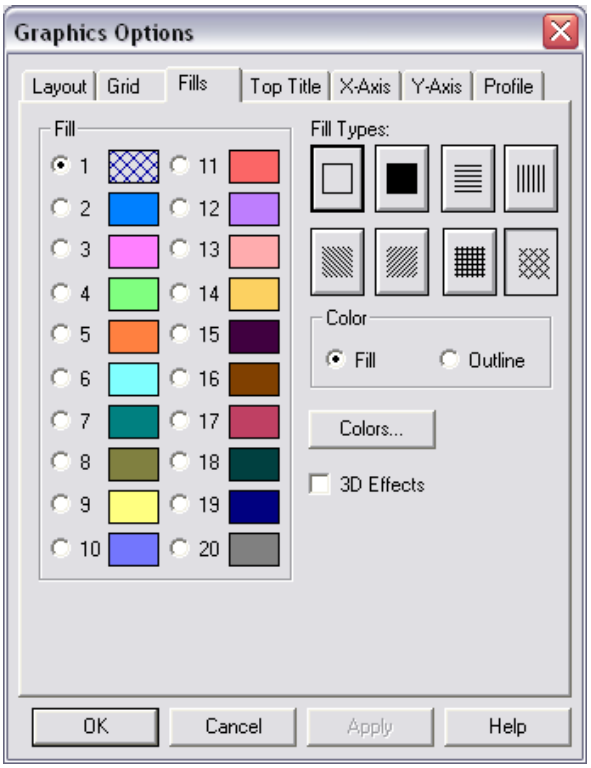


Figure 4-14. Fills Tab on Graphics Options Dialog Box

Radio button #1 controls the first fill type on a graph. In a histogram, all of the bars use the first fill type. On some graphs, such as piecharts, more than one fill type is used. In those cases, radio buttons #2 through #20 control the other fill types.

For plots such as histograms, setting a non-solid fill type is often a good idea when printing the results in black and white:

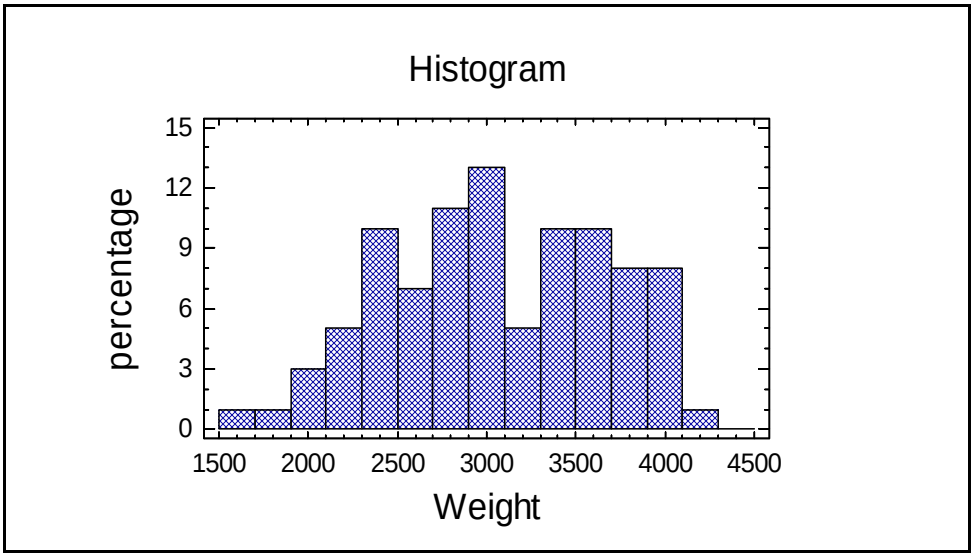


Figure 4-15. Frequency Histogram with Modified Fill Type

4.1.8 Text, Labels and Legends Options

For graphs containing additional legends or labels, tabs are included on the *Graphics Options* dialog box that allow you to change the text and fonts.

4.1.9 Adding New Text

Additional text may also be added to any graph by pressing the *Add text* button  on the analysis toolbar. A dialog box will be generated in which to enter the new text:

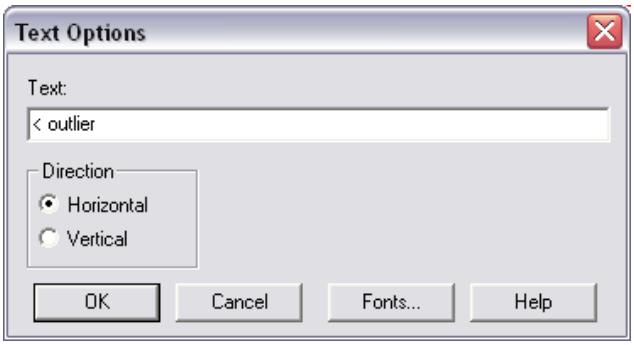


Figure 4-16. Dialog Box for Adding New Text

The text string will be initially positioned under the top title, but may be dragged to any location with the mouse:

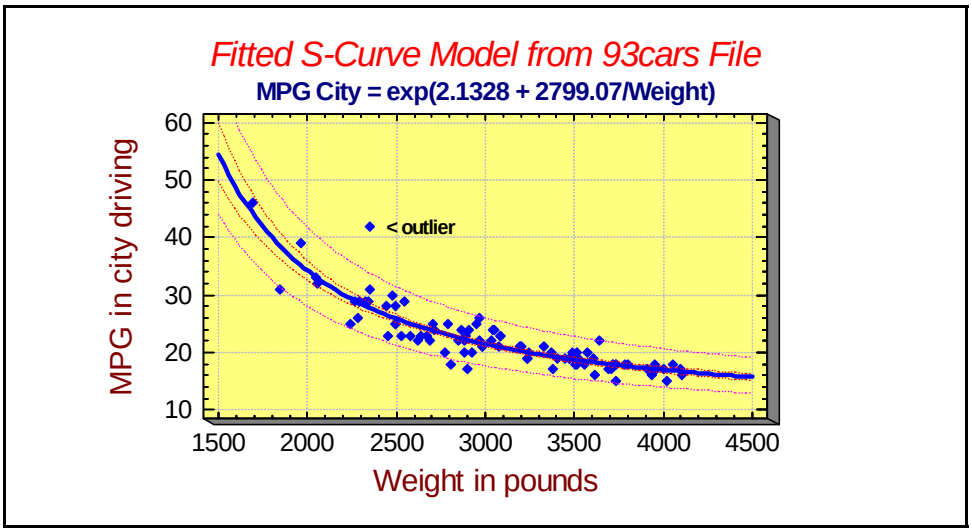


Figure 4-17. Plot after Adding New Text String

After text is added, click on it and then press the *Graphics Option* button if changes need to be made.

4.2 Jittering a Scatterplot

When one or both of the variables in a scatterplot are discrete, the chance of points being exactly in the same location and obscuring each other can be large. The analysis toolbar has a *Jitter* button that overcomes this problem by randomly offsetting points in the horizontal and/or vertical direction. For example, consider the following plot of the data in the *93cars.sgd* file:

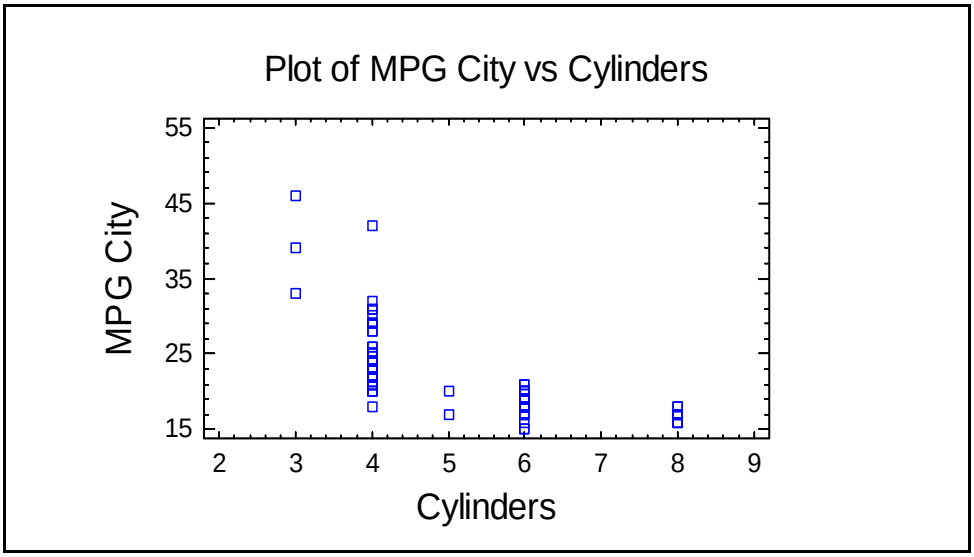


Figure 4-18. Scatterplot of Miles per Gallon versus Cylinders

Although there are 93 rows in the datasheet, there are many less points than that on the plot.

If you press the *Jitter* button, a dialog box will appear allowing you to add a little jitter (random offset) to the points:

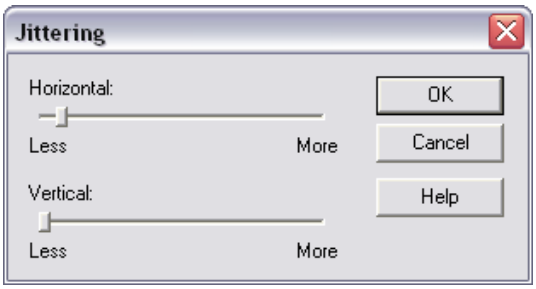


Figure 4-19. Jittering Dialog Box

In this case, adding a small amount of horizontal jitter gives a much better picture of the location of the points:

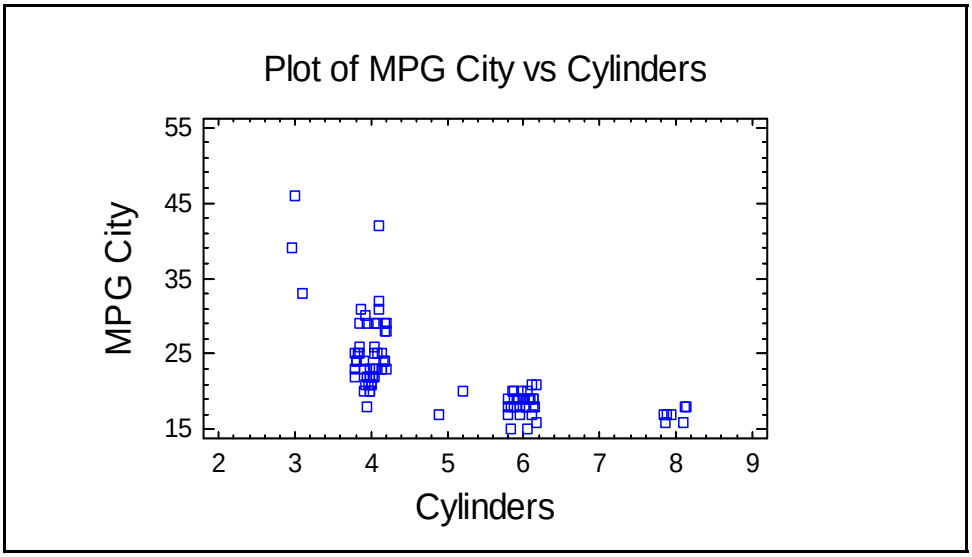


Figure 4-20. Scatterplot after Horizontal Jitter

Each point has been offset a small random amount along the horizontal axis. Jittering a plot affects only the display. It has no affect on the data in the datasheet or any calculations made with it.

4.3 Brushing a Scatterplot

An interesting method of visualizing relationships between variables is to color the points of a scatterplot according to the value of another variable. For example, consider the following *Matrix Plot* for selected variables from the *93cars.sgd* file:

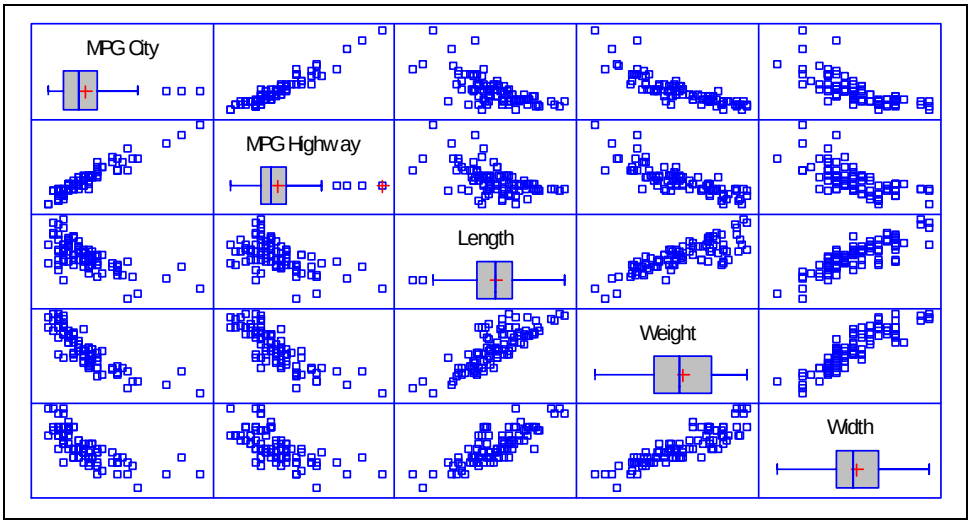


Figure 4-21. Matrix Plot for Data from the 93cars File

The scatterplot in each cell of the matrix plots the values of variables corresponding to its row and column identifiers.

Suppose you wished to visualize how the horsepower of the automobiles was related to the 5 plotted variables. If you press the *Brush* button  on the analysis toolbar, the following dialog box will be displayed:

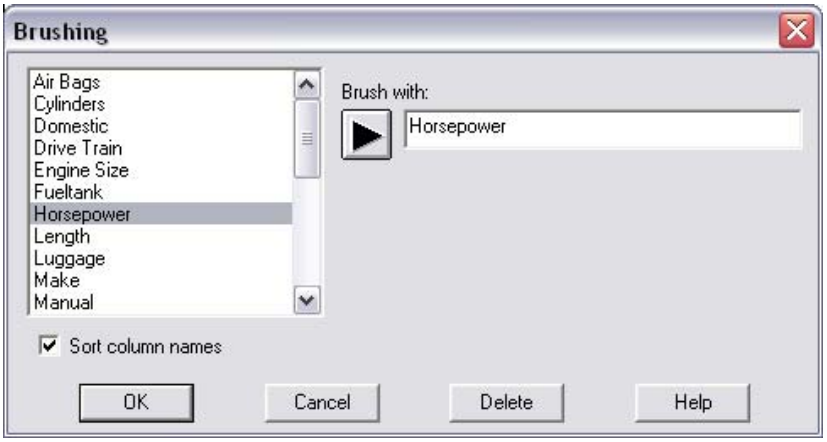


Figure 4-22. Dialog Box for Selecting Brushing Variable

Select a quantitative variable to use to code the points. After selecting the variable to brush with, a floating dialog box will appear:

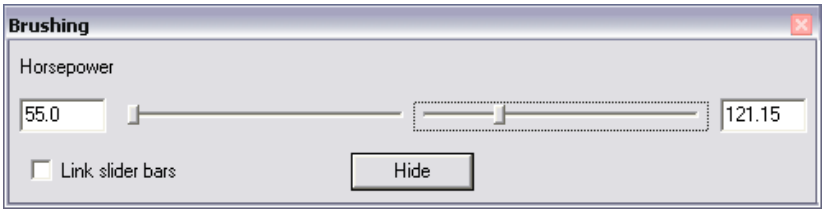


Figure 4-23. Floating Dialog Box for Selecting Brushing Interval

The two slider bars are used to specify lower and upper limits for the variable. All points in the plot are colored light blue if they fall within the specified interval. For example, in the plot below, all automobiles with horsepower between 55.0 and 121.15 are colored light blue:

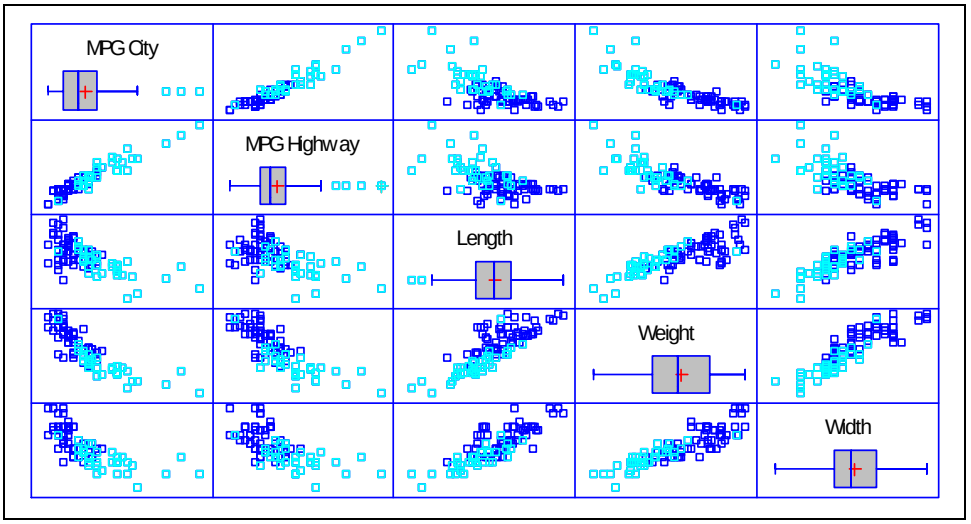


Figure 4-24. Matrix Plot after Brushing Points

It is evident from the above plot that *Horsepower* is strongly correlated with the other variables.

4.4 Smoothing a Scatterplot

To help visualize the relationships between the variables in a scatterplot, a smoother may be added. To smooth a scatterplot, press the *Smooth/Rotate* button  on the analysis toolbar. This will display the following dialog box:

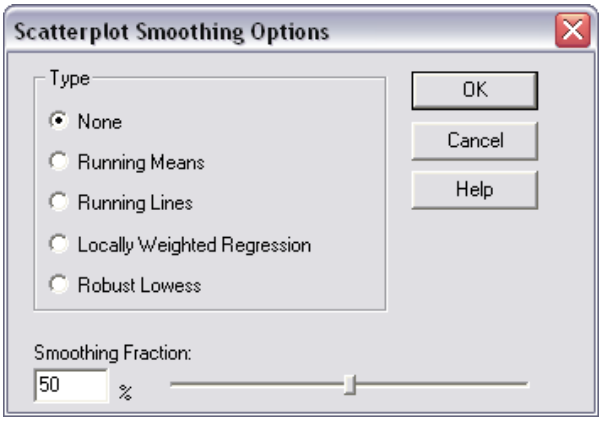


Figure 4-25. Scatterplot Smoothing Dialog Box

Smoothing a scatterplot is done by selecting a set of locations along the X-axis and at each location plotting a weighted average of the specified fraction of the points that are closest to that location. One of the best methods for smoothing is called LOWESS (LOcally WEighted Scatterplot Smoothing), usually with a smoothing fraction between 40% and 60%. The result of smoothing the *Matrix Plot* of the automobile data is shown below:

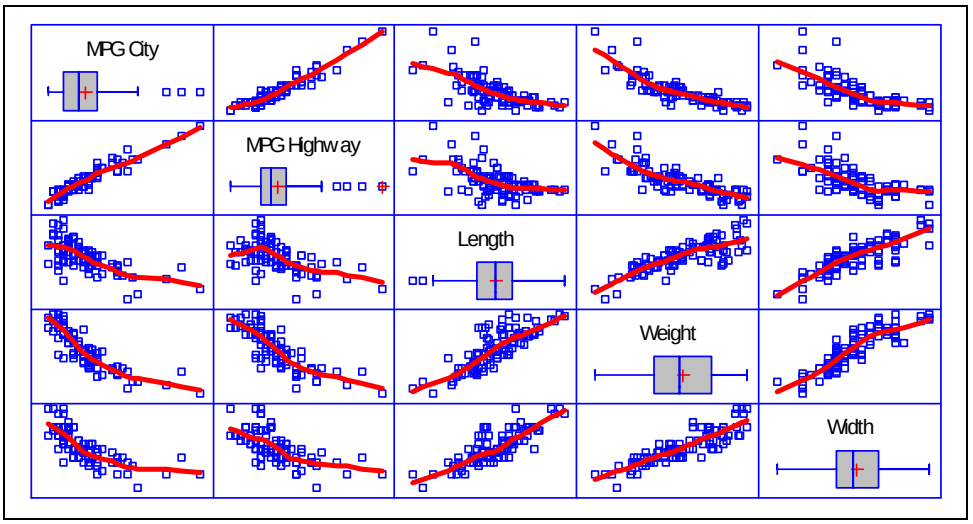


Figure 4-26. Smoothed Matrix Plot using Lowess with a 50% Smoothing Fraction

The smooth helps illustrate the type of relationships between the variables.

4.5 Identifying Points

To display the row number and coordinates corresponding to any point on a graph, you may hold the mouse button down on the point. A small box will be displayed in the upper right corner of the plot, showing the row number and coordinates of the point:

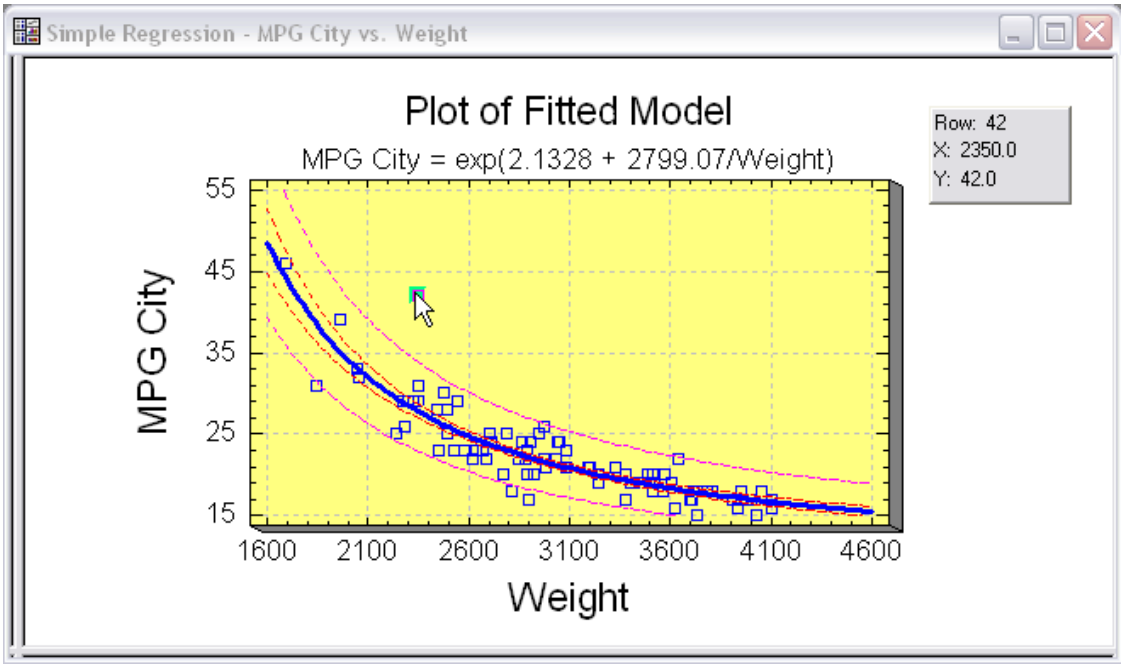


Figure 4-27. Displaying Information about a Selected Point

At the same time, the row number of the point will be placed into the *Row* field on the analysis toolbar:

Figure 4-28. Analysis Toolbar Showing Row Number of Selected Point

Additional information about the point may be obtained by pressing the *Identify* button  and selecting a column from the DataBook:

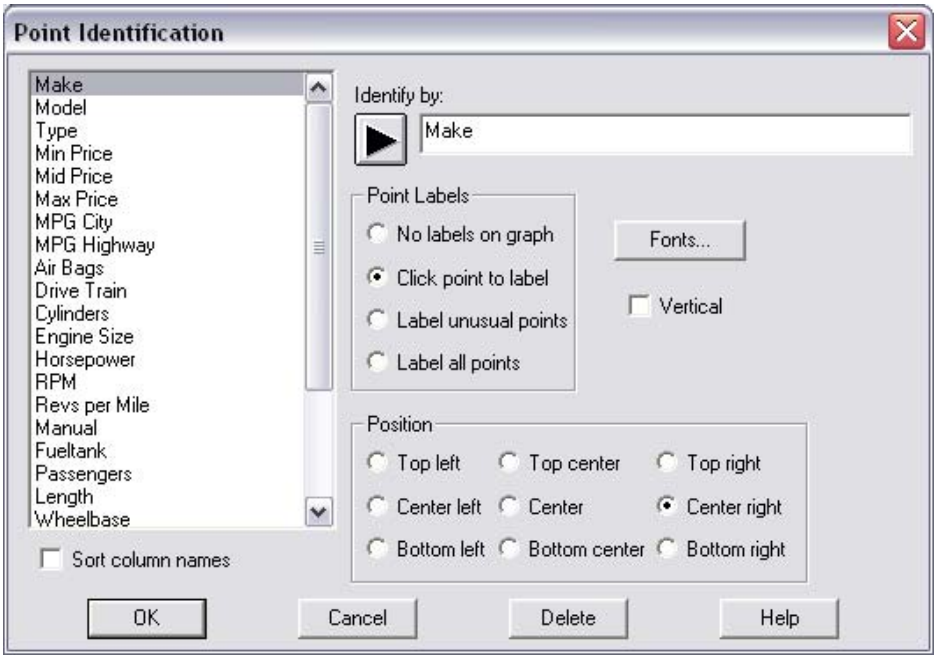


Figure 4-29. Point Identification Dialog Box

After selecting a variable, clicking on any point will add the value of that variable to the *Label* field on the analysis toolbar:

Figure 4-30. Analysis Toolbar Showing Make of Selected Point

The binoculars buttons  to the right of the *Label* and *Row* fields may be used to locate points on a graph. If you enter a value into either edit field and then press the corresponding *Locate* button, all points in the graph matching the entered value will be highlighted. For example, the plot below colors the points for all Hondas light blue:

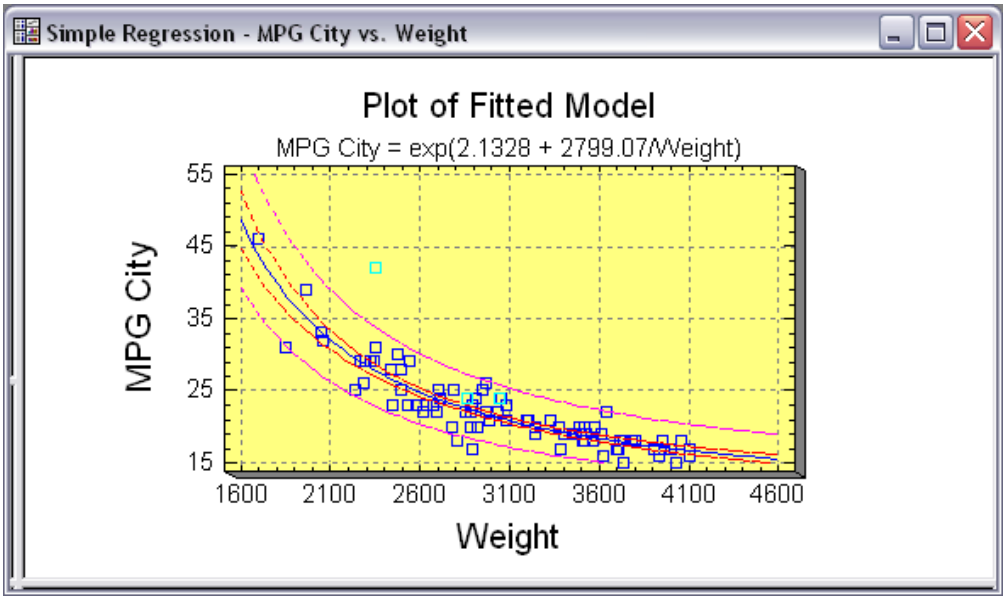


Figure 4-31. Plot Highlighting All Hondas

This technique is also quite effective on a *Matrix Plot*. In the following display, all points corresponding to row #42 have been highlighted:

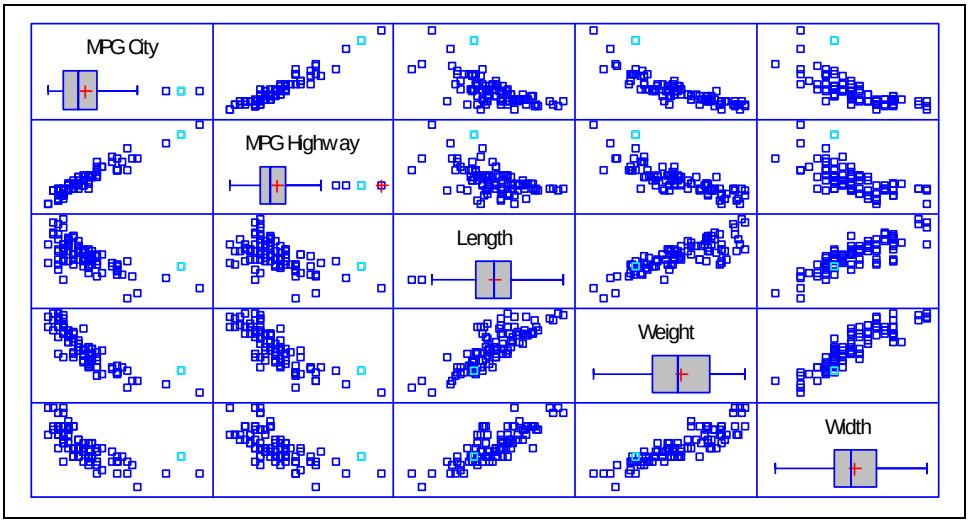


Figure 4-32. Matrix Plot Highlighting Row #42

Locating a point in a *Matrix Plot* can help identify whether it is an outlier with respect to more than one variable.

NOTE: the color used to highlight the points is specified on the *Graphics* tab of the *Preferences* dialog box, accessible from the *Edit* menu.

4.6 Copying Graphs to Other Applications

Once a graph has been created in STATGRAPHICS Centurion XVI, it can be easily copied to other programs such as Microsoft Word or PowerPoint by:

1. Maximizing the pane containing the graph.

2. Selecting *Copy* from the STATGRAPHICS Centurion XVI *Edit* menu.

3. Selecting *Paste* while in the other application.

By default, graphs are pasted in "Picture" format, which corresponds to a Windows metafile. In rare cases when you wish to paste the graph in some other format, you can select *Paste Special* instead of a simple *Paste*.

To copy an entire analysis to another application, including all tables and graphs, first copy the analysis to the StatReporter using the alternate mouse button popup menu, and then copy the StatReporter to the other application. This technique is illustrated in Chapter 7.

To copy both the graph and its enclosing window, as in *Figure 4-31* above, a third-party screen capture tool is recommended. In producing this manual, a program called *SnagIt* has been used, which is available for purchase at *www.techsmith.com*. If you use *SnagIt*, we recommend that you set the *Input* option to "Window" and the *Output* option to "Clipboard". You can then paste images directly into any document.

4.7 Saving Graphs in Image Files

Individual graphs may also be saved in image files by maximizing a graph and then selecting *Save Graph* from the *File* menu. A dialog box will be displayed on which to specify a file name and image format:

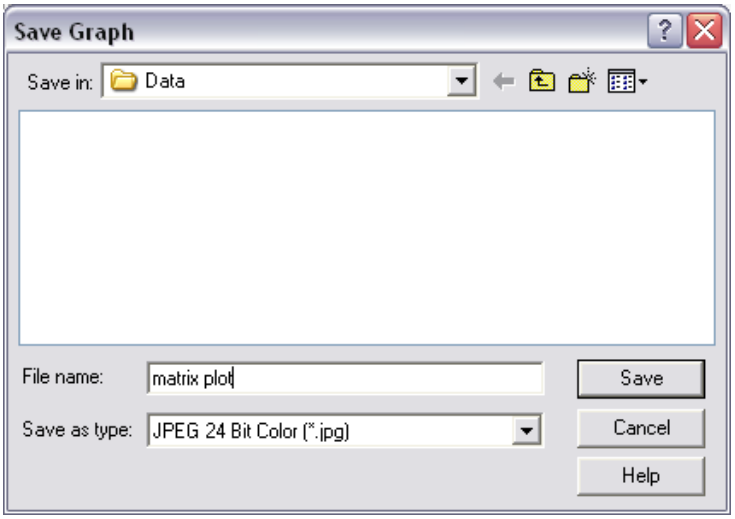


Figure 4-33. File Selection Dialog Box for Saving Graph in Image File

For saving graphs that are to be read into Word or PowerPoint, saving the graph as a Windows metafile gives the most flexibility. If the graph is to be displayed on a web page, saving it as a JPEG file is recommended.

Chapter

5

StatFolios

Saving your session, publishing results in HTML format, and automating analyses using start-up scripts.

Each time you select a statistical analysis from the STATGRAPHICS Centurion XVI menu, a new analysis window is created. You may save all of the analysis windows at any time by creating a *StatFolio*. A StatFolio is a file containing the definition of all statistical analyses that have been created, with pointers to the data used by them. By saving a StatFolio and later reopening it, you effectively save and retrieve your current STATGRAPHICS Centurion XVI session.

When a session is saved in a StatFolio, it is the definition of the analyses that is saved, not the output. When reopening a StatFolio, the data in the associated data sources is reread and all analyses recalculated. StatFolios thus provide a simple method for repeating analyses at a later time on different data.

You may also create a script that is executed whenever a StatFolio is loaded. Details of this and other StatFolio features are described in this chapter.

5.1 Saving Your Session

To save the current status of your STATGRAPHICS Centurion XVI session, select *File* – *Save* – *Save StatFolio* from the main menu. Enter a name for the StatFolio in the dialog box shown below:

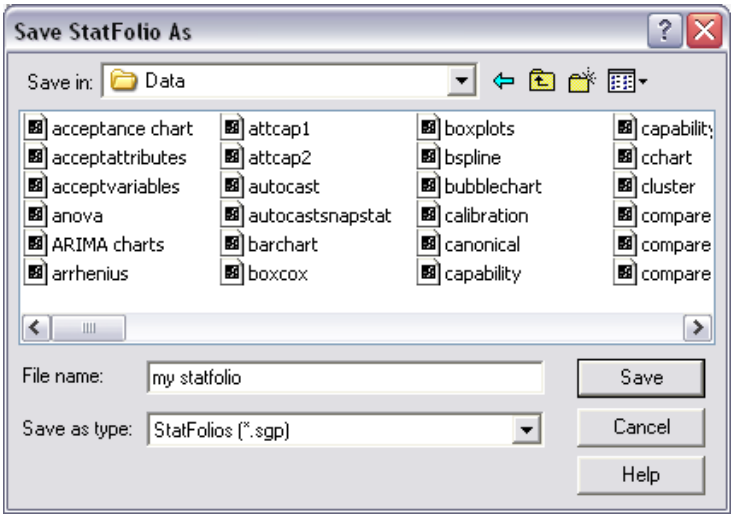


Figure 5-1. File Selection Dialog Box for Saving StatFolio

StatFolios are saved in files with the extension *.sgp*. They contain:

1. A definition of all analyses that have been created, including the input variables, the tables and graphs, settings of all options, changes made to graphs, etc. When a StatFolio is reopened, the analyses are recalculated and all tables and graphs updated.

2. Links to the data sources contained in the DataBook. If the data change between the time the StatFolio is saved and when it is reopened, the analysis windows will reflect the changes.

3. Links to a StatGallery and StatReporter file, if material has been placed in them before the StatFolio was saved. The program will ask you to supply names for the StatGallery and the StatReporter when the StatFolio is saved.

5.2 StatFolio Scripts

When a StatFolio is first loaded, all of the analysis windows are restored to their previous condition. STATGRAPHICS Centurion XVI then looks to see whether a Start-up script has been saved with the StatFolio and executes it if it has. A script may be created by selecting *StatFolio Start-up Script* from the *Edit* menu. A dialog box is displayed with fields to define a sequence of actions to be performed:

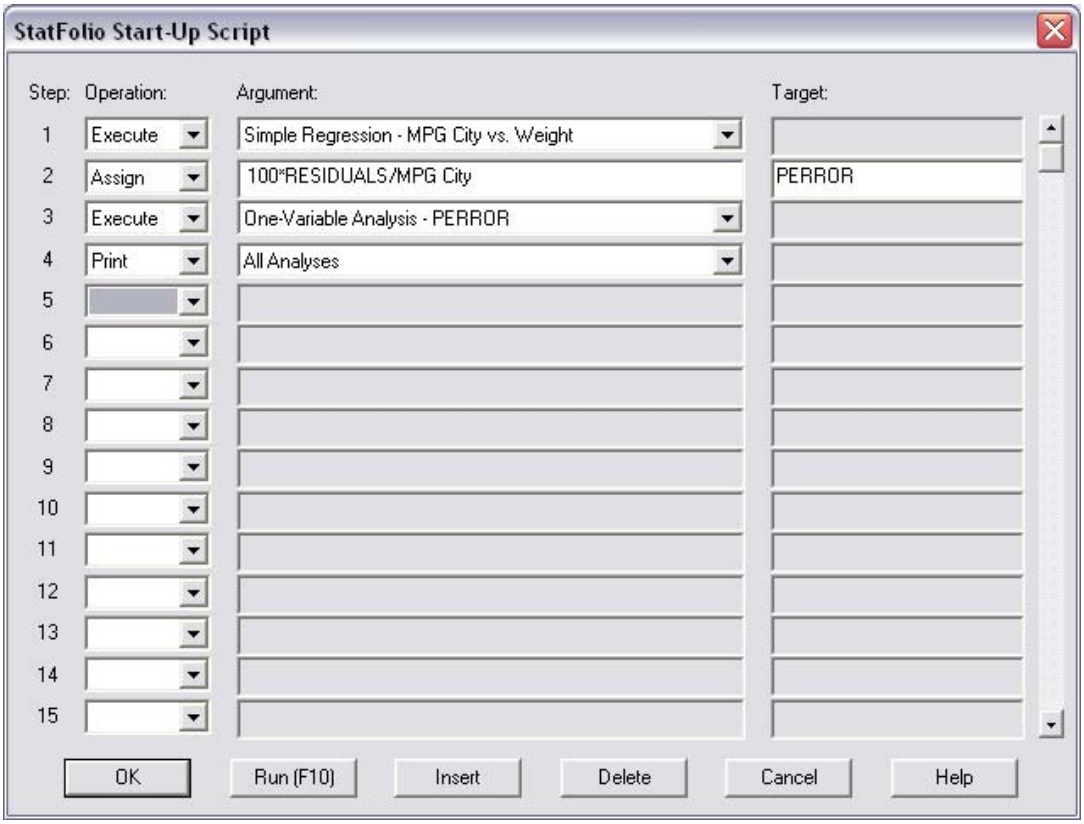


Figure 5-2. A Simple StatFolio Start-Up Script

The desired operations are specified in the order they should be performed. The available operations are:

Operation	*Argument*	*Target*	*Description*
Execute	Analysis title		Updates the indicated analysis.
Assign	STATGRAPHICS Centurion expression	Column name	Evaluates the expression and assigns it to the indicated column.
Print	Window(s) to print		Prints the contents of the indicated windows.
Publish			Runs StatPublish to publish the contents of the StatFolio in HTML format.
Shell	Windows command to execute	Command argument	Causes Windows to execute a command.
Delay	Number of seconds		Pauses for the specified time.
Load	Name of StatFolio		Specifies StatFolio to load after script is run. This allows StatFolios to be executed in a chain.
Exit			Exits STATGRAPHICS Centurion XVI

Figure 5-3. Start-Up Script Operators

In the example shown in *Figure 5-2*, a *Simple Regression* is performed. Within that analysis, it is assumed that *Save Results* has been set to automatically save the residuals from the fitted model in a column called *RESIDUALS*. The residuals are then divided by the original data values and multiplied by 100 to create percentage errors, which are assigned to a new variable called *PERROR*. The values in *PERROR* are then summarized using the *One-Variable Analysis* procedure, after which the results of both analyses are printed.

Note that StatFolios can be chained together using the *LOAD* operator in one script to load and start the script in another StatFolio. You can also automatically exit STATGRAPHICS Centurion XVI using the *EXIT* operator.

NOTE: You can suppress execution of scripts by selecting *Disable Start-Up Scripts* on the *General* tab of the *Preferences* dialog box, accessible from the *Edit* menu:

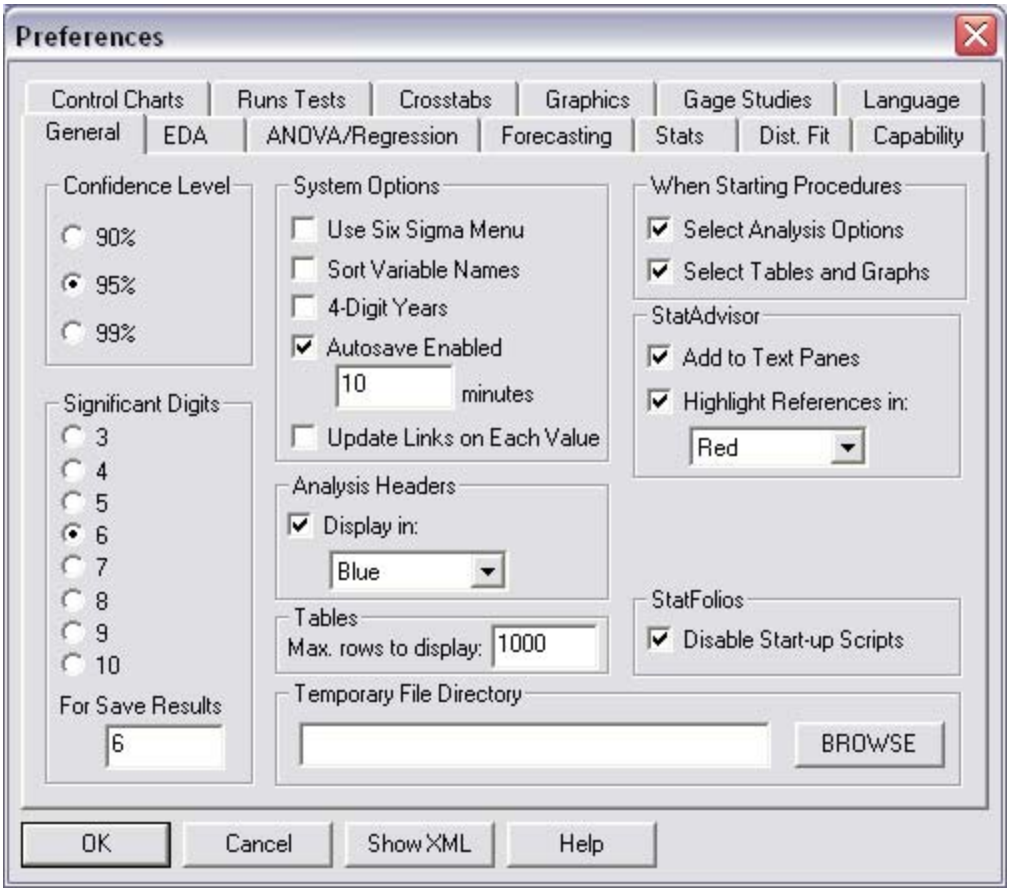


Figure 5-4. Disabling Start-Up Scripts

5.3 Polling Data Sources

Once a StatFolio has been created containing several analyses, the data in the data sources can be reread at fixed intervals of time and all of the analyses updated. This is accomplished using the *DataBook Properties* dialog box on the *Edit* menu, or by selecting *StatLink* from the *File* menu:

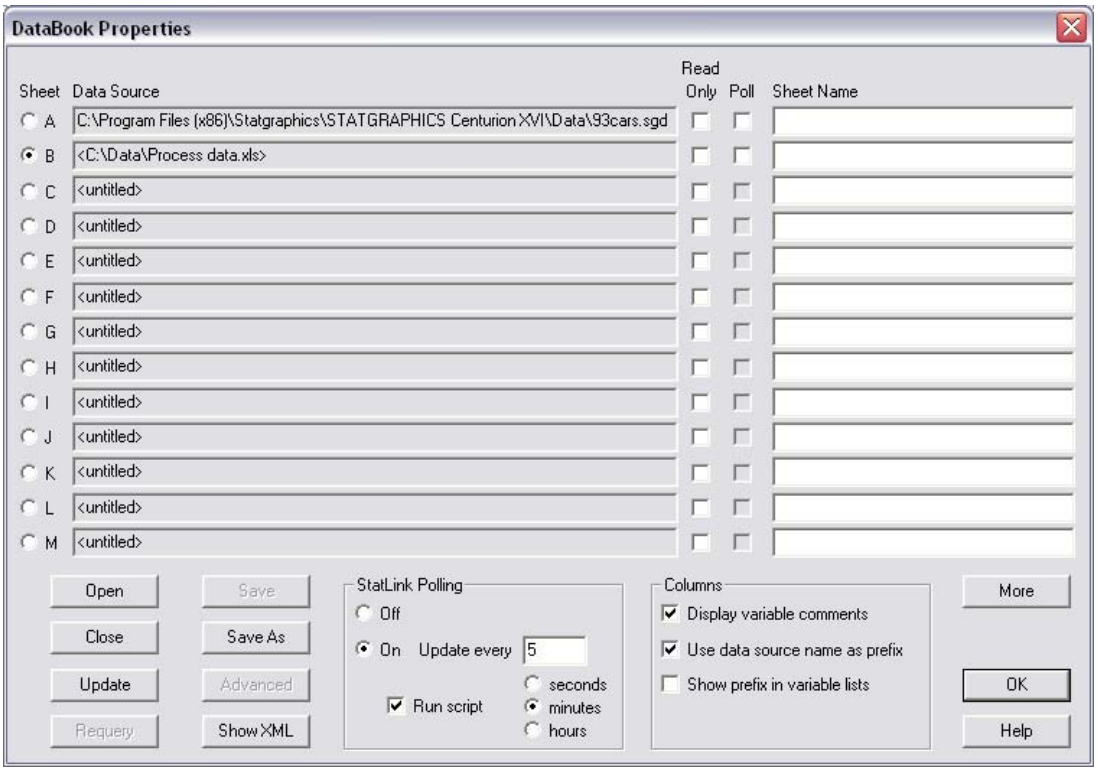


Figure 5-5. DataBook Properties Dialog Box for Polling Data Sources

To query the data sources repeatedly:

1. Place a checkmark in the *Poll* box for each data source to be reread.

2. Set the radio buttons in the *Polling* field to *On*.

3. Specify the frequency for requerying each data source.

4. Check *Run Script* if you wish to run the StatFolio start-up script each time the data is read.

By including a *Publish* step in the start-up script, you can have STATGRAPHICS Centurion XVI automatically upload the output to a network server.

5.4 Publishing Data in HTML Format

The output of a StatFolio may be published in a format that is viewable using only a standard web browser by selecting *StatPublish* from the *File* menu. A dialog box is displayed to specify which output to publish and where it should be placed:

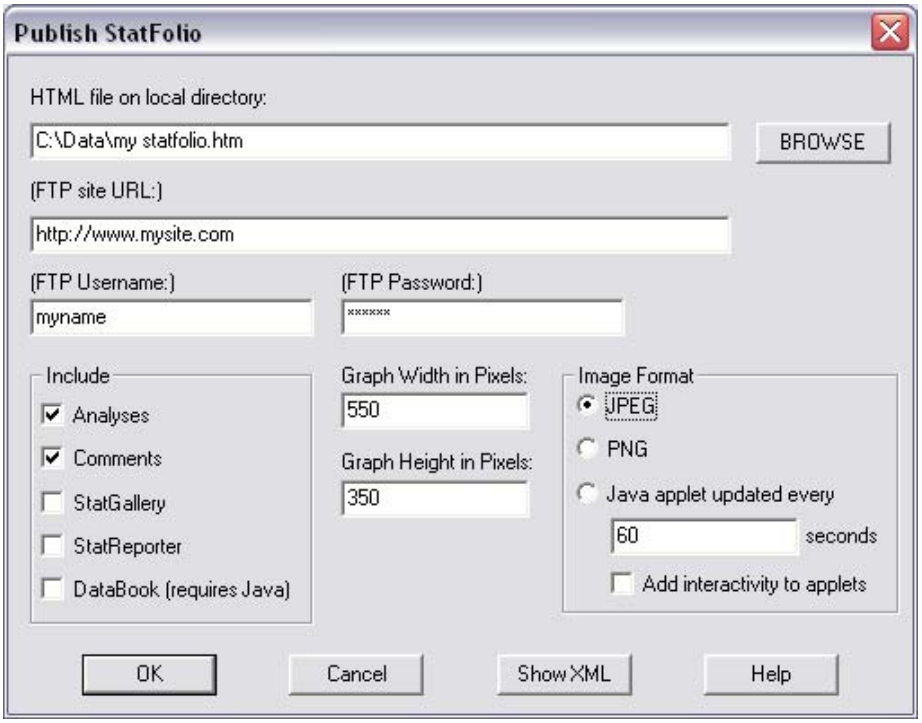


Figure 5-6. StatPublish Dialog Box for Creating HTML Output

The fields on this dialog box are used to specify:

- **HTML file on local directory**: This is the name of the HTML file that will hold the Table of Contents for the StatFolio. It will list the contents of the StatFolio and provide links to other HTML files corresponding to each window in the StatFolio. By default, it is placed in the same directory as the StatFolio itself, with the same name as the StatFolio but an extension of *.htm* rather than *.sgp*. To view a published StatFolio, a browser would normally be directed to open this file.

- **FTP site URL**: All published output is first placed in the local directory indicated above. This includes HTML files, image files containing the graphs, and other support files. If an entry is made in the *FTP Site URL* field, all of the files will also be uploaded to the location referred to by the URL. This would commonly be a directory on a server. Note that you must have FTP write access to the indicated URL, which may have to be set up by the network administrator.

- **FTP Username**: user name for FTP access to the indicated URL.

- **FTP Password**: password for FTP access to the indicated URL.

- **Include**: Check all StatFolio windows that are to be published.

- **Graph Width and Height in Pixels**: the size of the graphs when imbedded in the HTML files.

- **Image Format**: Graphs may be imbedded in the HTML files in one of three formats:

    1. *JPEG* – static images saved in JPEG format. Files are created with names such as pubexample_analysis1_graph1.jpg.

    2. *PNG* – static images saved in PNG format. Files are created with names such as pubexample_analysis1_graph1.png.

    3. *Java applets* – dynamic output that can be updated while being viewed by the browser. While in the browser, the graph will be updated at the specified increment by reading an auxiliary file with a name such as pubexample_analysis1_graph1.sgz. This option is designed to be used in conjunction with real-time polling of data using the StatLink feature, as described in the PDF document titled *Dynamic Data Processing and Analysis*. Note: not all graphs will publish properly using this option. If one or more graphs do not display correctly in the published output, select a different option.

- **Add interactivity to applets**: For graphs published as applets, selecting this feature allows the viewer to display information about data values by clicking on a point with the mouse while in the web browser.

After completing the input fields, press OK to publish the StatFolio.

To view a published StatFolio, start any web browser and use its *File* menu item to open the file specified in the top field in *Figure 5-6*. You can also view the output by selecting *View Published Results* from the STATGRAPHICS Centurion XVI *File* menu.

NOTE: Tables and graphs are imbedded in the HTML output files with names that are automatically generated by StatPublish. While in a web browser, you can view the HTML source code and easily determine the file names. These files can then be imbedded in your own web pages if you prefer.

Chapter

6

Using the StatGallery

Displaying graphs side-by-side, and overlaying graphs.

The StatGallery is a special window within STATGRAPHICS Centurion XVI where graphs created in other procedures may be pasted side-by-side or on top of one another. Side-by-side comparisons provide a powerful tool for comparing two sets of data, two statistical models, or two levels of a contour plot. Overlaying graphs creates unique displays not producible elsewhere in the system.

StatGallery output is saved in files with the extension *.sgg*. If you place output in the StatGallery, a pointer to the StatGallery file will be saved in the current StatFolio. When the StatFolio is reopened later, it will automatically load the associated StatGallery.

6.1 Configuring a StatGallery Page

The StatGallery is contained in a separate window that is created when STATGRAPHICS Centurion XVI is first loaded. It consists of one or more pages, each capable of displaying up to 9 graphs. By default, each page of the gallery is configured to display 4 graphs, as shown below:

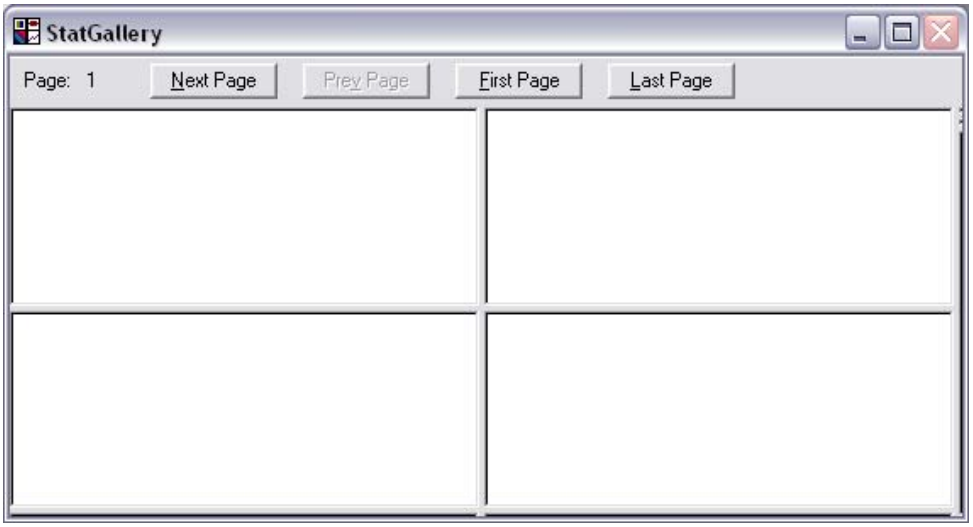


Figure 6-1. The StatGallery Window

The buttons along the top of the window permit you to navigate to other pages in the gallery. If you want to change the number of graphs displayed on a page, press the alternate mouse button and select *Arrange Panes*. Arrangements containing up to 9 graphs may be selected for a single page:

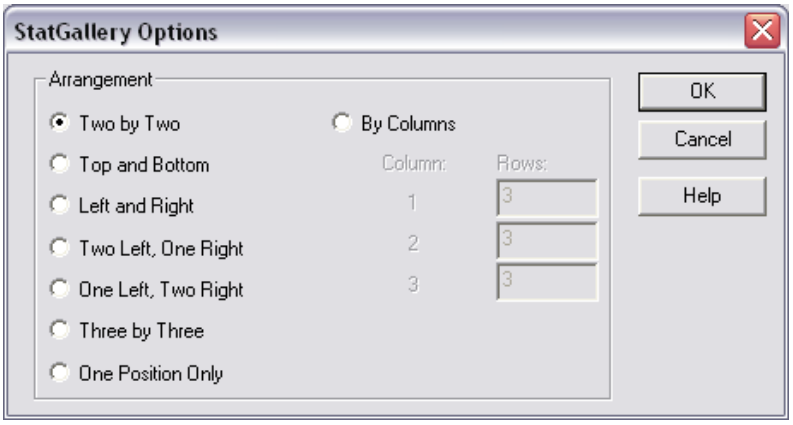


Figure 6-2. Alternative StatGallery Page Configurations

The seven arrangements on the left each correspond to rectangular sets of rows and columns. The *By Columns* option allows you to create an arrangement with different numbers of rows in each of 3 columns.

You may also use the slider bars in the StatGallery window to move the panes into any arrangement you wish.

6.2 Copying Graphs to the StatGallery

To place a graph in the StatGallery, you must first copy it to the Windows clipboard while in the analysis window where it was created. For example, suppose you wanted to display contour plots created in the DOE *Analyze Design* procedure at two different levels of a selected experimental factor. The steps to be followed are:

1. Configure a selected page of the StatGallery to show plots in a *Left and Right* format.

2. Generate a contour plot within *Analyze Design* for one level of the experimental factor and copy it to the Windows clipboard.

3. Activate the StatGallery window. Click on the leftmost pane with the alternate mouse button and select *Paste* from the popup menu to put the contour plot in the StatGallery.

4. Return to the *Analyze Design* window and generate a second contour plot at a different level of the experimental factor. Copy it to the Windows clipboard.

5. Return to the StatGallery window. Click on the rightmost pane with the alternate mouse button and select *Paste* from the popup menu. This will place the second contour plot in the StatGallery alongside the first.

The resulting display is similar to that shown below:

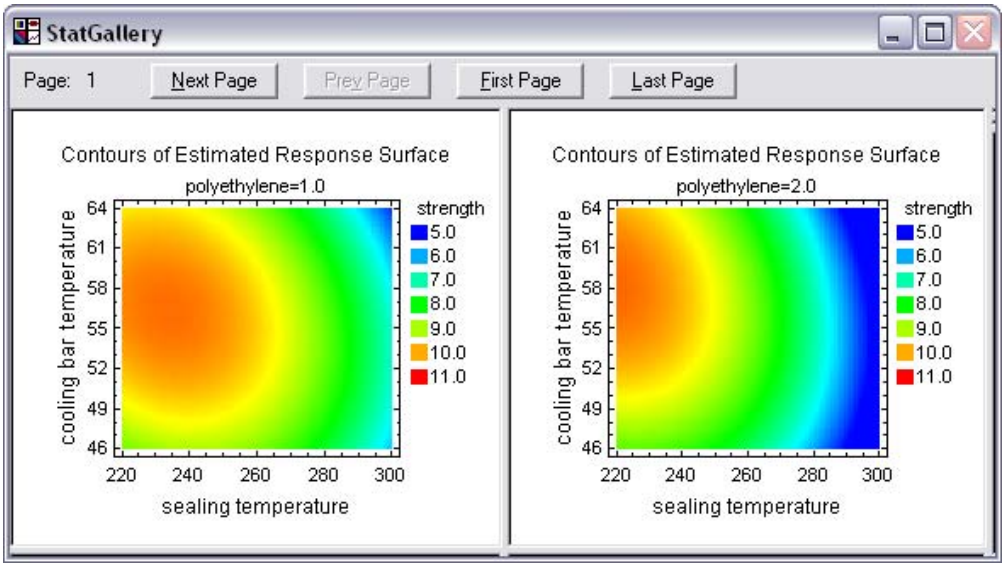


Figure 6-3. Side-by-Side Graphs in the StatGallery

In the plot above, the progression of the colors from one plot to the next shows a decrease in strength with increasing polyethylene.

When pasting a graph into the StatGallery, you may select *Paste Link* from the alternate mouse button popup menu rather than *Paste*. With paste link, the graph in the gallery is hot-linked back to the analysis window in which it was originally created and will change in the StatGallery whenever it changes in the original analysis window.

6.3 Overlaying Graphs

When a graph is pasted into a pane in the StatGallery that already contains a graph, you are given the choice of replacing the graph already there or overlaying the new graph on top of the old. Overlaying one graph on another can be useful, as when fitting two different statistical models:

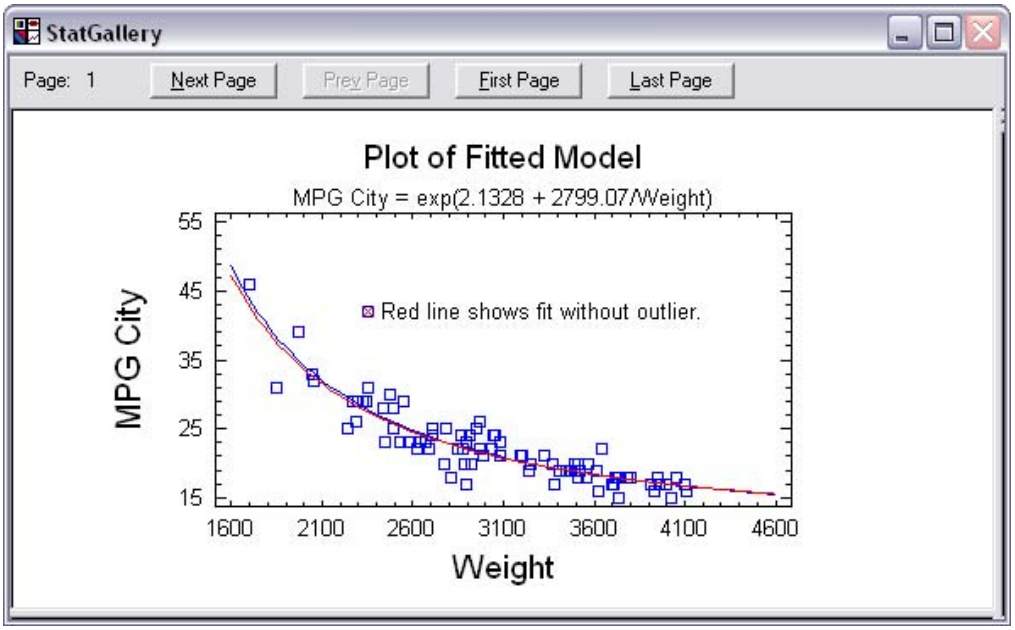


Figure 6-4. Overlaid Graphs in the StatGallery

When a graph is overlaid on top of another that is already in the StatGallery, only the contents of the second graph *inside* the axes are added to the display. Text from the second graph is **not** included.

> NOTE: If the scaling of the second graph is different than the first, the second plot will be adjusted so that it matches the first.

6.4 Modifying a Graph in the StatGallery

Certain aspects of a graph may be changed after it is pasted into the StatGallery.

6.4.1 Adding Items

To add an item to a graph:

1. Double-click on the desired graph to maximize its pane.

2. Press the alternate mouse button and select *Add Item* from the popup menu. The following floating dialog box will appear:

Figure 6-5. Add Item Dialog Box

3. Select the type of item that you want to add to the plot.

The first 5 buttons on the dialog box in *Figure 6-5* work by holding the mouse button down and stretching the line or figure until it fills the desired area. The last button activates text mode so that a text entry dialog box is displayed the next time you click on the graph. The added text may then be dragged to the desired location.

6.4.2 Modifying Items

To modify an item in the StatGallery:

1. Double-click on the desired graph to maximize its pane.

2. Click on the item to be changed with the mouse to mark it. Small rectangular blocks will be placed around an item that has been marked.

3. Press the alternate mouse button and select *Modify Item* from the popup menu.

A dialog box corresponding to the type of item marked will be displayed, on which desired changes may be indicated.

6.4.3 Deleting Items

To delete an item in the StatGallery:

1. Double-click on the desired graph to maximize its pane.

2. Click on the item to be deleted with the mouse to mark it.

3. Press the alternate mouse button and select *Delete Item* from the popup menu.

6.5 Printing the StatGallery

To print the items in the StatGallery:

1. Activate the StatGallery window by clicking on it with your mouse.

2. Press the *Print* icon on the main toolbar, or press the alternate mouse button and select *Print* from the popup menu.

You may print all of the pages or a selected set of pages.

Chapter 7

Using the StatReporter

Copying analyses to the StatReporter, annotating the output, and saving the results in an RTF file for import into Microsoft Word.

The StatReporter is a window in which output from different statistical procedures can be integrated into a formal report. It is a standalone version of WordPad, running within STATGRAPHICS Centurion XVI. The StatReporter allows you to:

1. Create a complete report within STATGRAPHICS Centurion XVI, without the necessity of using another application. This can be very useful where resources are limited, as on a production floor.

2. Save the contents of the StatReporter in an RTF (Rich Text Format) file, which can be read directly into programs such as Microsoft Word.

7.1 The StatReporter Window

The StatReporter consists of a separate window within STATGRAPHICS Centurion XVI, created automatically when the program is loaded. It consists of a single rich-edit control, together with a toolbar:

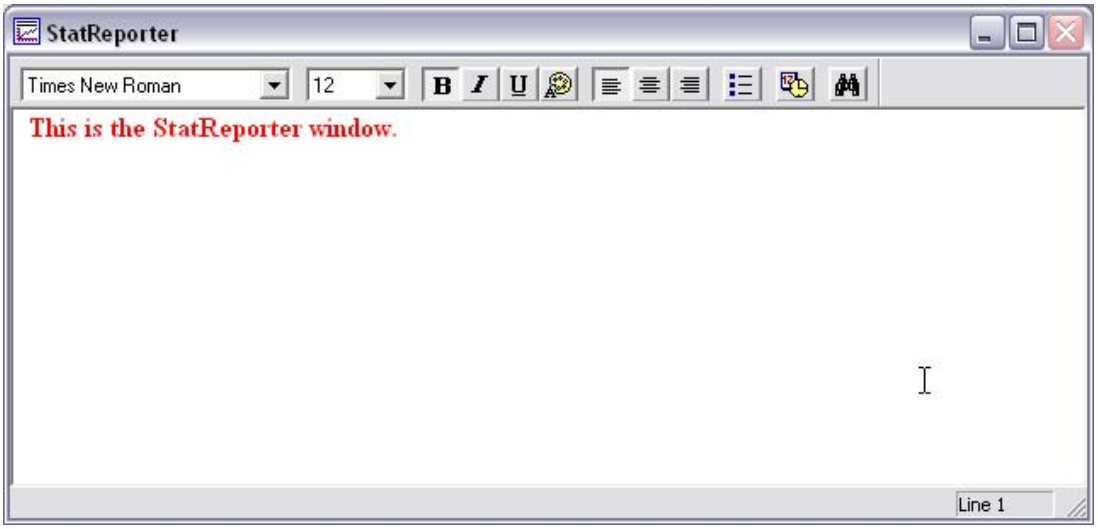


Figure 7-1. The StatReporter Window

You may type text in the window or paste output created elsewhere within STATGRAPHICS Centurion XVI.

7.2 Copying Output to the StatReporter

STATGRAPHICS Centurion XVI provides three methods for copying output to the StatReporter:

1. To copy a single table or graph to the StatReporter, first copy it to the Windows clipboard by maximizing its pane and selecting *Copy* from the *Edit* menu. Then move to the StatReporter window, put the cursor at the desired location, and select *Edit – Paste.*

2. Alternatively, maximize the pane containing the table or graph to be moved by double-clicking on it. Then press the alternate mouse button and select *Copy Pane to StatReporter* from the popup menu. This automatically pastes the table or graph into the StatReporter wherever the cursor is currently located.

3. To copy all of the output in an analysis window, press the alternate mouse button and select *Copy Analysis to StatReporter* from the popup menu. All tables and graphs in the analysis window will be pasted into the StatReporter.

Each of the above operations does a static paste (the output in the StatReporter will never change). You can link a table or graph to its source using method #1 above but selecting *Paste*

Link instead of *Paste*. The pasted table or graph in the StatReporter will then be "hot", in the sense that it will change automatically whenever the source output changes in the analysis window from which the table or graph was copied.

7.3 Modifying StatReporter Output

The StatReporter toolbar allows you to modify output once it has been placed in the window. To change text, select the text to be changed and push any of the buttons on the StatReporter toolbar. You may also insert the current date and time by pressing the *Date/Time* button.

7.4 Saving the StatReporter

To save the StatReporter output, select *File – Save – Save StatReporter* from the main menu and enter a name for the file to be saved. StatReporter contents are saved in files of type *.rtf*, which may be read directly into programs such as Microsoft Word.

Whenever a StatFolio is opened, it automatically loads the StatReporter that was present when the StatFolio was saved. You may also open a StatReporter independently using the *File – Open* menu.

Chapter

8

Using the StatWizard

Selecting the right statistical analysis, searching for desired statistics and tests, and generating multiple windows by factor levels.

The StatWizard is a special feature of STATGRAPHICS Centurion XVI designed to assist you in several ways:

1. It can help you create a new datasheet or read an existing data source.
2. It can suggest analyses based on the type of data to be analyzed.
3. It can search for desired statistics or tests and take you to analysis procedures that calculate them.
4. It can help in defining data transformations or in selecting subsets of the data.
5. It can repeat desired analyses for each unique value in a data column.

The wizard can be accessed at any time by pressing the StatWizard button  on the main toolbar.

8.1 Accessing Data or Creating a New Study

If the DataBook is empty when the StatWizard is activated, it displays a dialog box inquiring about your data needs:

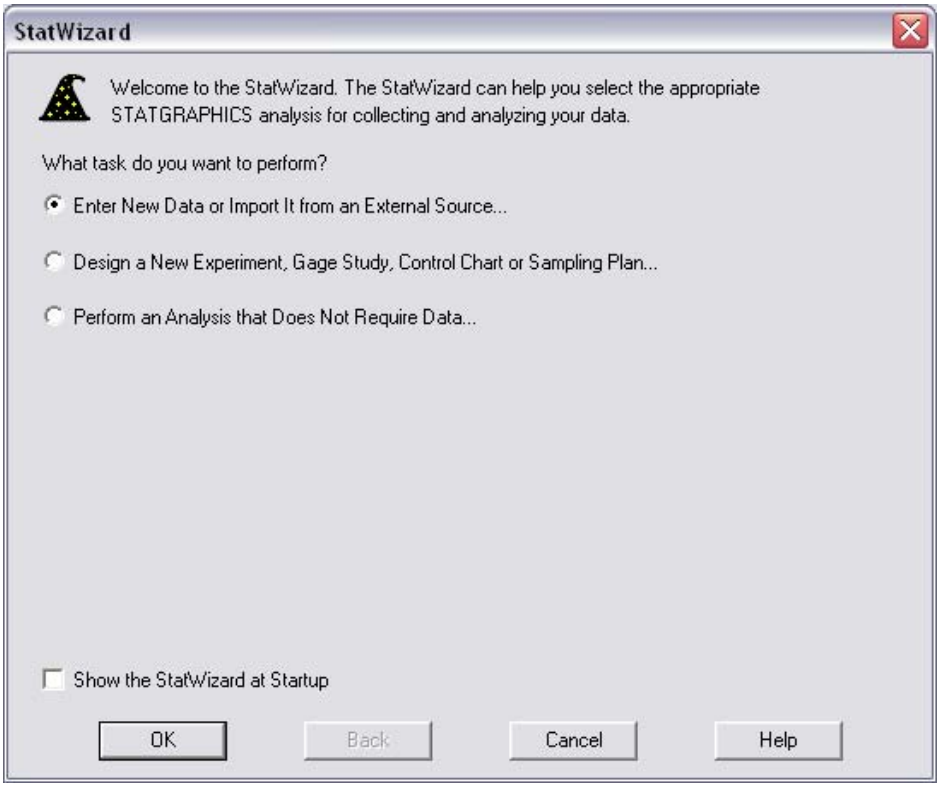


Figure 8-1. StatWizard Data Input Dialog Box

There are 3 choices:

1. You wish to load new data into the STATGRAPHICS Centurion XVI DataBook. The wizard will then take you through a sequence of additional dialog boxes in order to define the columns of a datasheet or select a data source, as described in earlier chapters of this manual.

2. You wish to design a new study before you collect data. In this case, the wizard will ask you to specify the type of study to be created and step through a sequence of dialog boxes in which you define the study to be created.

3. You wish to perform an analysis that does not require data. In this case, the wizard will list all such analyses, ask you to select one, and then take you immediately to that analysis.

For example, suppose you want to set up a new gage study in order to estimate the repeatability and reproducibility of a measurement process. Selecting the second radio button in Figure 8-1 and pressing *OK* displays the options shown below:

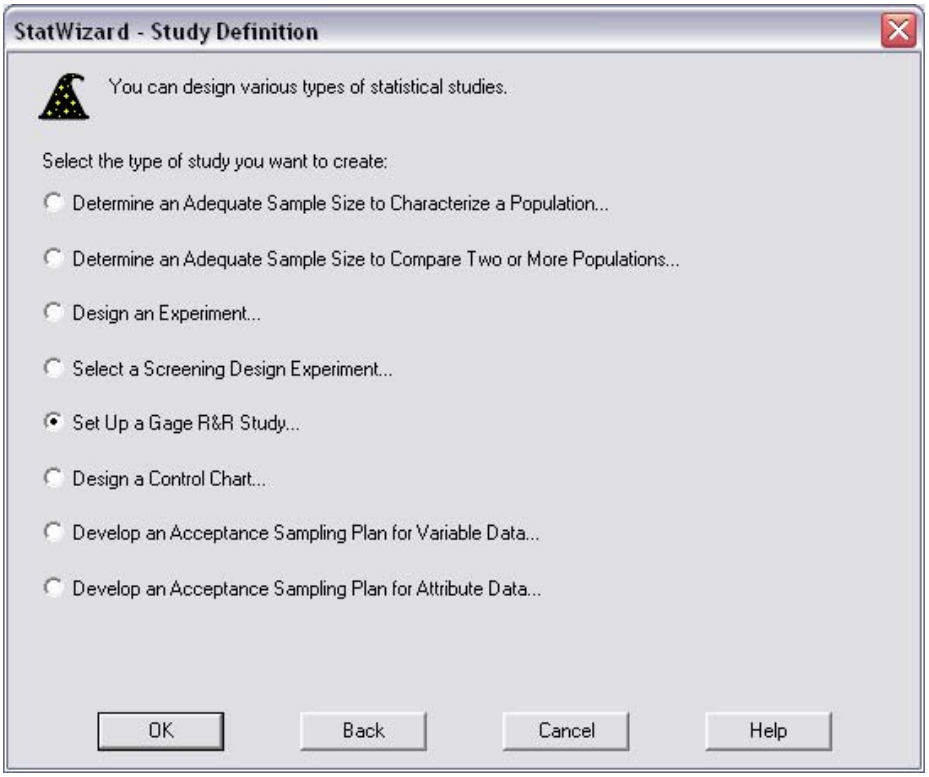


Figure 8-2. StatWizard Study Definition Dialog Box

Select *Set Up a Gage R&R Study* and press *OK* to display a third dialog box requesting information about the study:

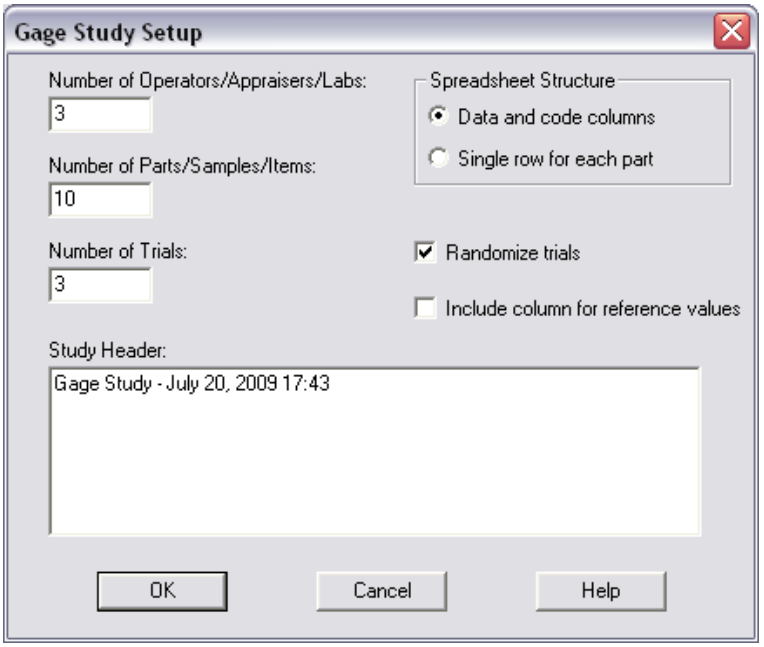


Figure 8-3. StatWizard Gage Study Setup Dialog Box

In the dialog box, enter the number of operators who will be involved in the study, the number of parts that will be measured, and the number of times each operator will measure each part. You may also specify a header for the study.

A final dialog box requests names for the operators, appraisers, or labs that will be making the measurements:

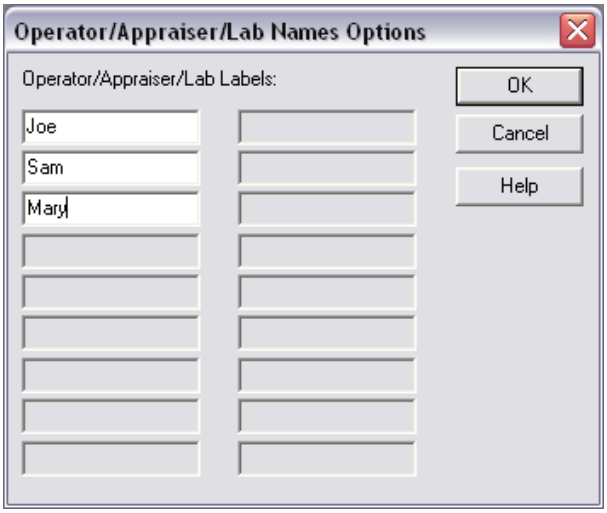


Figure 8-4. Dialog Box for Specifying Operator Names

The StatWizard then creates the desired study and places it into a datasheet in the DataBook:

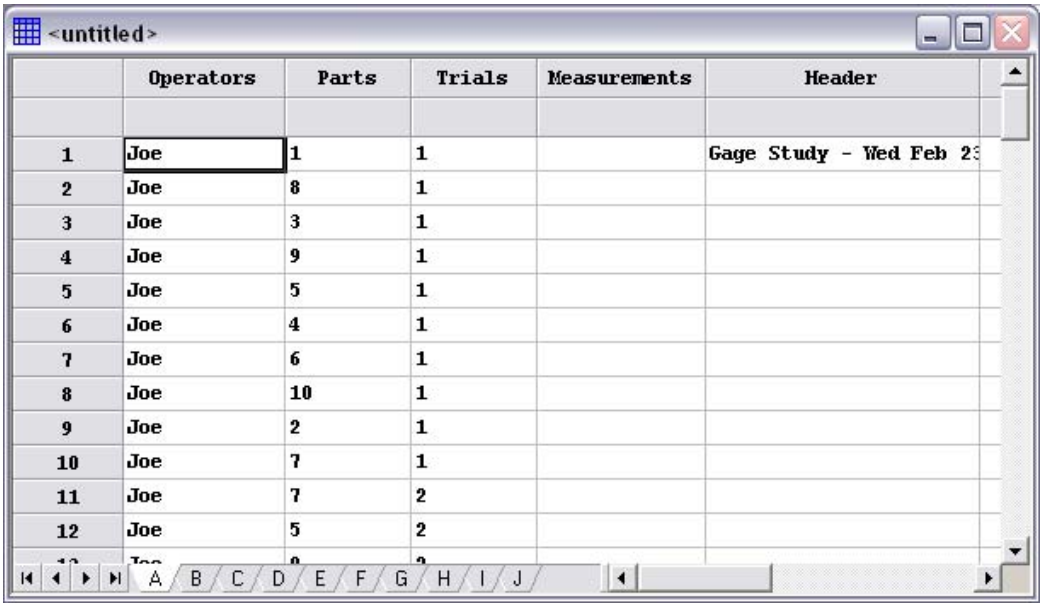


Figure 8-5. Gage Study Created by the StatWizard

The study would then be performed and measurements entered in the datasheet. The StatWizard could then be accessed again to select an analysis procedure (or you could go directly to the relevant analyses on the main menu).

8.2 Selecting Analyses for Your Data

If data has already been loaded into the DataBook, clicking on the StatWizard button displays a dialog box from which to select one or more analyses to perform:

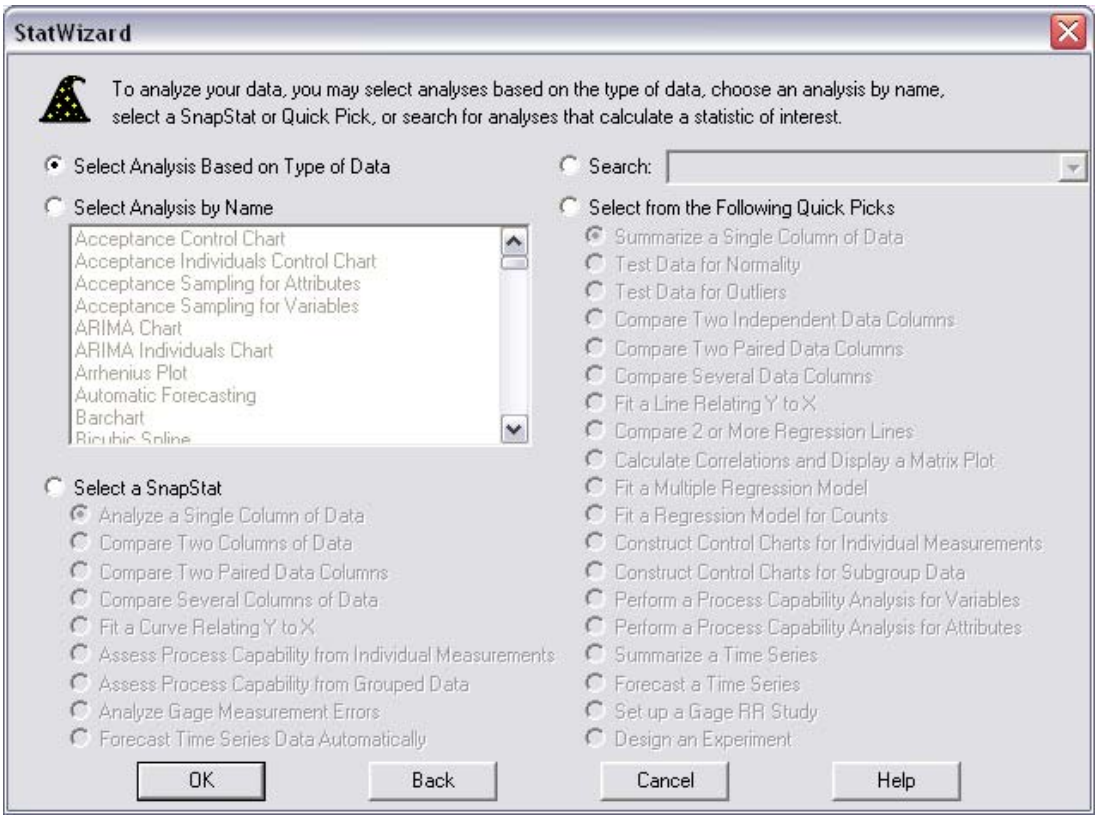


Figure 8-6. StatWizard Dialog Box for Selecting Analyses

There are five options:

1. **Select Analysis Based on Type of Data**: Displays additional dialog boxes requesting information about the data to be analyzed, after which a list of relevant procedures is presented.

2. **Select Analysis by Name**: Displays all analyses in alphabetical order. Selecting an analysis by name and pressing *OK* takes you directly to the data input dialog box for that analysis, bypassing the usual menus.

3. **Select a SnapStat**: Allows you to select a SnapStat. SnapStats are streamlined analyses that produce a single page of preformatted output. They have fewer options than other analyses but are very easy to create.

4. **Search**: Displays a pulldown list of statistics, tests, graphs, and other output that may be created in STATGRAPHICS Centurion XVI. Selecting an item from the list changes the display in the *Select Analysis by Name* field to list only those analyses that calculate the desired item.

5. **Select from the Following Quick Picks**: Lists some of the more commonly used analyses. Selecting an analysis and pressing *OK* takes you directly to the data input dialog box for that analysis.

If you elect option #1, the StatWizard will next display a dialog box in which to indicate the data to be analyzed. For example, if the *93cars.sgd* file is loaded into the DataBook, the dialog box takes the following form:

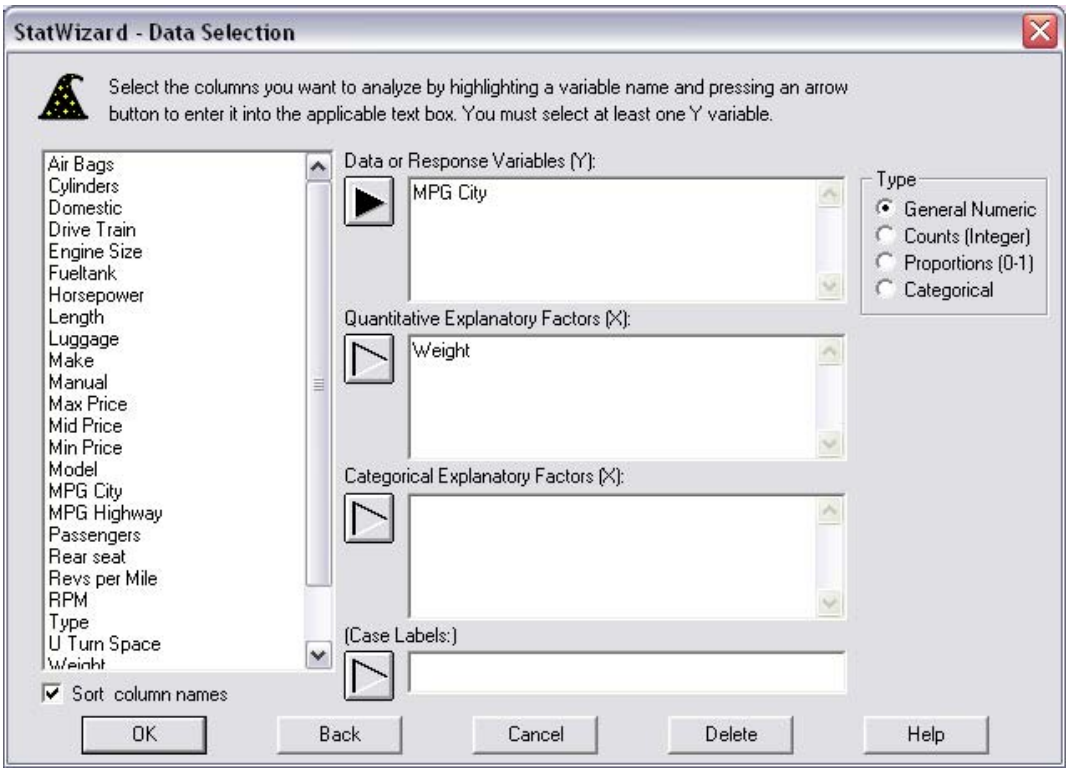


Figure 8-7. StatWizard Data Selection Dialog Box

The fields on this dialog box are:

- *Data or Response Variables (Y)*: one or more response variables containing the values to be analyzed. If only one column contains data to be analyzed, it must be entered here.

- *Type*: the type of data contained in the response variable(s). The analyses displayed in subsequent dialog boxes depend on this choice.

- *Quantitative Explanatory Factors (X)*: any quantitative factors that are to be used to predict the response variables. In a regression, the independent variables go here.

- *Categorical Explanatory Factors (X)*: any non-quantitative factors that are to be used to predict the response variables. In an ANOVA, the explanatory factors go here.

- *Case Labels*: a column containing labels for each of the observations (rows).

The procedures offered in subsequent dialog boxes depend on the data entries made in *Figure 8-7*.

The next dialog box asks you which rows of the file to analyze:

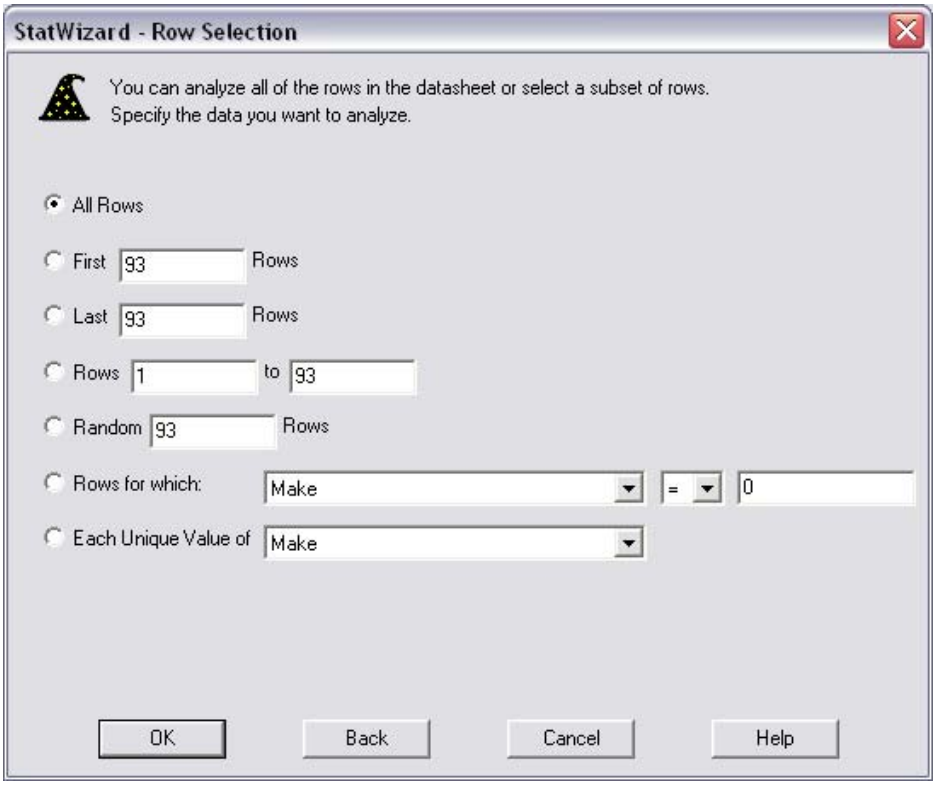


Figure 8-8. StatWizard Row Selection Dialog Box

The first six options assume that you wish to create only a single analysis. The last option will create multiple analysis windows, one for each unique value contained in the indicated column. This is an easy way to specify a "BY" variable for a set of analyses.

You will next be asked whether you wish to transform any of the indicated variables. If you reply affirmatively, the following dialog box will be displayed:

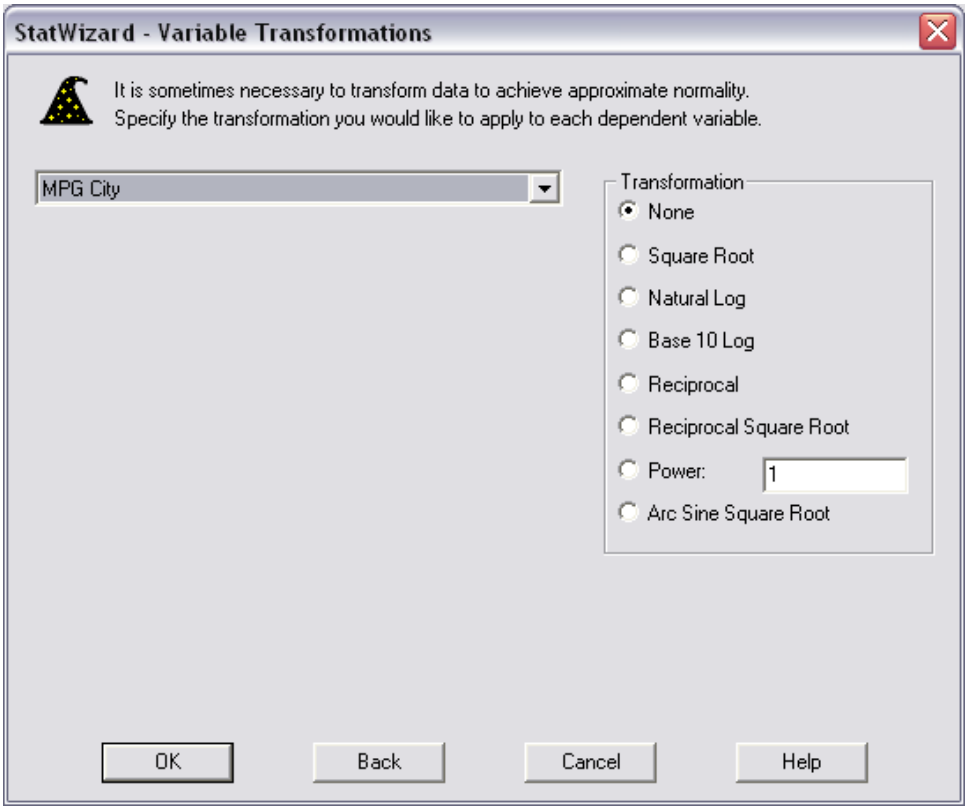


Figure 8-9. StatWizard Variable Transformation Dialog Box

You may select a transformation for one or more variables. If a transformation is requested, the appropriate expression will be created. For example, requesting a square root for *MPG City* would create the expression *SQRT(MPG City)* for use by the analysis procedures.

A final dialog box will then be displayed listing all analyses appropriate for the type of data you have specified:

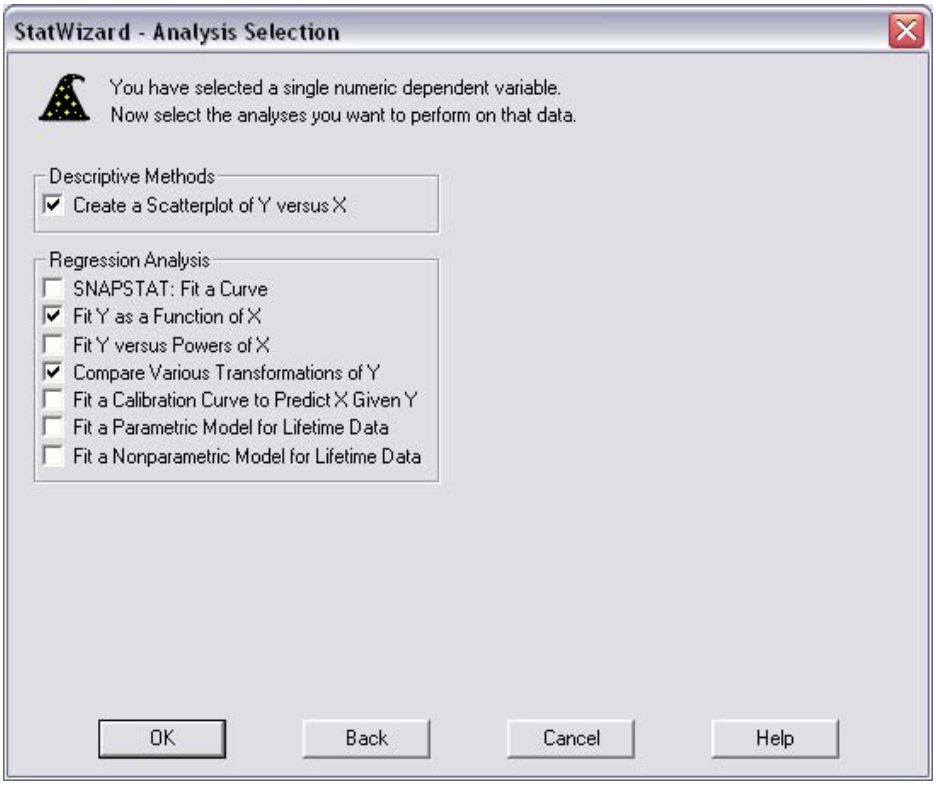


Figure 8-10. StatWizard Analysis Selection Dialog Box

Select one or more analyses from the list. When you press *OK*, an analysis window will be created for each selected analysis.

8.3 Searching for Desired Statistics or Tests

If you wish to calculate a particular statistic or test and are unsure which of the analyses calculates it, you may enter your data into a datasheet and then press the *StatWizard* button on the main toolbar. On the initial StatWizard dialog box, select *Search* and pull down the list. A list of all statistics, tests and other calculations performed by STATGRAPHICS Centurion XVI will be displayed:

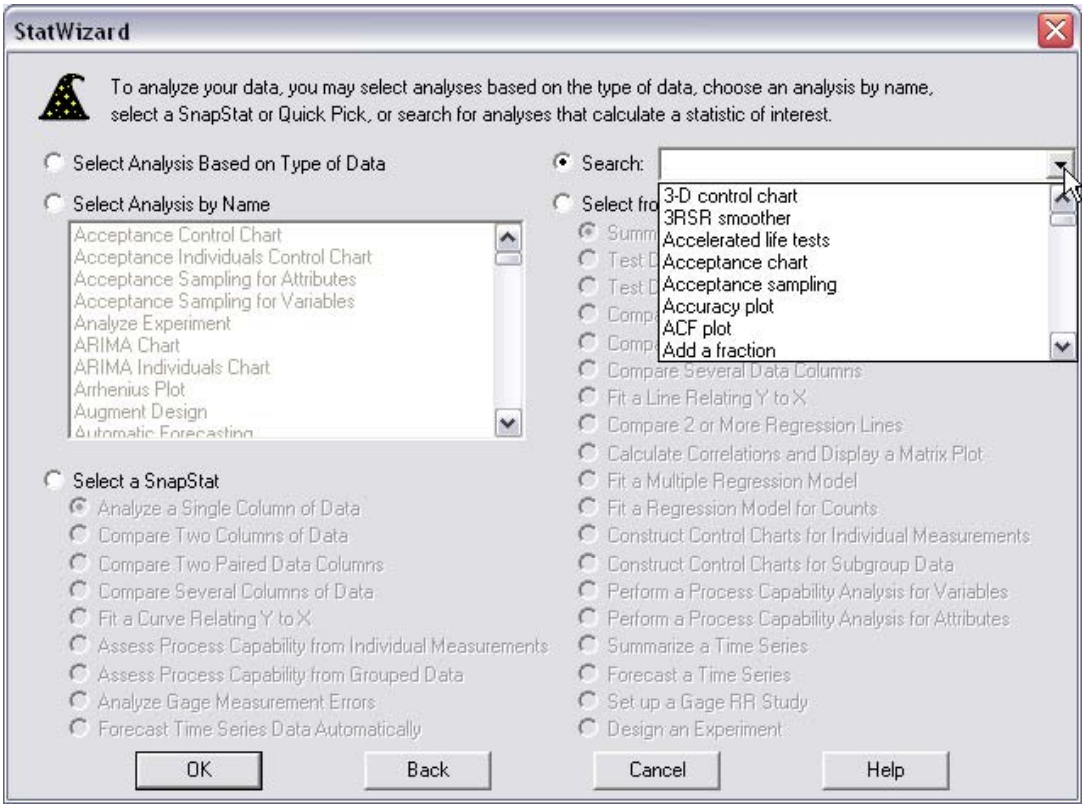


Figure 8-11. Using the StatWizard Search Option

If you select an item from the list, all analyses that calculate the selected item will be displayed in the *Select Analysis by Name* field:

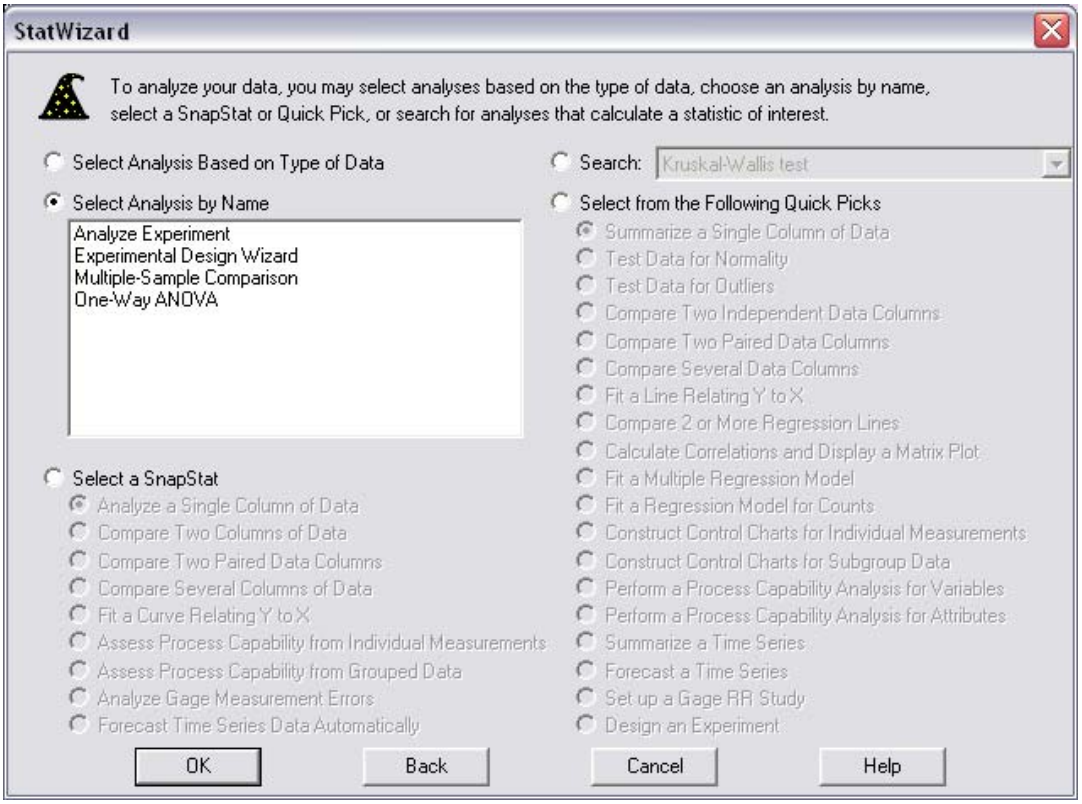


Figure 8-12. List of All Analyses Matching the Search Option

To run a selected analysis:

1. Click on the *Select Analysis by Name* radio button.
2. Highlight an analysis.
3. Press *OK*.

You will be taken directly to the data input dialog box for the selected analysis, bypassing the usual menus.

Chapter

9

System Preferences

Setting preferences for system behavior.

STATGRAPHICS Centurion XVI contains hundreds of options, each of which has a default value that has been selected to meet most users' needs. If desired, you can set new defaults for many of these options. There are 3 main places in the program to do this:

1. **General system behavior**: set on the *Preferences* dialog box accessible from the *Edit* menu.

2. **Printing options**: set on the *Page Setup* dialog box accessible from the *File* menu.

3. **Graphs**: set by selecting *Graphics Options* while viewing any graph. The *Profile* tab of the *Graphics Options* dialog box allows you to save multiple sets of graphics attributes.

9.1 General System Behavior

The default values for general system behavior and selected statistical procedures may be changed by selecting *Preferences* from the *Edit* menu. This displays a tabbed dialog box with a *General* tab for overall system behavior and other tabs for statistical analysis defaults:

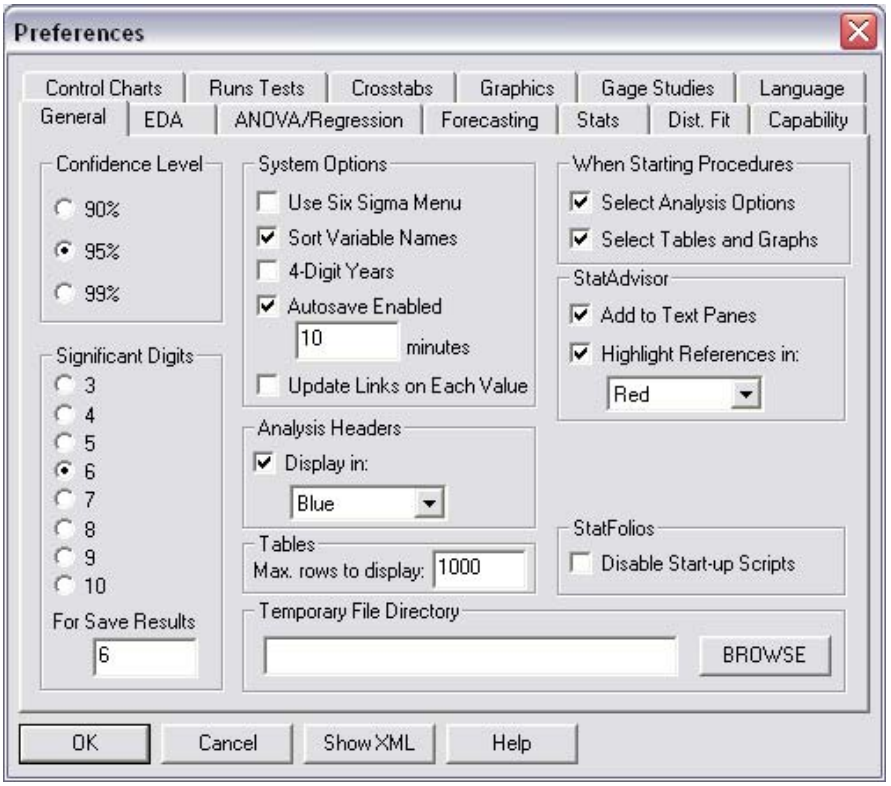


Figure 9-1. Preferences Dialog Box

Some of the most important options that may be set are:

- **Confidence Level**: default percentage used for confidence limits, prediction limits, hypothesis tests, and interpretation of *P*-values by the StatAdvisor.

- **Significant Digits**: number of significant digits used when displaying numerical results. The indicated number of digits will be displayed, except for trailing zeroes that will be dropped. A separate entry is provided for saving numerical results back to the datasheet.

- **System Options**: options that apply system-wide.

    - **Use Six Sigma Menu**: display the menu selections under headings corresponding to the Six Sigma DMAIC arrangement (Define, Measure, Analyze, Improve, Control). The same selections are available as with the classic menu, except they are arranged under different menu headings.

- **Sort Variable Names**: whether to list column names in alphabetic order on data input dialog boxes. Otherwise, column names will be listed in the same order as in the datasheets.

- **4-Digit Years**: whether dates should be displayed with 4-digit years rather than 2-digit years. By default, 2-digit years such as 2/1/05 are assumed to represent dates between the years 1950-2049. Changes to this option will not take effect until the program is restarted.

- **Autosave Enabled**: whether to save the current *StatFolio* and data files automatically in the background, and the duration of time between saves. If enabled and there is a computer or program malfunction, you will be given the chance to restore the state of the StatFolio and datasheets when the program is next restarted.

- **Update Links on Each Value**: whether to recalculate all statistics whenever a data value changes in one of the datasheets. Normally, statistics are not recalculated until an analysis receives the focus, is printed or published, or the StatFolio is saved.

- **StatAdvisor**: sets the default behavior of the StatAdvisor.

  - **Add to Text Panes**: whether StatAdvisor output should automatically be added to the bottom of text panes. StatAdvisor output is always available by pressing the button on the main toolbar showing the graduation cap.

  - **Highlight References in …**: whether to highlight in a special color values on text panes that are referred to by the StatAdvisor.

- **Analysis Headers**: whether to use a blue font to display the analysis title at the top of the *Analysis Summary* pane.

- **StatFolios**: check *Disable Start-Up Scripts* to prevent start-up scripts from being run when StatFolios are loaded.

- **Temporary File Directory**: If specified, StatFolios, data files, and other files will first be written to this directory before being copied to their final location. By specifying a local drive, this can greatly speed up the time required to save a file over some networks, since it reduces the number of network requests.

For a description of the options on the other tabs, refer to the PDF document titled *Preferences*.

9.2 Printing

Two selections on the *File* menu control printed output:

1. *Print Setup*: accesses the standard printer options dialog box supplied with your printer driver. This dialog box typically sets paper size and chooses between *landscape* and *portrait* mode for the output.

2. *Page Setup*: a STATGRAPHICS Centurion XVI specific dialog box that sets margins, headers, and other options. This dialog box was discussed in Section 3.3.

9.3 Graphics

Maximizing a pane containing a graph within any analysis window activates the *Graphics Options* button on the analysis toolbar. That button displays a tabbed dialog box that allows you to change the appearance of a graph, as described in detail in Chapter 4. Also included on that dialog box is a tab labeled *Profile*, which enables you to save sets of graphics attributes in user profiles and change the default profile that is used when a new graph is created:

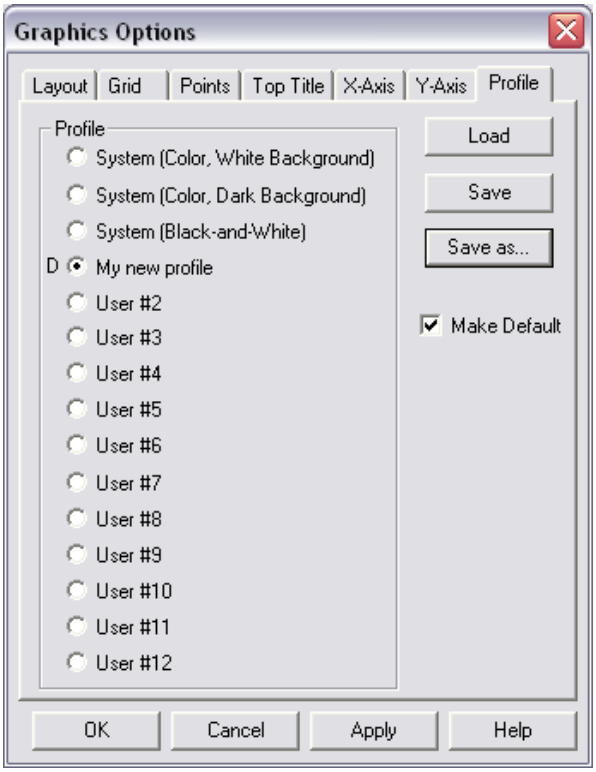


Figure 9-2. Profile Tab on Graphics Options Dialog Box

To change the system defaults:

1. Modify the features of a graph in any analysis window. Set the colors, fonts, and other options that you want future graphs to reflect.

2. Select *Graphics Options* from the analysis toolbar and go to the *Profile* tab.

3. Check *Make Default.*

4. Select any of the 12 user profiles and press the *Save as* button (the system profiles are read-only).

5. Enter a name for the profile to be saved:

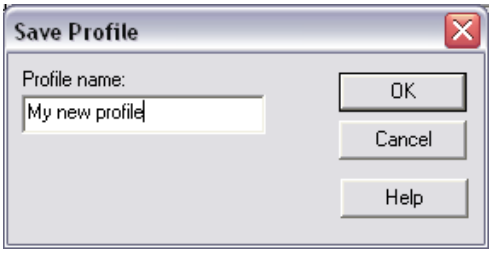


Figure 9-3. Save Profile Dialog Box

6. Press *OK* to save the current set of graphics attributes (colors, fonts, point and line styles, etc.) in a new profile.

The next graph created will use the newly saved profile.

You can also apply other saved profiles to a new graph by creating the graph with default settings and then:

1. Select *Graphics Options* from the analysis toolbar and go to the *Profile* tab.

2. Select any of the 15 profiles and press the *Load* button.

The current graph will be immediately updated to reflect the settings in the selected profile.

Chapter

10

Tutorial #1: Analyzing a Single Sample

Summary statistics, histogram, box-and-whisker plot, confidence intervals, and hypothesis tests.

A very common problem in statistics is that of analyzing a sample of n observations taken from a single population. For example, consider the following body temperatures taken from $n = 130$ individuals:

98.4	98.4	98.2	97.8	98	97.9	99	98.5	98.8	98
97.4	98.8	99.5	98	100.8	97.1	98	98.7	98.9	99
98.6	97.7	96.7	98.8	98.2	97.5	97.2	97.4	97.1	96.7
99.2	97.9	98.8	97.6	98.6	98.8	98.5	98.7	97.5	97.9
97.1	98.4	97.4	98.6	97.8	98.2	98	98	98.3	98.6
98.8	98.7	98.8	98.1	96.4	98.8	98.7	97.9	98.6	99.2
98.6	98	99.1	97.8	97.2	98.2	98.7	98.4	98.2	97.7
98.3	98.7	96.8	98	97.2	97.9	96.9	98.3	97.8	97
98.6	98.4	98.2	98	98	98.2	97.8	99	98.1	97.7
97.4	98.8	99.3	98.9	96.3	97.8	99.9	98.4	99.4	98.7
98.4	98.2	99.3	98.5	98.3	99	99.2	97.6	99.1	97.6
98.4	97.6	98.4	98	98.8	97.3	98.7	98.6	99.4	100
98.6	98.3	98.6	97.4	98.1	97.8	98.2	99	99.1	98.2

The data were obtained from the Journal of Statistical Education Data Archive (www.amstat.org/publications/jse/jse_data_archive.html) and are used by permission. It has

been placed in a file named *bodytemp.sgd*, in a column called *Temperature* that contains 130 rows, one row for each person in the study.

The primary procedure in STATGRAPHICS Centurion XVI for summarizing a sample taken from a population is the *One-Variable Analysis* procedure. The *One-Variable Analysis* procedure summarizes the data in both numerical and graphical form and tests hypotheses about the population mean, median, and standard deviation.

10.1 Running the One-Variable Analysis Procedure

To analyze the body temperature data, first load the *bodytemp.sgd* file into a datasheet. To accomplish this:

1. Select *File – Open – Open Data Source* from the main menu.
2. On the *Open Data Source* dialog box, indicate that you wish to open a *STATGRAPHICS Data File*.
3. Select *bodytemp.sgd* from the list of files on the *Open Data File* dialog box.

The data should appear as shown below:

bodytemp

	Temperature	Gender	Heart Rate	Col_4	Col_5
	degrees		beats per minute		
1	98.4	Male	84		
2	98.4	Male	82		
3	98.2	Female	65		
4	97.8	Female	71		
5	98	Male	78		
6	97.9	Male	72		
7	99	Female	79		
8	98.5	Male	68		
9	98.8	Female	64		
10	98	Male	67		
11	97.4	Male	78		
12	98.8	Male	78		

A B C D E F G H I J

Figure 10-1. Datasheet with Body Temperature Data

The body temperatures are in the leftmost column, measured in degrees Fahrenheit.

The *One-Variable Analysis* procedure can be accessed from the main menu as follows:

1. If using the Classic menu, select *Describe – Numeric Data – One-Variable Analysis*.
2. If using the Six Sigma menu, select *Analyze – Variable Data – One-Variable Analysis*.

On the data input dialog box, indicate the column to be analyzed:

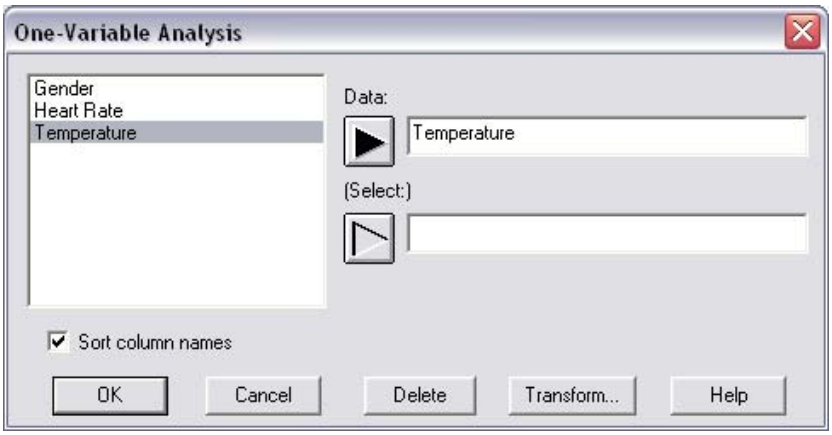


Figure 10-2. One-Variable Analysis Data Input Dialog Box

Leave the *Select* field blank to analyze all 130 rows. Press *OK*.

When *OK* is pressed, the *Tables and Graphs* window appears. This window shows the tables and graphs that are available. For now, the default setting will be acceptable.

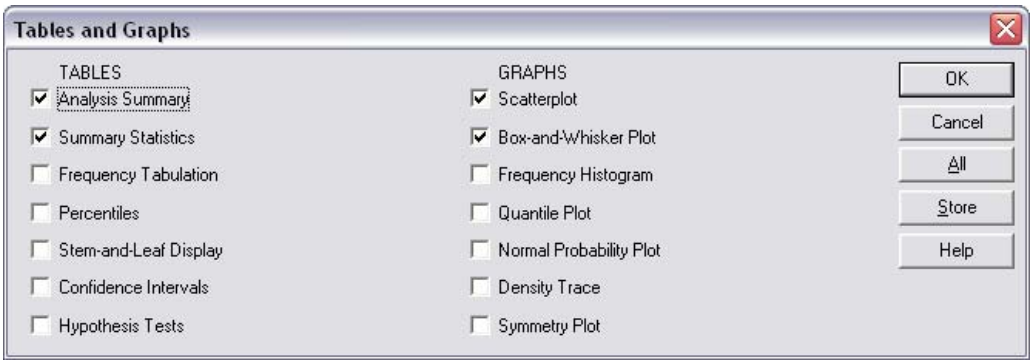


Figure 10-3. Tables and Graphs Dialog Box

An analysis window will appear with four panes:

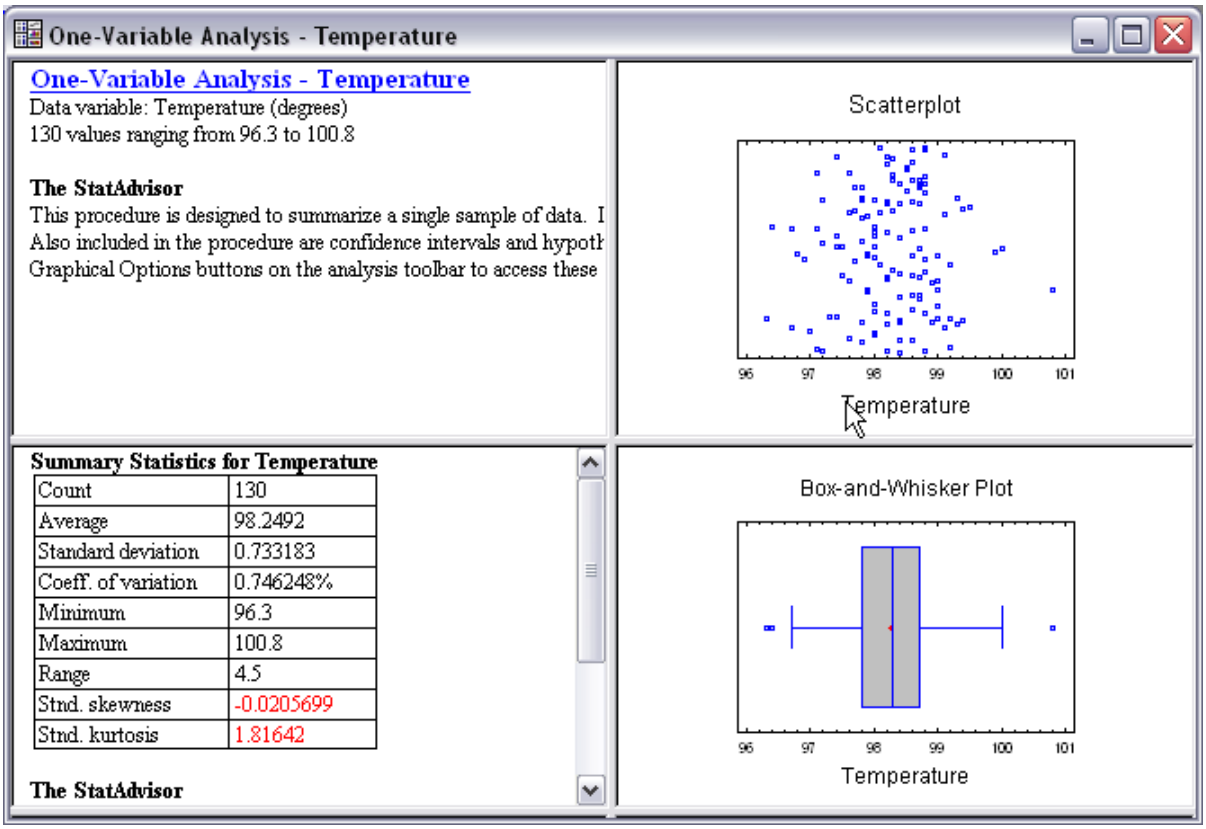


Figure 10-4. One-Variable Analysis Window

The top left pane indicates that the sample has $n = 130$ values ranging between 96.3 and 100.8 degrees. The top right pane shows a scatterplot of the data, with the points randomly scattered in the vertical direction. Note that the points are densest between 98 and 99 degrees, thinning out at either end. This type of behavior is typical of data that are sampled from a population whose distribution has a well-defined central peak.

The bottom panes show summary statistics and a box-and-whisker plot, described in the following sections.

10.2 Summary Statistics

The table in the bottom left pane displays several sample statistics. Additional statistics can be added by maximizing that pane (double-click on it with your mouse) and selecting *Pane Options*:

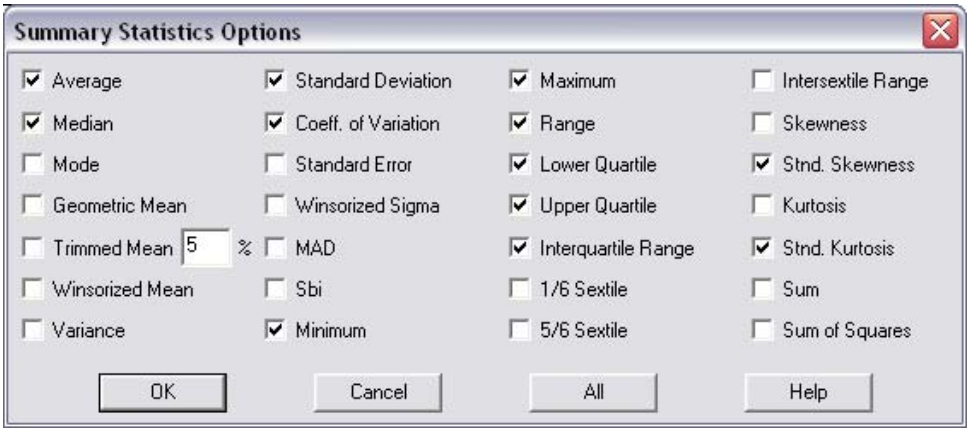


Figure 10-5. Summary Statistics Options Dialog Box

Including the sample median, quartiles, and the interquartile range results in:

Summary Statistics for Temperature	
Count	130
Average	98.2492
Median	98.3
Standard deviation	0.733183
Coeff. of variation	0.746248%
Minimum	96.3
Maximum	100.8
Range	4.5
Lower quartile	97.8
Upper quartile	98.7
Interquartile range	0.9
Stnd. skewness	-0.0205699
Stnd. kurtosis	1.81642

Figure 10-6. Summary Statistics Table

A common assumption for measurement data is that it comes from a normal or Gaussian distribution, i.e., from a bell-shaped curve. Data from a normal distribution are fully described by two statistics:

1. The sample *mean* or *average* $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 98.25$, which estimates the center of the distribution.

2. The sample *standard deviation* $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = 0.733$, which is related to the spread of the distribution.

For a normal distribution, approximately 68% of all values will lie within one standard deviation of the population mean, approximately 95% within two standard deviations, and approximately 99.73% within three standard deviations.

The sample mean and standard deviation fully describe the sample only if it comes from a normal distribution. Two statistics that may be used to check this assumption are the standardized skewness and standardized kurtosis. These statistics measure shape:

1. *Skewness* measures symmetry or lack thereof. A symmetric distribution, such as the normal, has zero skewness. Distributions in which values tend to lie farther *above* the peak than below have positive skewness. Distributions in which values tend to lie farther *below* the peak than above have negative skewness.

2. *Kurtosis* measures the shape of a symmetric distribution. A normal, or bell-shaped curve, has zero kurtosis. A distribution that is more *peaked* than the normal has positive kurtosis. A distribution that is *flatter* than the normal has negative kurtosis.

If the data come from a normal distribution, both the standardized skewness and standardized kurtosis should be within the range of -2 to +2. In this case, the normal distribution appears to be a reasonable model for the data.

Another useful summary of the data is provided by John Tukey's five number summary:

Minimum (smallest data value) = 96.3
Lower quartile (25th percentile) = 97.8
Median (50th percentile) = 98.3
Upper quartile (75th percentile) = 98.7
Maximum (largest data value) = 100.8

These five numbers divide the sample into quarters and form the basis of his box-and-whisker plot, described in the next section.

NOTE: Selecting additional summary statistics using *Pane Options* changes the selection for the current analysis only. To change the default statistics for future analyses, go to the *Edit* menu and select *Preferences*. The *Stats* tab on that dialog box allows you to change the statistics calculated by default when the *One-Variable Analysis* is run (as well as several other procedures that display summary statistics):

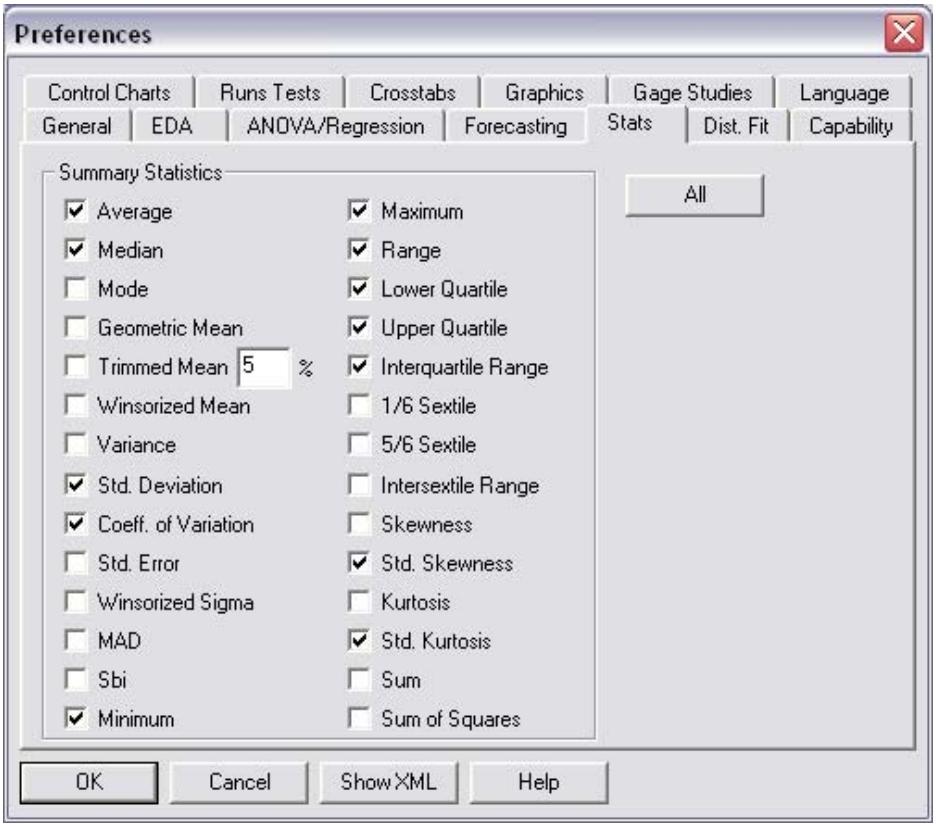


Figure 10-7. Preferences Dialog Box Used to Select Default Statistics

10.3 Box-and-Whisker Plot

A useful graphical display for summarizing data, invented by John Tukey, is the box-and-whisker plot displayed in the lower right corner of *Figure 10-4* and enlarged below:

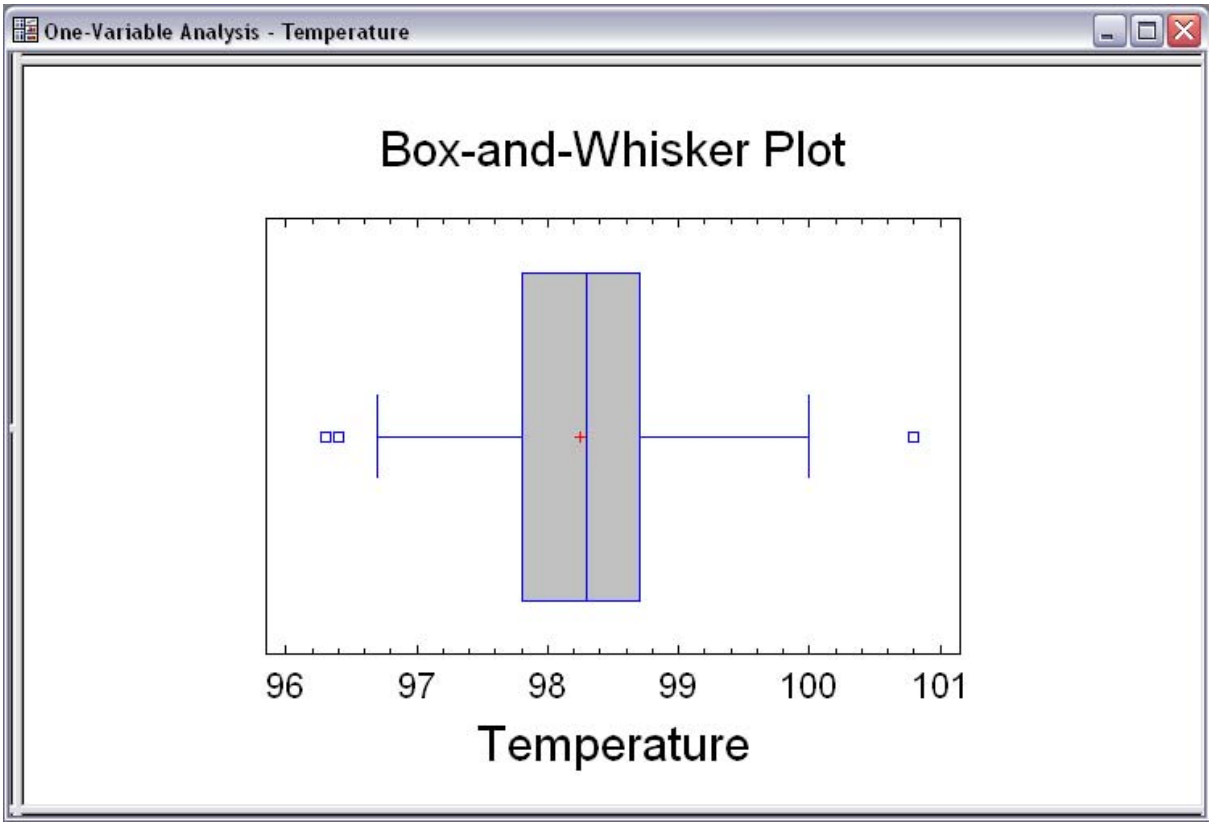


Figure 10-8. Box-and-Whisker Plot for Body Temperatures

The box-and-whisker plot is constructed by:

1. Drawing a box extending from the lower quartile to the upper quartile. The middle 50% of the data values are thus covered by the box.

2. Drawing a vertical line at the location of the sample median, which divides the data in half. If the data come from a symmetric distribution, this line should be close to the center of the box.

3. Drawing a plus sign at the location of the sample mean. Any substantial difference between the median and the mean usually indicates either the presence of an outlier (a data value that does not come from the same population as the rest) or a skewed distribution. In the case of a skewed distribution, the mean will be pulled in the direction of the longer tail.

4. Whiskers extending from the quartiles to the largest and smallest observations in the sample, unless some values are far enough from the box to be classified as "outside points", in which case the whiskers extend to the most extreme points that are not classified as "outside". STATGRAPHICS Centurion XVI follows Tukey in flagging two types of unusual points:

    a. "Far outside" points – points more than 3 times the interquartile range above or below the limits of the box. (Note: the interquartile range is the distance between the quartiles, which is equal to the width of the box.) Far outside points are denoted by a point symbol (usually a small square) with a plus sign superimposed on it. If the data come from a normal distribution, the chance that any point will be far enough away from the box to be classified as far outside is only about 1 in 300 in a sample of the current size. Unless there are thousands of observations in the sample, far outside points are usually indicative of true outliers (or of a non-normal distribution).

    b. "Outside" points - points more than 1.5 times the interquartile range above or below the limits of the box. Outside points are denoted by a point symbol but no superimposed plus sign. Even when data come from a normal distribution, the chance of observing 1 or 2 outside points in a sample of $n = 100$ observations is about 50% and does not necessarily indicate the presence of a true outlier. These points should be considered simply worthy of further investigation.

The box-and-whisker plot in *Figure 10-8* is reasonably symmetric. The whiskers are about the same length and the sample mean and median are similar and close to the center of the box. Three outside points are marked, but no far outside points. Clicking on the rightmost outlier with the mouse indicates that it corresponds to row #15 in the file.

If you select *Pane Options* from the analysis toolbar, you can add a median notch to the plot:

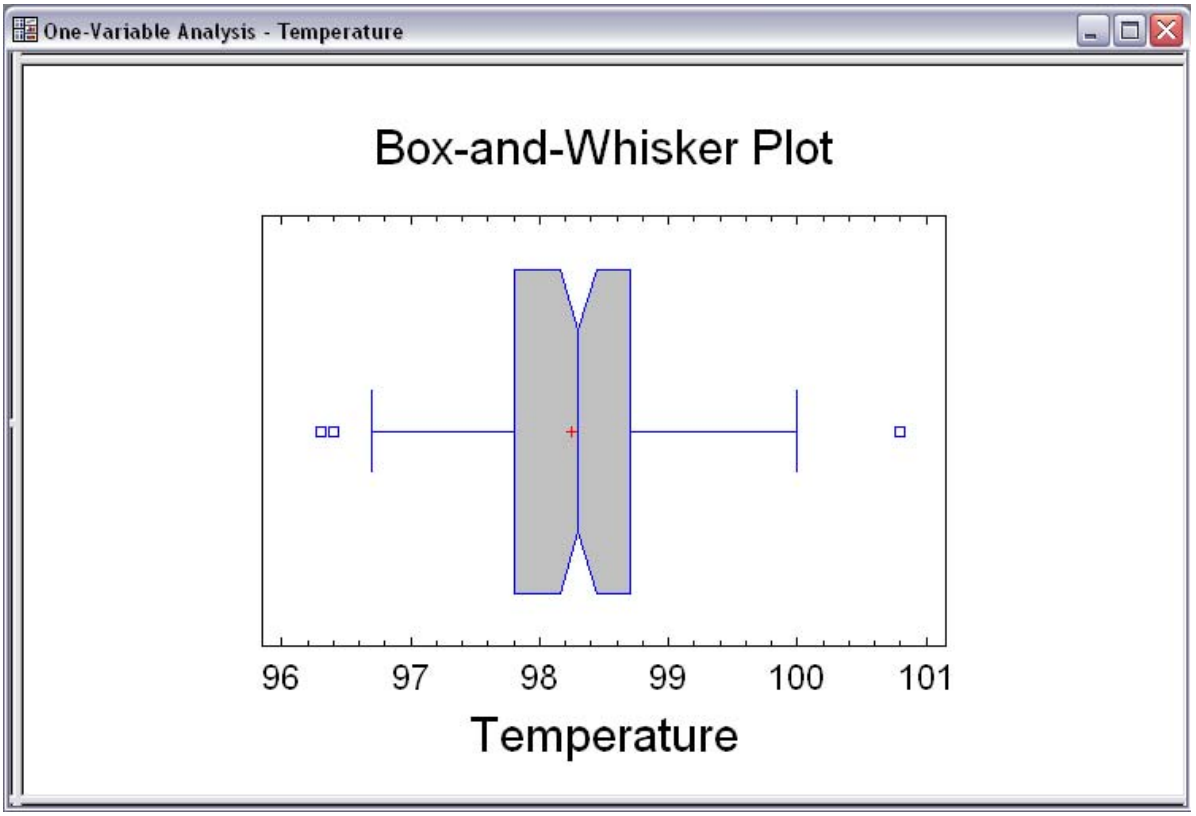


Figure 10-9. Box-and-Whisker Plot with a 95% Median Notch

This adds a notch to the display covering an approximate confidence interval for the population median, at the default system confidence level (usually 95%). It shows the margin of error when attempting to estimate the median temperature of the population from which the sample was taken. In this case, the sampling error is about 0.15 degrees in either direction. A larger sample would result in a smaller margin of error.

10.4 Testing for Outliers

Before estimating any additional statistics, it is worth taking a moment to investigate whether row #15 should be considered a true outlier and potentially removed from the data set. STATGRAPHICS Centurion XVI includes a procedure that performs a formal test to determine whether an observation could reasonably have come from a normal distribution. The test is available on the main menu by selecting:

1. If using the classic menu, select *Describe – Numeric Data – Outlier Identification*.

2. If using the Six Sigma menu, select *Analyze – Variable Data – Outlier Identification.*

Specifying *Temperature* in the *Data* field generates the *Options* window, then the *Tables and Graphs* window. After all desired options are selected, a large table of statistics is generated and displayed in the lower half of the left pane. Of particular interest is the table showing the 5 smallest values in the sample and the 5 largest values:

Sorted Values

		Studentized Values	*Studentized Values*	*Modified*
Row	*Value*	*Without Deletion*	*With Deletion*	*MAD Z-Score*
95	96.3	-2.65859	-2.74567	-2.698
55	96.4	-2.52219	-2.59723	-2.5631
23	96.7	-2.11302	-2.15912	-2.1584
30	96.7	-2.11302	-2.15912	-2.1584
73	96.8	-1.97663	-2.01521	-2.0235
...				
99	99.4	1.56955	1.59096	1.4839
13	99.5	1.70594	1.7323	1.6188
97	99.9	2.25151	2.30628	2.1584
120	100.0	2.3879	2.45231	2.2933
15	100.8	3.47903	3.67021	3.3725

Grubbs' Test (assumes normality)
Test statistic = 3.47903
P-Value = 0.0484379

Figure 10-10. Selected Output from Outlier Identification Procedure

The most unusual value is row #15, which is highlighted in red. It has a *Studentized Value Without Deletion* of 3.479. Studentized values are calculated from:

$$z_i = \frac{x_i - \bar{x}}{s}$$

A value of 3.479 indicates that an observation is 3.479 sample standard deviations above the sample mean, when the observation is included in the calculation of $\bar{x}$ and s. The *Studentized Values With Deletion* indicate how many standard deviations each observation lies from the sample mean when that observation is *not* used in the calculations. If not included in the calculation, row #15 is 3.67 standard deviations out.

Observations more than 3 standard deviations from the mean are unusual, unless the sample size n is very large or the distribution is not normal. A formal test may be made of the following hypotheses:

Null hypothesis: The most extreme value comes from the same normal distribution as the other observations.

Alternative hypothesis: The most extreme value does not come from the same normal distribution as the other observations.

A widely used test of these hypotheses is Grubbs' test, also called the *Extreme Studentized Deviate* test. STATGRAPHICS Centurion XVI conducts this test and displays a *P-value*. In general, a *P*-value quantifies the probability of obtaining a statistic as unusual or more unusual than that observed in the sample, if the null hypothesis were true. If the *P*-value is small enough, the null hypothesis can be rejected, since the sample would have been an extremely rare event. "Small enough" is usually defined as less than 0.05, which is called the "significance level" or "alpha risk" of the test procedure. If there is less than a 5% chance that the sample would have arisen given that the null hypothesis was true, then the null hypothesis is rejected.

In this example, the test statistic equals the largest absolute *Studentized Value Without Deletion*, 3.479. It has a *P*-value equal to 0.0484. Since the *P*-value is less than 0.05, we would reject the null hypothesis, thereby concluding that row #15 is an outlier compared to the rest of the data sample.

You can remove row #15 by pressing the *Input Dialog* button on the analysis toolbar and entering an expression in the *Select* field such as that shown below:

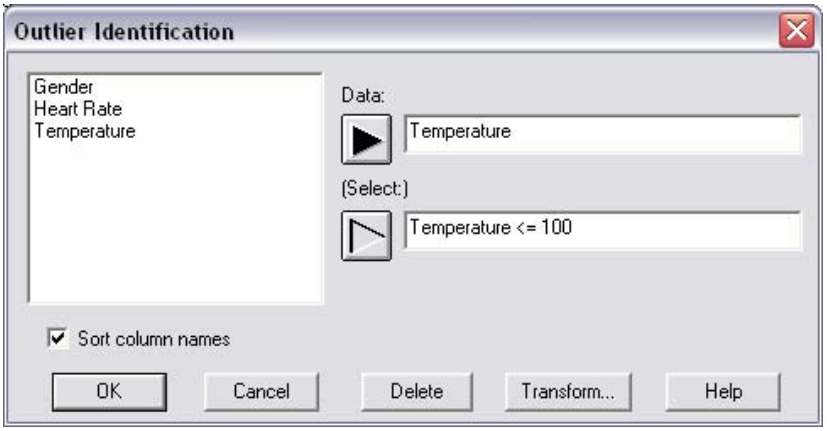


Figure 10-11. Outlier Identification Dialog Box with Entry Removing Outlier

Since row #15 is the only observation that exceeds 100 degrees, the *Select* field entry above will select only the other $n = 129$ rows. The modified *Outlier Identification* output is shown below:

Sorted Values

		Studentized Values	*Studentized Values*	*Modified*
Row	*Value*	*Without Deletion*	*With Deletion*	*MAD Z-Score*
95	96.3	-2.75487	-2.85205	-2.698
55	96.4	-2.61209	-2.6956	-2.5631
23	96.7	-2.18375	-2.23455	-2.1584
30	96.7	-2.18375	-2.23455	-2.1584
73	96.8	-2.04097	-2.08332	-2.0235
...				
119	99.4	1.6713	1.69652	1.4839
99	99.4	1.6713	1.69652	1.4839
13	99.5	1.81408	1.84516	1.6188
97	99.9	2.3852	2.44992	2.1584
120	100.0	2.52798	2.60411	2.2933

Grubbs' Test (assumes normality)
Test statistic = 2.75487
P-Value = 0.676064

Figure 10-12. Outlier Identification Output after Removing Row #15

The most extreme value among the remaining observations is row #95. Since the *P*-value for Grubbs' test is well above 0.05, all of the remaining observations appear to have come from the same population.

Ideally, one would go back to the original study and attempt to find an assignable cause for the abnormal value for individual #15. Since that is impossible to do now, we will accept the results of Grubbs' test and remove row #15 from all subsequent calculations. Modifying the data input dialog box for the *One-Variable Analysis* in the same manner as in *Figure 10-11*, the modified summary statistics are shown below:

Summary Statistics for Temperature

Count	129
Average	98.2295
Median	98.3
Standard deviation	0.70038
Coeff. of variation	0.713004%
Minimum	96.3
Maximum	100.0
Range	3.7
Lower quartile	97.8
Upper quartile	98.7
Interquartile range	0.9
Stnd. skewness	-1.40217
Stnd. kurtosis	0.257075

Figure 10-13. Summary Statistics after Removing Row #15

10.5 Histogram

Another common graphical display that illustrates a sample of measurement data is the frequency histogram. Returning to the *One-Variable Analysis* procedure, a histogram may be created by pressing the *Tables and Graphs* button  on the analysis toolbar and selecting *Frequency Histogram*. The default histogram is shown below:

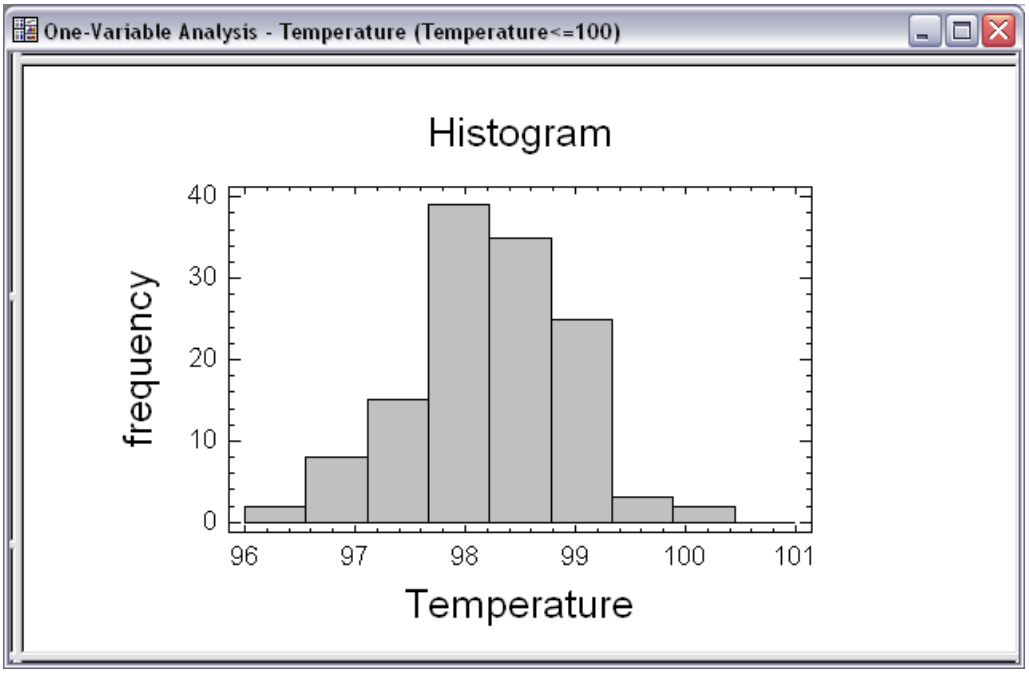


Figure 10-14. Frequency Histogram with Default Classes

The height of each bar in the histogram represents the number of observations that fall in the interval of *temperature* covered by the bar. The number of bars and their range is set by default based on the sample size *n*, using whatever rule has been selected on the *EDA* (Exploratory Data Analysis) tab of the *Edit - Preferences* dialog box:

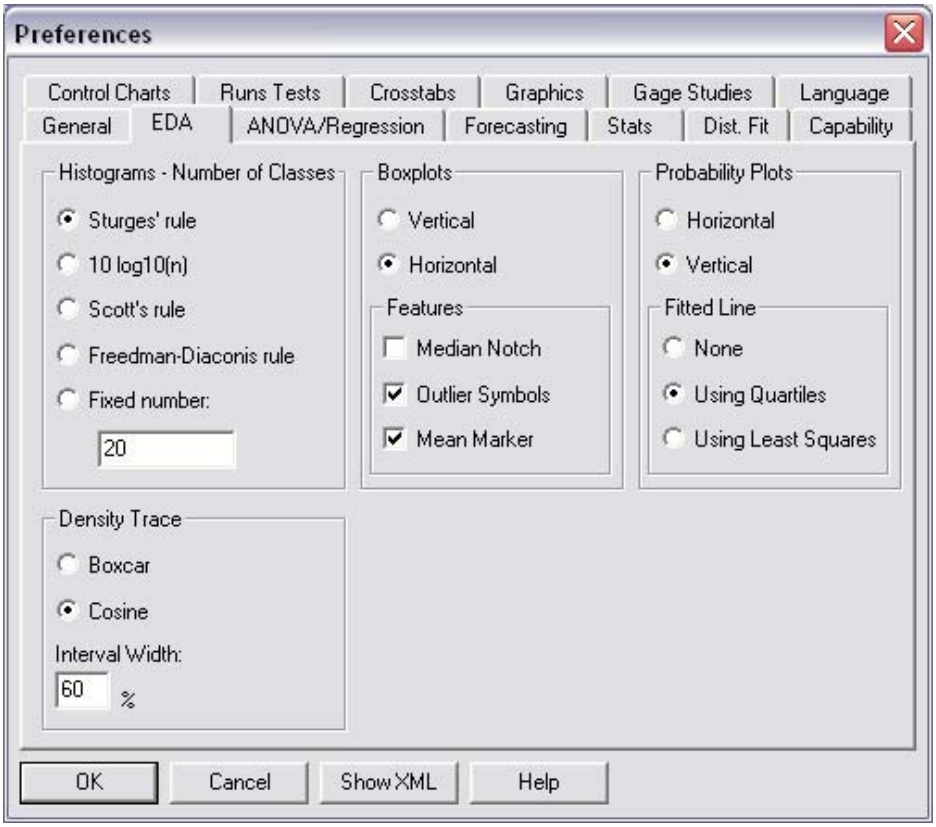


Figure 10-15. EDA Tab of the Preferences Dialog Box

Using Sturges' rule, the number of bars is set to the smallest integer that is not less than $(1+3.322\log_{10}(n))$. Other rules, such as the $10\log_{10}(n)$ rule, tend to produce more bars by default and may be preferable if you tend to work with large data sets.

A temporary override for a histogram once it has been created is available by double-clicking on the histogram to maximize its pane and then selecting *Pane Options*:

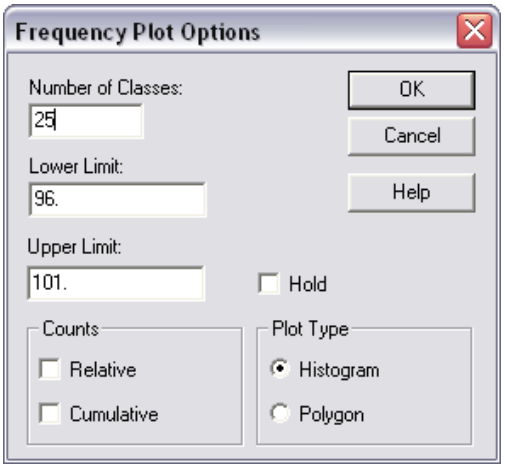


Figure 10-16. Pane Options Dialog Box for Frequency Histogram

In setting the classes, the number of significant digits in the data should be considered. For example, body temperatures were measured only to the nearest 0.1 of a degree. The width of the intervals covered by the bars should thus be an integer multiple of 0.1. That way, each bar will cover the same number of possible measurements. The plot below shows 25 intervals between 96 and 101 degrees, each covering an interval of 0.2 degrees:

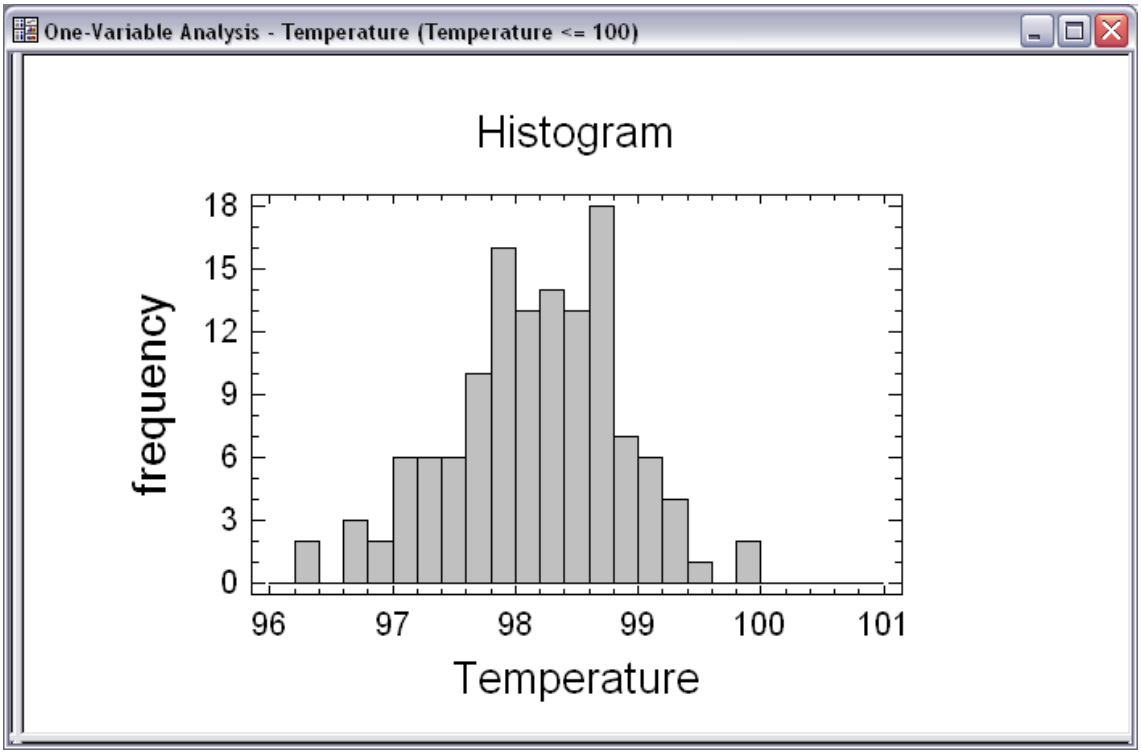


Figure 10-17. Frequency Histogram with Redefined Classes

With the greater number of classes, more detail is apparent. The general shape of the distribution is similar to that of a bell-shaped normal curve.

The data displayed in the histogram may be shown in tabular form by pressing the *Tables and Graphs* button on the analysis toolbar and selecting *Frequency Tabulation*:

Frequency Tabulation for Temperature

	Lower	*Upper*			*Relative*	*Cumulative*	*Cum. Rel.*
Class	*Limit*	*Limit*	*Midpoint*	*Frequency*	*Frequency*	*Frequency*	*Frequency*
	at or below	96.0		0	0.0000	0	0.0000
1	96.0	96.2	96.1	0	0.0000	0	0.0000
2	96.2	96.4	96.3	2	0.0155	2	0.0155
3	96.4	96.6	96.5	0	0.0000	2	0.0155
4	96.6	96.8	96.7	3	0.0233	5	0.0388
5	96.8	97.0	96.9	2	0.0155	7	0.0543
6	97.0	97.2	97.1	6	0.0465	13	0.1008
7	97.2	97.4	97.3	6	0.0465	19	0.1473
8	97.4	97.6	97.5	6	0.0465	25	0.1938
9	97.6	97.8	97.7	10	0.0775	35	0.2713
10	97.8	98.0	97.9	16	0.1240	51	0.3953
11	98.0	98.2	98.1	13	0.1008	64	0.4961
12	98.2	98.4	98.3	14	0.1085	78	0.6047
13	98.4	98.6	98.5	13	0.1008	91	0.7054
14	98.6	98.8	98.7	18	0.1395	109	0.8450
15	98.8	99.0	98.9	7	0.0543	116	0.8992
16	99.0	99.2	99.1	6	0.0465	122	0.9457
17	99.2	99.4	99.3	4	0.0310	126	0.9767
18	99.4	99.6	99.5	1	0.0078	127	0.9845
19	99.6	99.8	99.7	0	0.0000	127	0.9845
20	99.8	100.0	99.9	2	0.0155	129	1.0000
21	100.0	100.2	100.1	0	0.0000	129	1.0000
22	100.2	100.4	100.3	0	0.0000	129	1.0000
23	100.4	100.6	100.5	0	0.0000	129	1.0000
24	100.6	100.8	100.7	0	0.0000	129	1.0000
25	100.8	101.0	100.9	0	0.0000	129	1.0000
	above	101.0		0	0.0000	129	1.0000

Mean = 98.2295 Standard deviation = 0.70038

Figure 10-18. Frequency Tabulation Table

Note that observations are counted as falling within an interval if they are greater than the lower limit of the interval and less than or equal to the upper limit.

The column on the far right is also of considerable interest, since it shows the cumulative probability that an individual will fall in a selected class or earlier classes. For example, 89.92% of all data values are equal to or less than 99.0 degrees.

10.6 Quantile Plot and Percentiles

Another way to display cumulative probabilities is by selecting *Quantile Plot* from the list of *Graphs* in the *One-Variable Analysis* procedure:

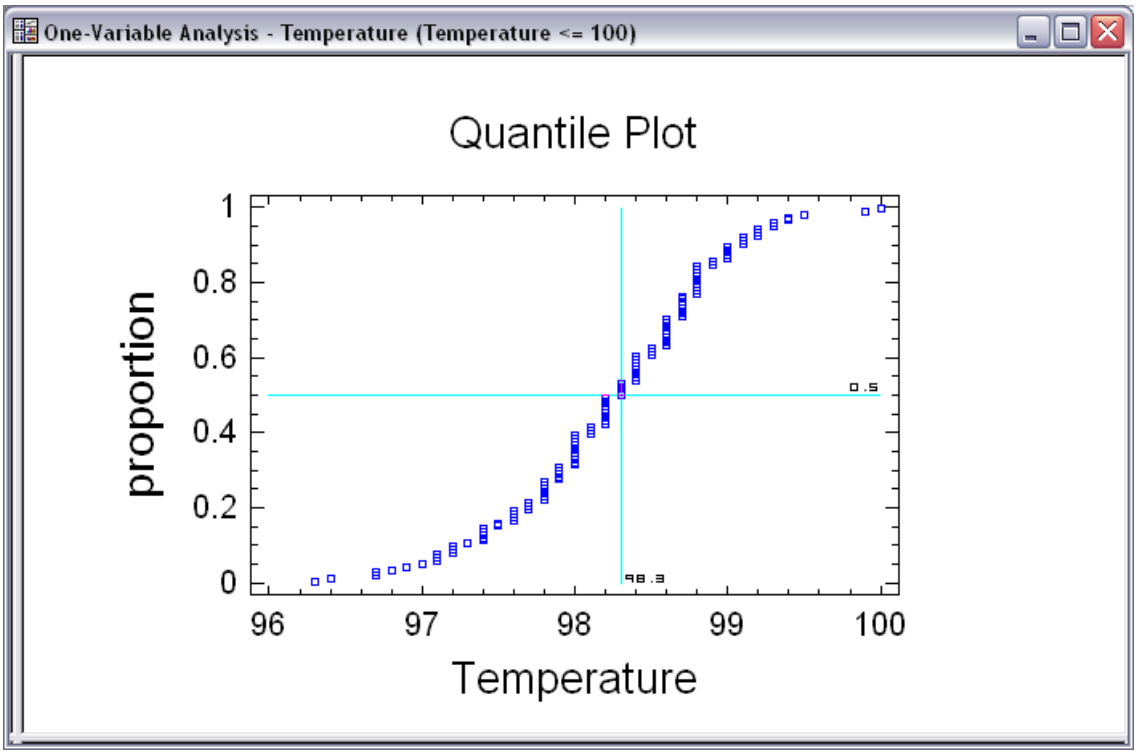


Figure 10-19. Quantile Plot

In this plot, the data are first sorted from smallest to largest. The j^{th} largest data value is then plotted at Y = *(j+0.5)/n*. This estimates the proportion of the population at or below the observed temperature. Like the rightmost column in the frequency table, the curve represents the cumulative probability of an individual having a temperature less than or equal to that shown on the horizontal axis. Since the temperature data were only measured to the nearest 0.1 degrees, there are vertical jumps in the above display.

Figure 10-19 also shows a set of crosshair cursors. These are created by pressing the alternate mouse button while viewing the graph and selecting *Locate* from the popup menu. You can then use your mouse to drag the crosshairs to any location. The small numbers near the crosshairs indicate their position. In the above plot, the crosshairs have been used to locate the median or

50th *percentile*, which is the value of *temperature* at which the proportion displayed on the vertical axis equals 0.5.

A table of percentiles may also be created by selecting *Percentiles* from the *Tables* list:

Percentiles for Temperature

	Percentiles
1.0%	96.4
5.0%	97.0
10.0%	97.2
25.0%	97.8
50.0%	98.3
75.0%	98.7
90.0%	99.1
95.0%	99.3
99.0%	99.9

Output includes 95.0% normal confidence limits.

Figure 10-20. Percentiles Table

The p^{th} percentile estimates the value of temperature below which p% of the population lies. *Pane Options* has been used to add 95% confidence limits to those percentiles, based on the assumption that the sample comes from a normal distribution.

For example, the 90th percentile is the value of temperature exceeded by only 10% of the individuals in the population. The best estimate of that percentile based on the sample of data is 99.1 degrees. However, given the limited size of the sample, the 90th percentile could lie anywhere between 98.98 and 99.31 degrees, with 95% confidence.

10.7 Confidence Intervals

Having removed the outlier from the sample, we can proceed to establish final estimates for the parameters of the distribution from which the data came. Selecting *Confidence Intervals* from the *Tables and Graphs* dialog box displays:

Confidence Intervals for Temperature
95.0% confidence interval for mean: 98.2295 +/- 0.122015 [98.1074,98.3515]
95.0% confidence interval for standard deviation: [0.624081,0.798114]

Figure 10-21. 95% Confidence Intervals for the Mean and Standard Deviation

The confidence intervals provide a bound on the potential error in estimating the mean and standard deviation of the population. Given the remaining $n = 129$ observations, we can declare

with 95% confidence that the mean population temperature is somewhere between 98.11 degrees and 98.35 degrees. Likewise, the standard deviation of the population is somewhere between 0.624 degrees and 0.798 degrees.

Selecting *Pane Options*, additional confidence intervals can be requested using the bootstrap method:

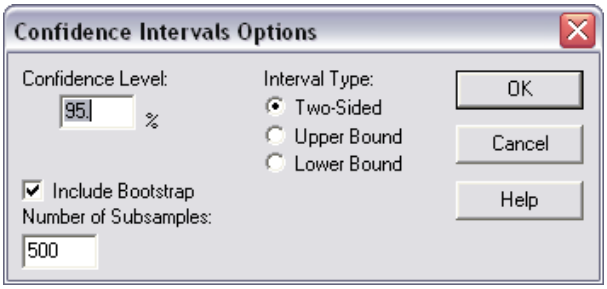


Figure 10-22. Confidence Intervals Options Dialog Box

Bootstrap intervals, unlike the intervals in *Figure 10-21*, do not rely on the assumption that the population follows a normal distribution. Instead, random samples of $n = 129$ observations are taken from the data, sampling with replication (the same observations may be selected more than once). This is repeated 500 times, sample statistics are calculated, and the most central 95% of the results are used to calculate the confidence intervals. The table below shows bootstrap intervals for the population mean, standard deviation, and median:

Confidence Intervals for Temperature
95.0% confidence interval for mean: 98.2295 +/- 0.122015 [98.1074,98.3515]
95.0% confidence interval for standard deviation: [0.624081,0.798114]

Bootstrap Intervals
Mean: [98.1132,98.3519]
Standard deviation: [0.621373,0.785949]
Median: [98.1,98.4]

Figure 10-23. Bootstrap 95% Confidence Intervals

NOTE: Your results might vary slightly from the results above.

The earlier intervals, calculated using Student's t distribution and the chi-square distribution, are closely matched by the bootstrap intervals. This is not unexpected, since the data do not show significant skewness or kurtosis.

10.8 Hypothesis Tests

Formal hypothesis tests may also be performed. For example, it is often asserted that normal human temperature is 98.6 degrees Fahrenheit. To test whether or not the current data come from a normal distribution with such a mean, a hypothesis test may be created to test between:

Null hypothesis: $\mu = 98.6$ degrees

Alternative hypothesis: $\mu \neq 98.6$ degrees

To run the test within the *One-Variable Analysis* procedure, select *Hypothesis Tests* from the list of *Tables and Graphs*. Before examining the results, select *Pane Options* and specify the attributes of the desired test:

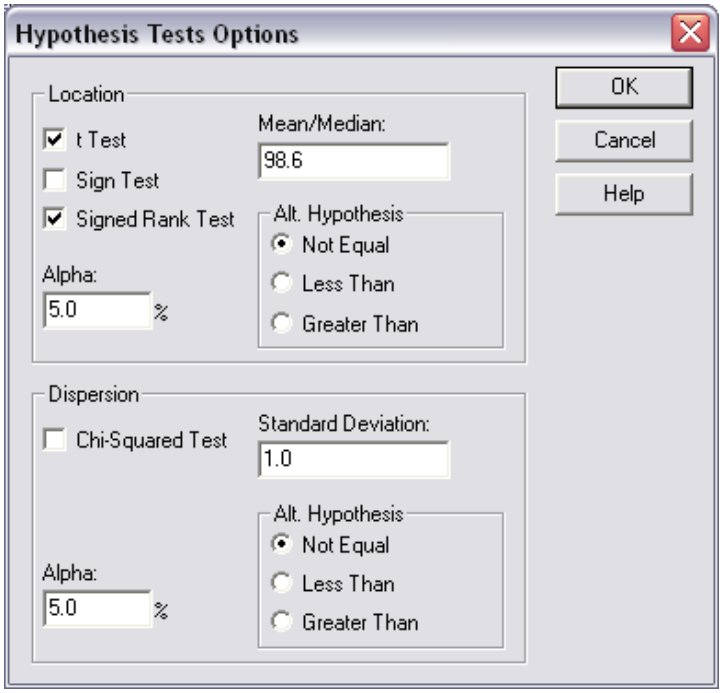


Figure 10-24. Pane Options for Hypothesis Tests

The value entered for *Mean* represents the null hypothesis. Under *Alt. Hypothesis*, you may select any of three alternative hypotheses:

1. *Not equal*: $\mu \neq 98.6$
2. *Less than*: $\mu < 98.6$
3. *Greater than*: $\mu > 98.6$

Even though the sample suggests a lower mean temperature, a two-sided alternative has been selected. Creating a one-sided test with an alternative hypothesis of $\mu < 98.6$ degrees would be considered "data snooping" at this point, since we would be formulating the hypothesis after having looked at the data.

The results of the test are shown below:

Hypothesis Tests for Temperature
Sample mean = 98.2295
Sample median = 98.3
Sample standard deviation = 0.70038

t-test
Null hypothesis: mean = 98.6
Alternative: not equal

Computed t statistic = -6.00896
P-Value = 1.81264E-8
Reject the null hypothesis for alpha = 0.05.

signed rank test
Null hypothesis: median = 98.6
Alternative: not equal

Average rank of values below hypothesized median: 67.7099
Average rank of values above hypothesized median: 43.5658

Large sample test statistic = 5.07771 (continuity correction applied)
P-Value = 3.82663E-7
Reject the null hypothesis for alpha = 0.05.

Figure 10-25. Hypothesis Tests Results

The results of two tests are shown:

1. A standard *t*-test, which assumes that the data come from a normal distribution (although it is not overly sensitive to departures from this assumption).

2. A nonparametric signed rank test, based on the ranks of the distance of each observation from the hypothesized median. This test does not assume normality and is less sensitive to outliers than the *t*-test.

In both cases, the *P*-value is way below 0.05, soundly rejecting the hypothesis that the sample comes from a population with a mean of 98.6 degrees.

> NOTE: the notation E-8 after a number means that the number is to be multiplied by 10^{-8}. The *P*-value shown as 1.81264E-8 therefore equals 0.0000000181264.

It should be noted that the confidence interval for the mean, given in Section 10.8, did not include the value 98.6. Any values not within the confidence interval would have been rejected by the *t*-test considered here. You can thus think of the confidence interval as containing all possible values for the population that are supportable by the data sample.

10.9 Tolerance Limits

One additional analysis is useful for the body temperature data. It creates normal tolerance limits, which are limits within which a selected percentage of the population is estimated to fall with a given confidence level. Tolerance limits are available on the main menu by selecting:

1. If using the Classic menu, select *Describe –Numeric Data – Statistical Tolerance Limits*
2. If using the Six Sigma menu, select *Analyze – Variable Data – Statistical Tolerance Limits*

The procedure begins by displaying a dialog box in which you enter the sample size *n* and the sample mean and standard deviation. Using the results in *Figure 10-13*, the proper entries are:

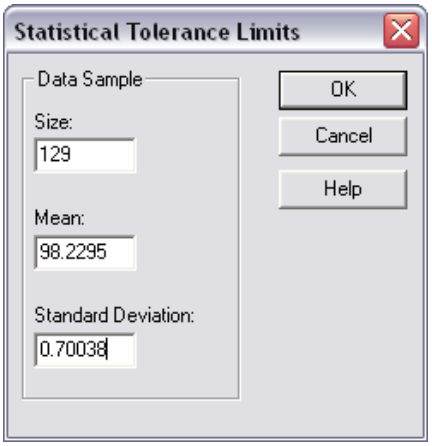


Figure 10-26. Dialog Box for Statistical Tolerance Limits

When you hit OK, the *Options* menu appears, and then the *Tables and Graphs* dialog box appears. The resulting output is shown below:

Statistical Tolerance Limits
Sample size = 129
Sample mean = 98.2295
Sample standard deviation = 0.70038

95.0% tolerance interval for 99.0% of the population
Xbar +/- 2.88436 sigma
Upper: 100.25
Lower: 96.2093

The StatAdvisor
Assuming that the data comes from a normal distribution, the tolerance limits state that we can be 95.0% confident that 99.0% of the distribution lies between 96.2093 and 100.25. This interval is computed by taking the mean of the data +/-2.88436 times the standard deviation.

Figure 10-27. Analysis Summary for Statistical Tolerance Limits

The interpretation of the StatAdvisor summarizes the results succinctly. The confidence level and percentage of the population that is bound may be changed using *Pane Options*.

Also created by the *Statistical Tolerance Limits* procedure is a *Tolerance Plot*, which displays the tolerance limits:

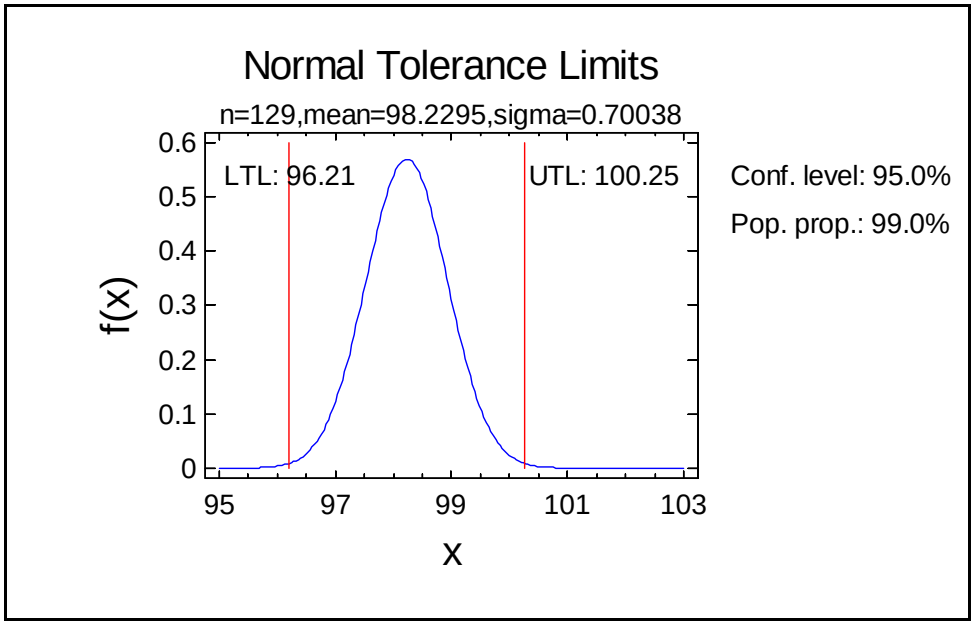


Figure 10-28. Tolerance Plot

No more than one individual out of every 100 is likely to lie outside the calculated limits.

Chapter

11

Tutorial #2: Comparing Two Samples

Graphical comparisons and hypothesis tests.

Often, data to be analyzed consists of two samples, possibly from different populations. In such cases, it is useful to:

1. Display the data in such a way that visual comparisons are possible.

2. Test hypotheses to determine whether or not there are statistically significant differences between the samples.

Tutorial #1 in the last chapter analyzed a set of body temperatures taken from 130 subjects. Of those subjects 65 were female and 65 were male. In this tutorial, we will compare the data of the women to those of the men.

To analyze the body temperatures, open the *bodytemp.sgd* data file using *Open Data Source* on the *File – Open* menu.

11.1 Running the Two Sample Comparison Procedure

The main procedure for comparing data from two samples is the *Two-Sample Comparison* procedure, accessed from the main menu as follows:

1. If using the Classic menu, select *Compare – Two Samples – Independent Samples*.

2. If using the Six Sigma menu, select *Analyze – Variable Data – Two Sample Comparisons – Independent Samples*.

The data input dialog box for that procedure is shown below:

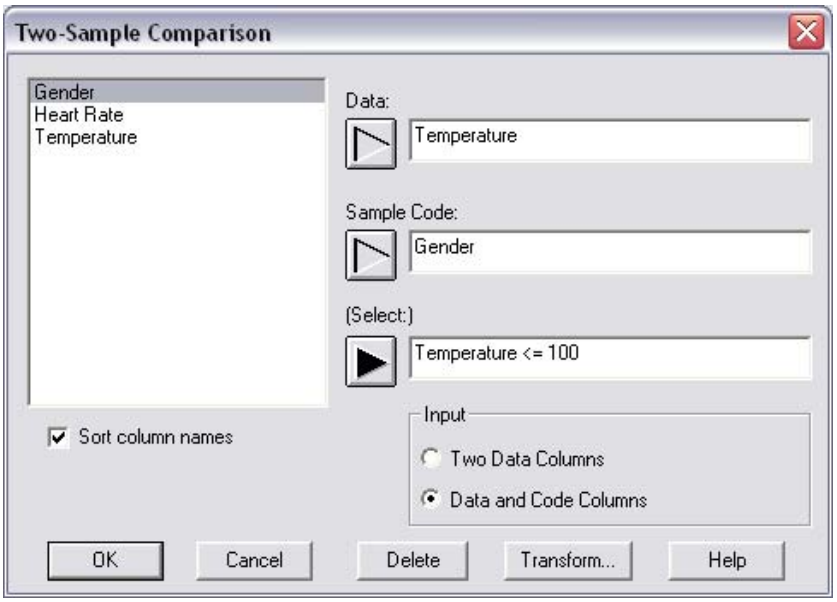


Figure 11-1. Two Sample Comparison Dialog Box

The *Input* box indicates how the data for the two samples have been entered:

1. *Two Data Columns* – the data for each sample is in a different column.

2. *Data and Code Columns* – the data for both samples is in the same column, and a second column contains codes that differentiate between the two samples.

The *bodytemp.sgd* file has the second type of structure, with all $n = 130$ observations in one column named *Temperature*, while a second column named *Gender* contains the label "Female" or "Male". In the *Select* field, an entry has been made to select only rows for which *Temperature* is less than or equal to 100. This will exclude row #15 from the analysis, which was determined in Chapter 10 to be an outlier.

After the *Tables and Graphs* box, the initial analysis window contains four panes, with a summary of the data, a dual histogram, summary statistics by group, and a dual box-and-whisker plot:

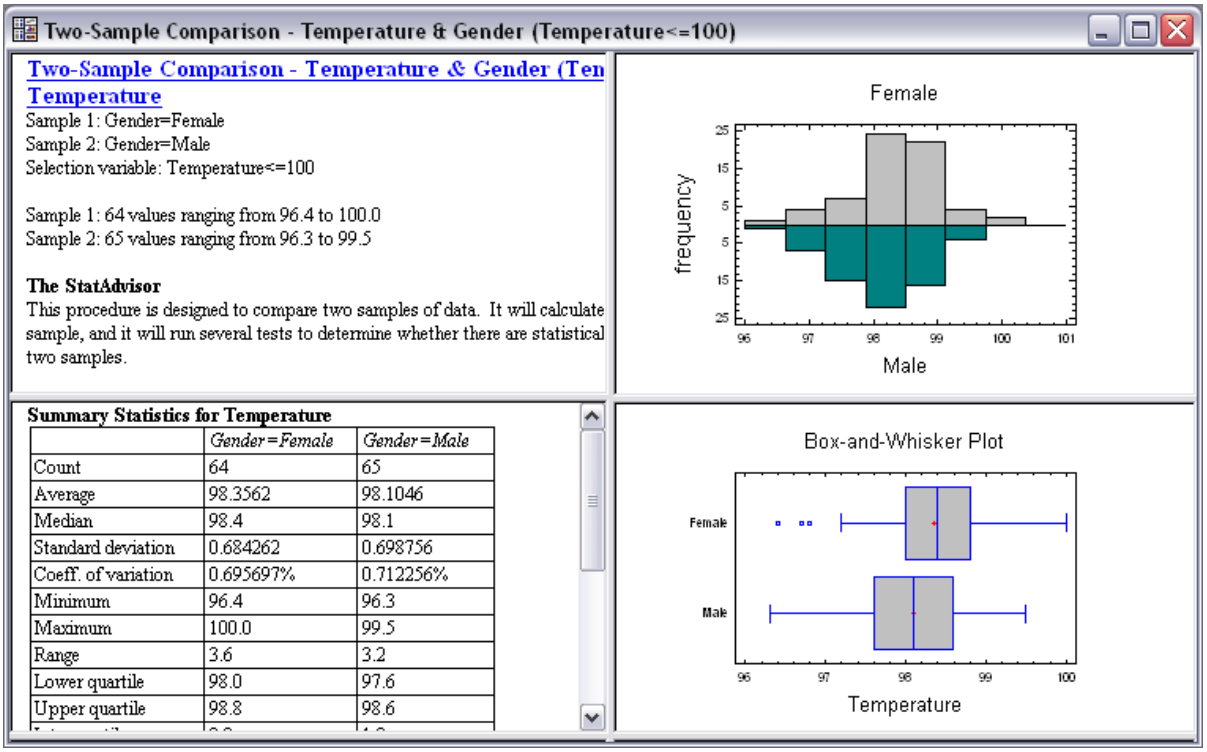


Figure 11-2. Two Sample Comparison Analysis Window

After removing the outlier, there are $n_1 = 64$ observations for females, ranging from 96.4 to 100.0 degrees, and $n_2 = 65$ observations for males, ranging from 96.3 degrees to 99.5 degrees.

11.2 Summary Statistics

The *Summary Statistics* table shows statistics calculated for each sample:

Summary Statistics for Temperature

	Gender=Female	*Gender=Male*
Count	64	65
Average	98.3562	98.1046
Median	98.4	98.1
Standard deviation	0.684262	0.698756
Coeff. of variation	0.695697%	0.712256%
Minimum	96.4	96.3
Maximum	100.0	99.5
Range	3.6	3.2
Lower quartile	98.0	97.6
Upper quartile	98.8	98.6
Interquartile range	0.8	1.0
Stnd. skewness	-1.35246	-0.702297
Stnd. kurtosis	1.49635	-0.610877

Figure 11-3. Summary Statistics by Sample

Several facts are of particular interest:

1. The average temperature of the females is about 0.25 degrees higher than that of the males. The difference between the medians is 0.30 degrees.

2. The standard deviation of the females is slightly less than that of the males, indicating that the body temperatures of the females may be less variable than those of the males.

3. Both samples have standardized skewness and standardized kurtosis values within the range of -2 to 2. As explained in Chapter 10, values within that range are consistent with the hypothesis that the data come from normal distributions.

Whether or not the apparent difference between the temperatures of females and males is statistically significant remains to be determined.

11.3 Dual Histogram

The frequency histogram provides a back-to-back comparison of the two samples. Using *Pane Options* to rescale the class intervals so that there are 25 intervals between 96 and 101 degrees generates the following plot:

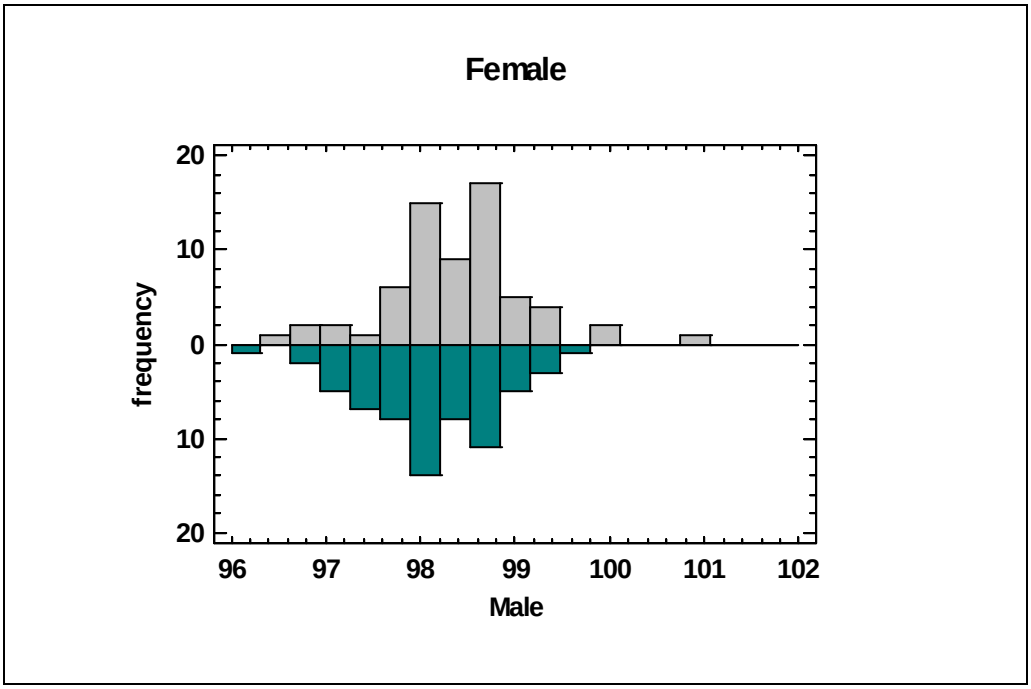


Figure 11-4. Dual Frequency Histogram

The histogram for the females is displayed above the horizontal line. The histogram for the males is inverted and displayed below the line. The shapes of the distributions are similar, with a possible shift of the females' distribution to the right of the males.

11.4 Dual Box-and-Whisker Plot

The analysis window also displays box-and-whisker plots for the two samples. As explained in Chapter 10, the central boxes cover the middle half of each sample. The whiskers extend to the largest and smallest data values in each sample, except for any points that are unusually far from the boxes. A vertical line is drawn within each box at the sample median, while small plus signs indicate the locations of the sample means.

In this case, it is particularly useful to add median notches by accessing *Pane Options*. The resulting plot is shown below:

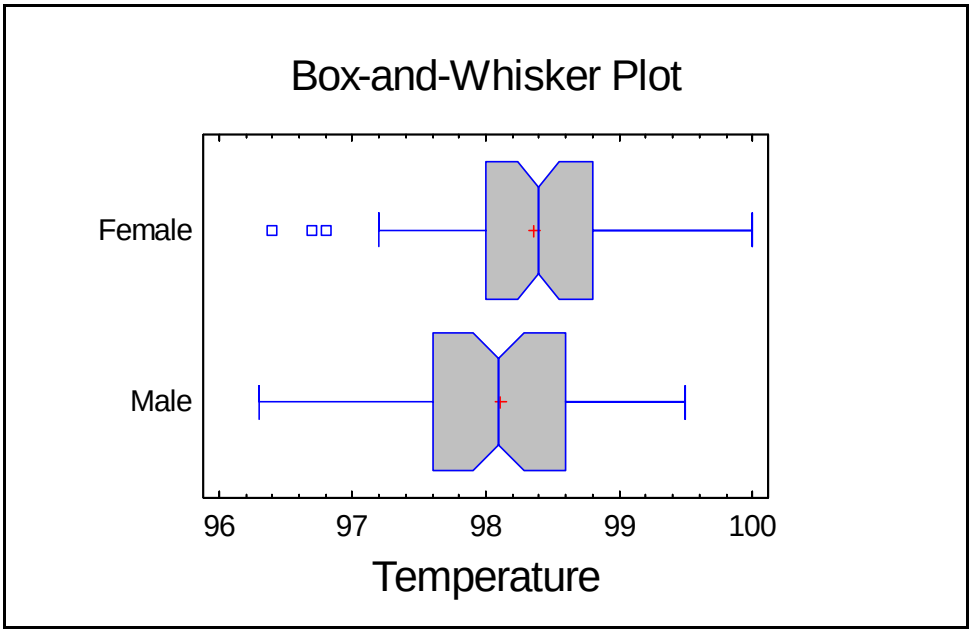


Figure 11-5. Dual Box-and-Whisker Plot with Median Notches

Evident in this plot are:

1. An apparent offset of the center of the females' distribution to the right of the males' distribution. Both the sample means and medians show a similar difference.

2. The range covered by the females is wider than the range covered by the males, but only if you include the lowest outside point.

3. The median notch for the females overlaps that of the males slightly. The notches are drawn in such a way that if the two notches did not overlap, one could declare the two medians to be significantly different at the default system significance level (which is currently 5%). A more formal comparison is described in a later section.

Based upon this plot, there appears to be a difference in the center of the two samples, though the statistical significance of that difference remains undetermined.

11.5 Comparing Standard Deviations

The first formal comparison between the two samples is to test the hypothesis that the standard deviations (σ) of the populations from which the data came are equal versus the hypothesis that they are different:

Null hypothesis: $\sigma_1 = \sigma_2$

Alternative hypothesis: $\sigma_1 \neq \sigma_2$

This will allow us to determine whether the apparent difference between the variability of the males and females is statistically significant, or whether it is within the range of normal random variability for samples of the current size.

To perform the test, press the *Tables and Graphs* button  on the analysis toolbar and select *Comparison of Standard Deviations*. The result is shown below:

Comparison of Standard Deviations for Temperature

	Gender=Female	*Gender=Male*
Standard deviation	0.684262	0.698756
Variance	0.468214	0.48826
Df	63	64

Ratio of Variances = 0.958945

95.0% Confidence Intervals
Standard deviation of Gender=Female: [0.582853,0.828723]
Standard deviation of Gender=Male: [0.595887,0.844885]
Ratio of Variances: [0.584028,1.57609]

F-test to Compare Standard Deviations
Null hypothesis: sigma1 = sigma2
Alt. hypothesis: sigma1 NE sigma2
F = 0.958945 P-value = 0.8684
Do not reject the null hypothesis for alpha = 0.05.

Figure 11-6. Two-Sample Comparison of Standard Deviations

The most important output in this table is highlighted in red:

1. *Ratio of Variances*: displays a 95% confidence interval for the ratio of the variance of the population of females, σ_1^2, divided by the variance of the population of males, σ_2^2. *Variance* is a measure of variability calculated by squaring the standard deviation. (NOTE: comparisons of variability amongst more than one sample are usually based on variances rather than standard deviations, since they have more attractive mathematical

properties.) The interval for σ_1^2 / σ_2^2 ranges from 0.58 to 1.58. This indicates that the variance of the females may well be anywhere between approximately 58% of the variance of the males to 158% of their variance. This lack of precision is typical when trying to compare the variability of relatively small samples.

2. The *P-value* associated with the *F* statistic of the hypotheses stated above. A *P*-value less than 0.05 would indicate a statistically significant difference between the variance of the females and the variance of the males at the 5% significance level. Since P is well above 0.05, there is no evidence upon which to reject the hypothesis of equal variances (and thus equal standard deviations).

Therefore there is no statistically significant evidence upon which to conclude that the variability of the female body temperatures is different that the variability of male body temperatures.

It should be noted that this test is quite sensitive to the assumption that the samples come from normally distributed populations, an assumption that was shown to be reasonable based on the standardized skewness and standardized kurtosis values.

11.6 Comparing Means

The second comparison between the two samples tests the hypothesis that the means (μ) of the two populations are equal:

Null hypothesis: $\mu_1 = \mu_2$

Alternative hypothesis: $\mu_1 \neq \mu_2$

To perform this test, press the *Tables* button again and select *Comparison of Means*. The results are:

Comparison of Means for Temperature
95.0% confidence interval for mean of Gender=Female: 98.3562 +/- 0.170924 [98.1853,98.5272]
95.0% confidence interval for mean of Gender=Male: 98.1046 +/- 0.173144 [97.9315,98.2778]
95.0% confidence interval for the difference between the means
assuming equal variances: 0.251635 +/- 0.240998 [0.0106371,0.492632]

t test to compare means
Null hypothesis: mean1 = mean2
Alt. hypothesis: mean1 NE mean2
assuming equal variances: t = 2.06616 P-value = 0.040846
Reject the null hypothesis for alpha = 0.05.

Figure 11-7. Two-Sample Comparison of Means

The most important output in this table is again highlighted in red:

1. *Difference between the Means (assuming equal variances)*: displays a 95% confidence interval for the mean of the population of females minus the mean of the population of males. The interval for $\mu_1 - \mu_2$ ranges from 0.01 to 0.49, indicating that the mean female body temperature is somewhere between 0.01 degrees and 0.49 degrees higher than the mean body temperature of the males.

2. The *P-value* associated with a t test of the hypotheses stated above. Since the *P*-value is less than 0.05, there is sufficient evidence upon which to reject the hypothesis of equal means and thus declare the two population means to be statistically different at the 5% significance level.

Note that this test was made assuming that the variances of the two populations are equal, which was validated by the *F* statistic in the previous section. Had the variances been shown to be significantly different, an approximate t test could have been requested by accessing *Pane Options* and removing the checkmark from the checkbox labeled *Assume Equal Sigmas*.

It thus appears that the females come from a population with a higher mean temperature than that of the males.

11.7 Comparing Medians

If it is suspected that the data may contain outliers, a nonparametric test can be performed to compare the medians rather than the means. Nonparametric tests do not assume that data come from normal distributions and tend to be less affected by outliers if any are present.

Selecting *Comparison of Medians* from the *Tables and Graphs* dialog box generates a Mann-Whitney (Wilcoxon) *W* statistic. In this test, the two samples are first combined. The combined data are then ranked from 1 to n_1+n_2, and the original data values are replaced by their respective ranks. A test statistic *W* is then constructed comparing the average ranks of the observations in the two samples:

Comparison of Medians for Temperature
Median of sample 1: 98.4
Median of sample 2: 98.1

Mann-Whitney (Wilcoxon) W-test to compare medians
Null hypothesis: median1 = median2
Alt. hypothesis: median1 NE median2

Average rank of sample 1: 71.9219
Average rank of sample 2: 58.1846

W = 1637.0 P-value = 0.0368312
Reject the null hypothesis for alpha = 0.05.

Figure 11-8. Two-Sample Comparison of Medians

Interpretation of the Mann-Whitney (Wilcoxon) test parallels that of the *t*-test described in the last section, with a small *P*-value leading to the conclusion that the medians of the two populations are significantly different.

11.8 Quantile Plot

To illustrate the difference between the two distributions, side-by-side quantile plots of each sample can be displayed by selecting *Quantile Plot* from the *Graphs* dialog box:

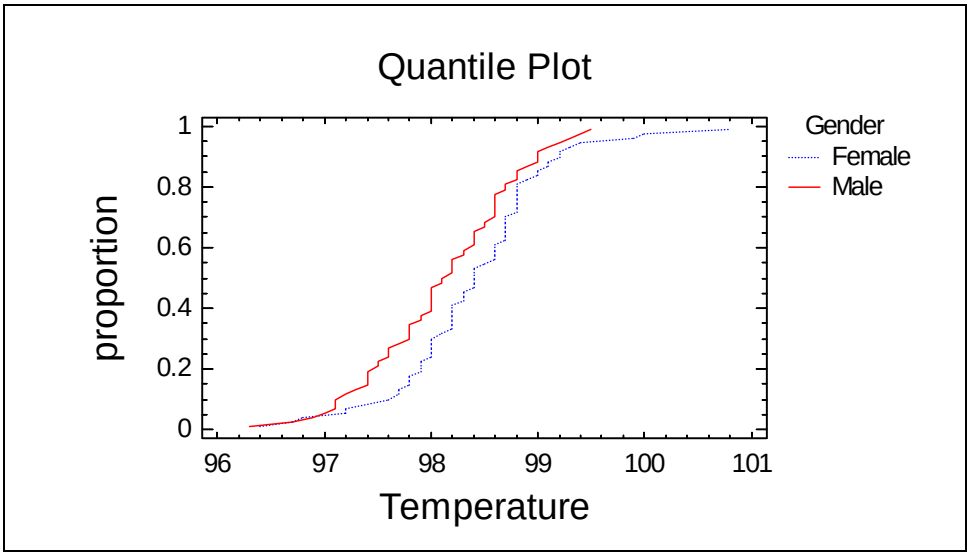


Figure 11-9. Side-by-Side Quantile Plots

The quantile plot illustrates the proportion of data in each sample that is below a given value of X, as a function of X. If the samples come from the same population, the quantile plots should be close together. Any offset of one plot to the right or left of the other indicates a difference between the two sample means. A difference in the slope of the curves indicates a difference between the standard deviations.

In the above plot, it is quite evident that the distribution of the females is shifted to the right of the males. The overall slopes, however, are similar.

11.9 Two-Sample Kolmogorov-Smirnov Test

One additional nonparametric test that may be performed if the assumption of normal distributions is not tenable is the two-sample Kolmogorov-Smirnov test. This test is based on calculating the maximum vertical distance between the cumulative distribution functions of the two samples, which is approximately the maximum distance between the two quantile plots in Figure 11-9. If the maximum distance is large enough, the two samples may be declared to come from significantly different populations.

Selecting *Kolmogorov-Smirnov Test* from the *Tables and Graphs* dialog box displays the following:

Kolmogorov-Smirnov Test for Temperature
Estimated overall statistic DN = 0.242548
Two-sided large sample K-S statistic = 1.37737
Approximate P value = 0.0449985

Figure 11-10. Output from Kolmogorov-Smirnov Test

The maximum vertical distance, denoted by DN, equals approximately 0.24 for the body temperature data.

The *P*-value is used to determine whether or not the distributions are significantly different from each other. A small *P*-value leads to the conclusion that there *is* a significant difference. Since the *P*-value for the sample data is less than 0.05, there is a significant difference between the male and female distributions at the 5% significance level.

> Warning: If data is heavily rounded, this test may not be reliable since the empirical Cumulative Distribution Function (CDF) may jump in large steps. When possible, it is best to rely on a comparison of selected distribution parameters such as the mean, standard deviation, or median.

11.10 Quantile-Quantile Plot

A final plot, available by selecting *Quantile-Quantile Plot* from the *Graphs* dialog box, plots the estimated quantiles of one sample versus the quantiles of the other sample:

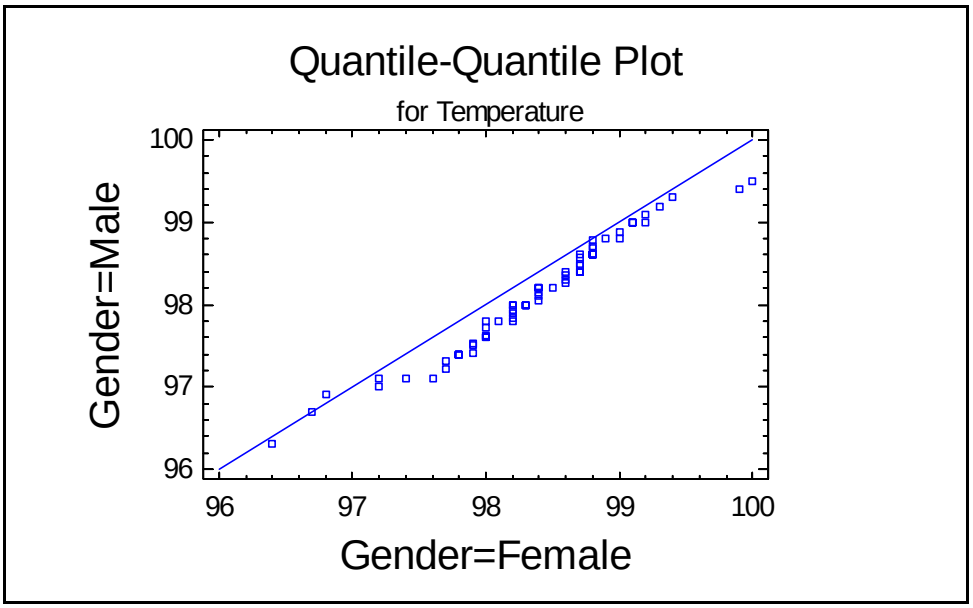


Figure 11-11. Q-Q Plot of Body Temperature Data

There is a point on this graph corresponding to each observation in the smaller of the two samples. Plotted on the other axis is the estimated quantile of the larger sample. If the samples come from identical populations, the points should lie close to the diagonal line. A constant shift left or right indicates that there is a significant difference between the centers of the two distributions. Points diverging from the line at a slope different than that of the diagonal line indicate a significant difference in variability. In this case, the difference between the populations may be a little more complicated than a simple shift in the mean, since the points are closer to the line at high and low temperatures than they are at central temperatures. It appears that the distribution of temperatures for the females is more concentrated in the center than the distribution for the males.

Chapter

12

Tutorial #3: Comparing More than Two Samples

Comparing means and standard deviations, one-way ANOVA, ANOM, and graphical methods.

When data fall into more than two groups, a different set of techniques need to be employed than in the previous chapter. For example, suppose you wished to compare the strength of widgets made from 4 different materials. In a typical experiment, you might make 12 widgets from each of the four materials in order to compare them. The following data represent the results of such an experiment:

Material A	*Material B*	*Material C*	*Material D*
64.7	60.4	58.3	60.8
64.8	61.8	62.1	60.2
66.8	63.3	62.4	59.8
67.0	61.6	60.3	58.3
64.9	61.0	60.6	56.4
63.7	63.8	60.0	61.6
61.8	60.9	60.3	59.5
64.3	65.1	62.4	62.0
64.3	61.5	61.9	61.4
65.9	60.0	63.1	58.6
63.6	62.9	60.2	59.5
64.6	60.6	58.6	60.0

It is of considerable interest to determine which of the materials produces the strongest widgets, as well as which materials are statistically different from each other.

There are two ways to enter data for multiple samples into a datasheet:

1. Use a separate column for each sample.

2. Use a single column for all of the data and create a second column to hold codes identifying which sample each observation comes from.

For this example, the first approach has been selected. The data for the widgets have been placed in four columns of a file called *widgets.sgd*, which you can open by selecting *Open - Open Data Source* from the *File* menu.

12.1 Running the Multiple Sample Comparison Procedure

The *Multiple Sample Comparison* procedure is available on the main menu under:

1. If using the Classic menu, select: *Compare – Multiple Sample Comparisons – Multiple-Sample Comparison.*

2. If using the Six Sigma menu, select *Analyze – Variable Data – Multiple Sample Comparisons – Multiple-Sample Comparison.*

The initial dialog box is used to indicate how the data have been structured:

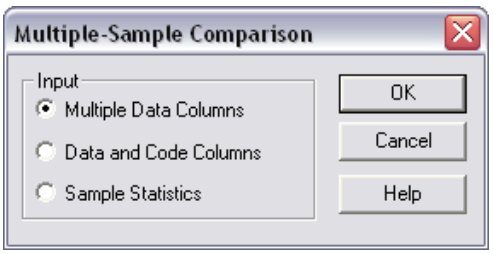


Figure 12-1. Initial Multiple Sample Comparison Dialog Box

In this case, the data have been placed in multiple columns of the datasheet.

The second dialog box requests the names of the columns containing the data:

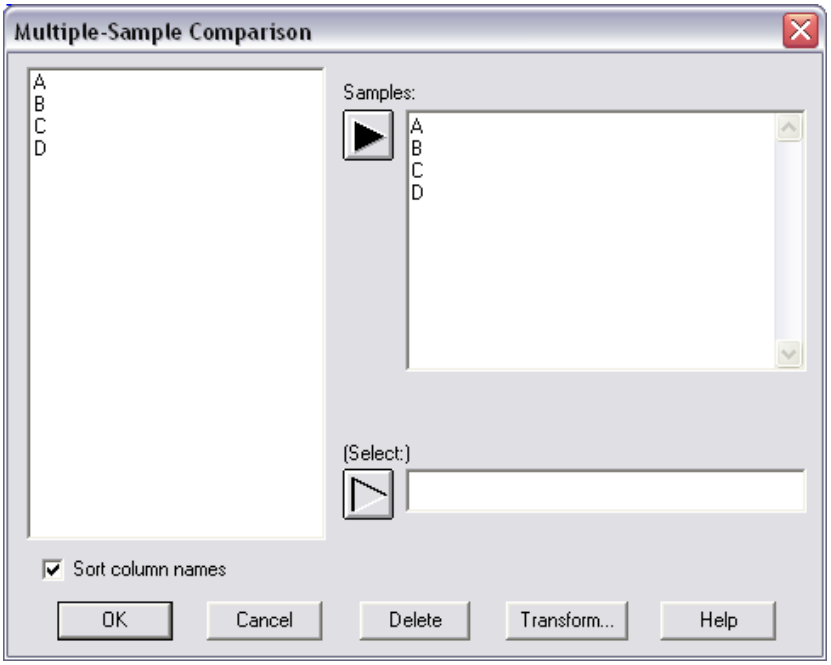


Figure 12-2. Multiple Sample Comparison Data Input Dialog Box

In the sample data file, the observations have been placed in four columns named A, B, C, and D.

When *OK* is pressed, the *Tables and Graphs* dialog box will appear. The default settings are acceptable right now for this tutorial.

When the analysis window opens, it will have four panes:

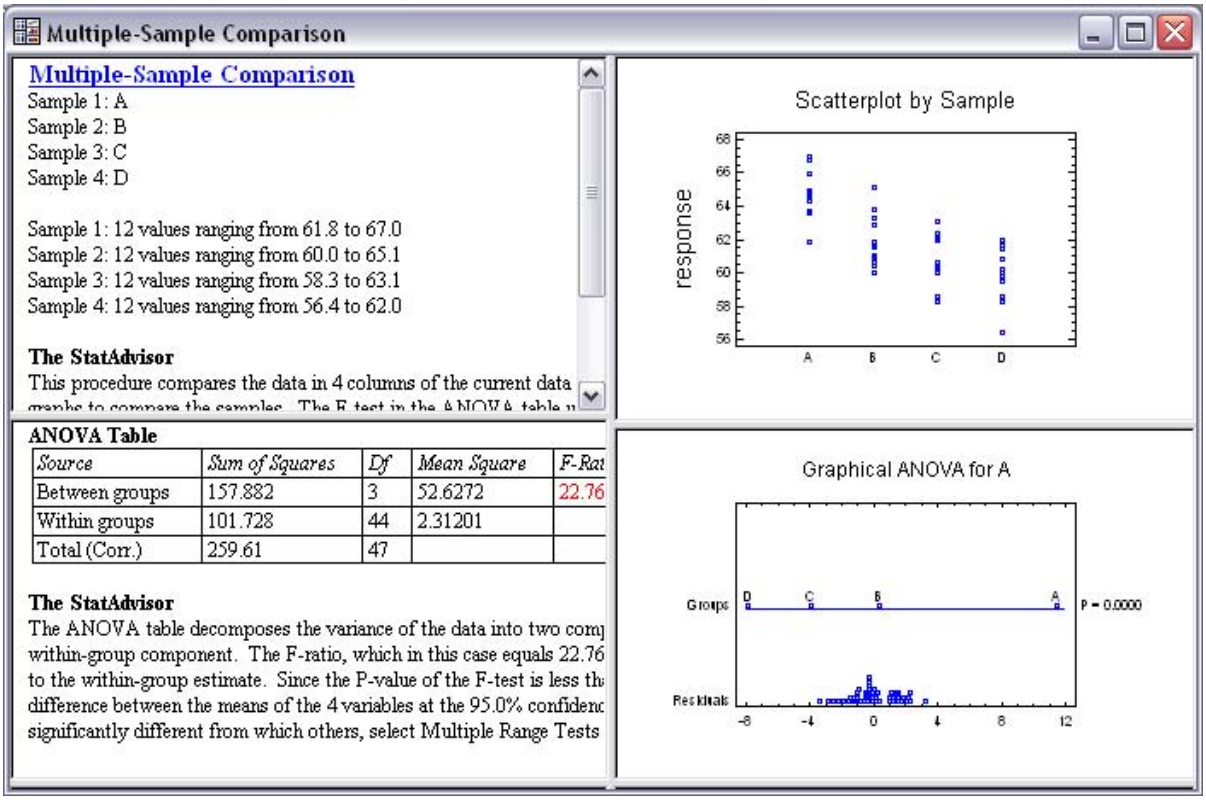


Figure 12-3. Multiple Sample Comparison Analysis Window

The top left pane summarizes the size of each sample and its range. The top right pane shows a scatterplot of the data, enlarged below:

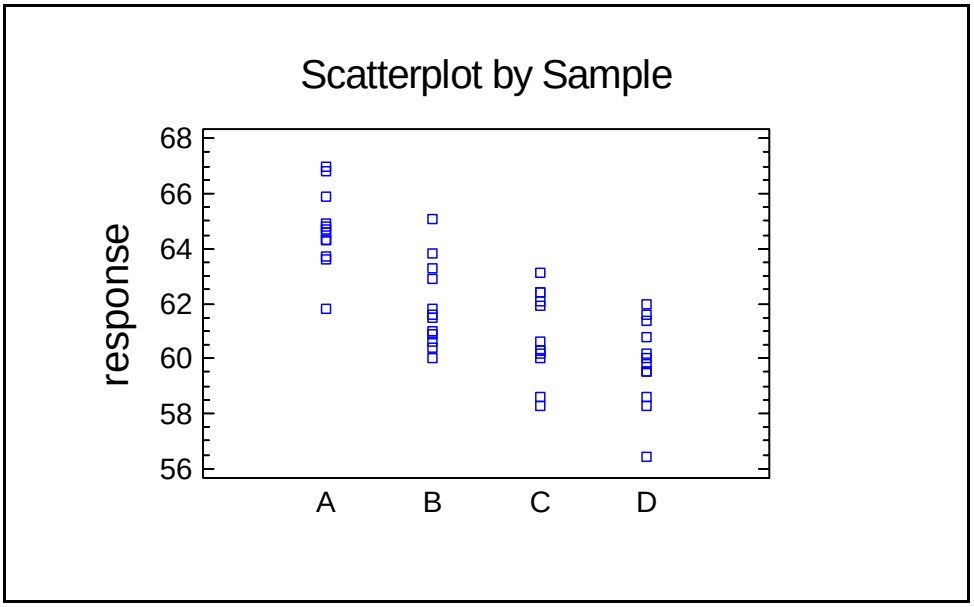


Figure 12-4. Scatterplot of Strength versus Material

Note that many of the observations plot on top of one another. To alleviate this problem, double-click on the graphics pane to maximize it and then press the *Jitter* button on the analysis toolbar and add a small amount of horizontal jitter by moving the top slider slightly to the right:

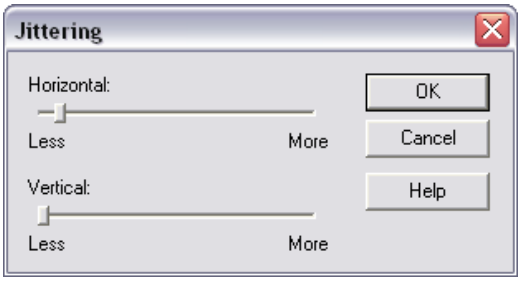


Figure 12-5. Jittering Dialog Box

This randomly offsets each point a small amount in the horizontal direction, making the individual points easier to see:

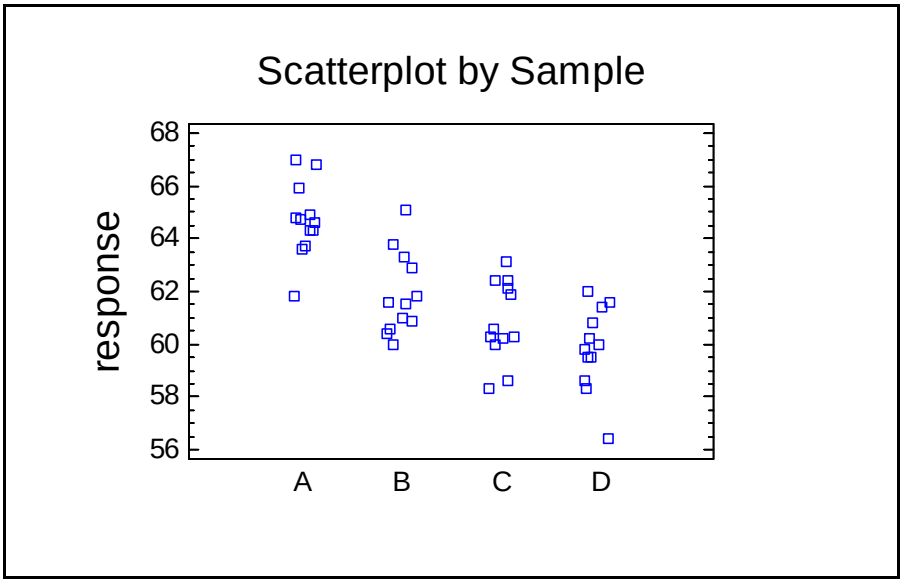


Figure 12-6. Scatterplot after Jittering

Jittering affects only the display, not the data or any calculations made from it.

12.2 Analysis of Variance

The first step when comparing multiple samples is usually to perform a oneway analysis of variance (ANOVA). The ANOVA is used to test the hypothesis of equal population means by choosing between the following two hypotheses:

Null hypothesis: $\mu_A = \mu_B = \mu_C = \mu_D$

Alternative hypothesis: the means are not all equal

where μ_j represents the mean of the population from which sample *j* was taken. Rejection of the null hypothesis indicates that the samples come from populations whose means are not all identical.

The output of the ANOVA is contained in the ANOVA table, which is initially displayed in the bottom left pane of the analysis window:

ANOVA Table

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	157.882	3	52.6272	22.76	0.0000
Within groups	101.728	44	2.31201		
Total (Corr.)	259.61	47			

Figure 12-7. Analysis of Variance Table

The analysis of variance decomposes the variability of the observed data into two components: a between-group component, quantifying differences between widgets made of different materials, and a within-group component, quantifying differences between widgets made of the same material. If the estimated variability between groups is significantly larger than the estimated variability within groups, it is evidence that the group means are not all the same.

The key quantity in *Figure 12-7* is the *P-value*. Small *P*-values (less than 0.05 if operating at the 5% significance level) lead to a rejection of the hypothesis of equal means. In the current example, there is little doubt that the means are significantly different.

In the latest edition of Statistics for Experimenters by Box, Hunter and Hunter (John Wiley and Sons, 2005), the authors present a new display designed to show the results of an ANOVA in graphical format. Their *Graphical ANOVA* is displayed by default in the lower right pane:

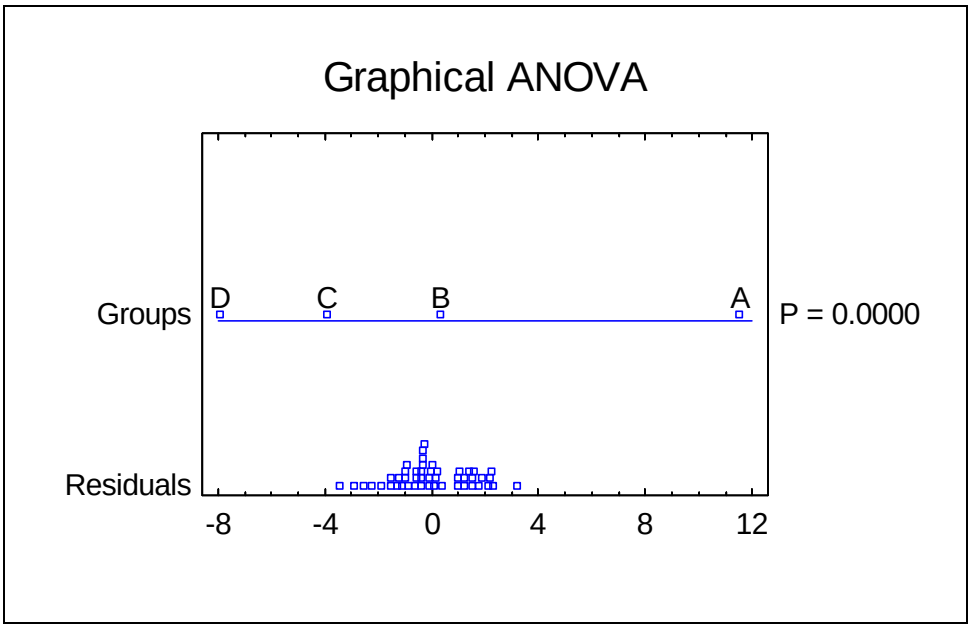


Figure 12-8. Graphical ANOVA

Along the bottom of the plot is a dot diagram of the model *residuals*. In a oneway ANOVA, the residuals are equal to the difference between each observation and the mean of all observations in its group. In the current example, the observed variability in the residuals is indicative of the natural variability amongst widgets made of the same material. Plotted above the central line are scaled deviations of the group means from the overall mean of all $n = 48$ observations. These group deviations are scaled so that their variability can be compared to that of the residuals. Any groups whose points are too far apart to have easily come from a distribution with spread similar to that of the residuals likely correspond to different populations.

In Figure 12-8, group A appears to be well separated from the other groups. Separation of the other three means is less clear. A more formal comparison of the four sample means is described in the next section.

12.3 Comparing Means

If the *P*-value in the ANOVA table is small, then the sample means should be examined to determine which means are significantly different from which others. A useful plot for this purpose is the *Means Plot*, available from the *Tables and Graphs* dialog box:

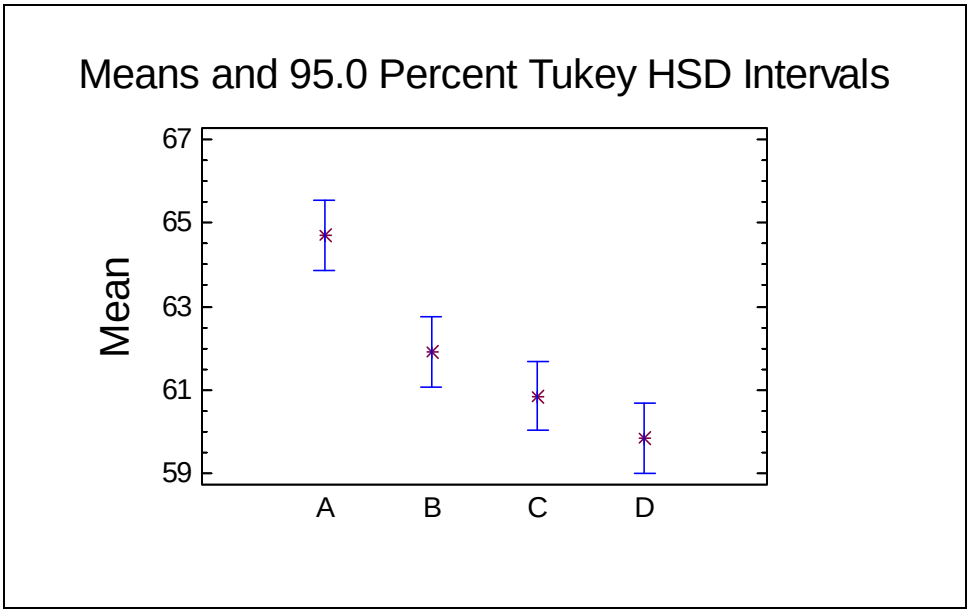


Figure 12-9. Means Plot

The means plot shows each sample mean, together with an uncertainty interval surrounding it. Interpretation of the intervals depends upon the type of interval plotted, which may be changed using *Pane Options*. The two most commonly used intervals are:

1. *Fisher's LSD (Least Significant Difference) Intervals*: These intervals are scaled in such a way that one can select a single pair of samples and declare their means to be significantly different if the intervals do not overlap in the vertical direction. While the chance of incorrectly declaring two samples to be different with this method is fixed at 5%, making comparisons amongst many pairs of means may result in an error on at least one pair with a considerably higher probability.

2. *Tukey's HSD (Honestly Significant Difference) Intervals*. These intervals are scaled to control the experiment-wide error rate at 5%. Using Tukey's method, you will not incorrectly declare *any* pair of means to be significantly different when they're really not in more than 5% of the analyses you do.

The intervals in *Figure 12-9* use Tukey's method. Since the interval for sample A does not overlap any of the other intervals, the mean of sample A is significantly different than that of all the other 3 samples. Sample B is also significantly different than sample D, since their intervals do not overlap. Sample C, however, is not significantly different than either B or D.

The same analysis can be displayed in tabular form by selecting *Multiple Range Tests* from the *Tables and Graphs* dialog box:

Multiple Range Tests

Method: 95.0 percent Tukey HSD

	Count	*Mean*	*Homogeneous Groups*
D	12	59.8417	X
C	12	60.85	XX
B	12	61.9083	X
A	12	64.7	X

Contrast	*Sig.*	*Difference*	*+/- Limits*
A - B	*	2.79167	1.65755
A - C	*	3.85	1.65755
A - D	*	4.85833	1.65755
B - C		1.05833	1.65755
B - D	*	2.06667	1.65755
C - D		1.00833	1.65755

* denotes a statistically significant difference.

Figure 12-10. Multiple Range Tests

The bottom section of the output shows each pair of means. The *Difference* column displays the sample mean of the first group minus that of the second. The +/- *Limits* column shows an uncertainty interval for the difference. Any pair for which the absolute value of the difference exceeds the limit is statistically significant at the selected significance level and is indicated by an * in the *Sig.* column. In the current example, four of the six pairs of means show significant differences.

The top section of the display arranges the samples into homogeneous groups, shown as columns of X's. A homogeneous group is a group within which there are no significant differences. In this case, sample A is in a group by itself, since it is significantly different than all of the others. Sample C falls in two groups, one with B and one with D. More data would be required to distinguish which group sample C actually belongs to.

12.4 Comparing Medians

If it is suspected that outliers may be present, a nonparametric procedure may be used as an alternative to the standard analysis of variance by selecting *Kruskal-Wallis* and *Friedman Tests* from the *Tables* dialog box. These tests compare the sample medians rather than the means:

Null hypothesis: the medians are all equal

Alternative hypothesis: the medians are not all equal

The type of test may be selected using *Pane Options*. Two types of tests are provided:

1. *Kruskal-Wallis test* – appropriate when each column contains a random sample from its population. In such a case, the rows have no intrinsic meaning.

2. *Friedman test* – appropriate when each row represents a block, i.e., the level of some other variable. Typical blocking variables are day of the week, shift, or manufacturing location.

In the example, row has no meaning, so the Kruskal-Wallis test is appropriate:

Kruskal-Wallis Test

	Sample Size	*Average Rank*
A	12	40.7917
B	12	25.7917
C	12	19.25
D	12	12.1667

Test statistic = 27.3735 P-Value = 0.00000491592

Figure 12-11. Multiple Range Tests

The important entry in the above table is the *P*-value. Since the *P*-value is small (less than 0.05), the hypothesis of equal medians is rejected.

Pairs of medians can also be compared by selecting *Box-and-Whisker Plot* from the *Tables and Graphs* dialog box and using *Pane Options* to add median notches:

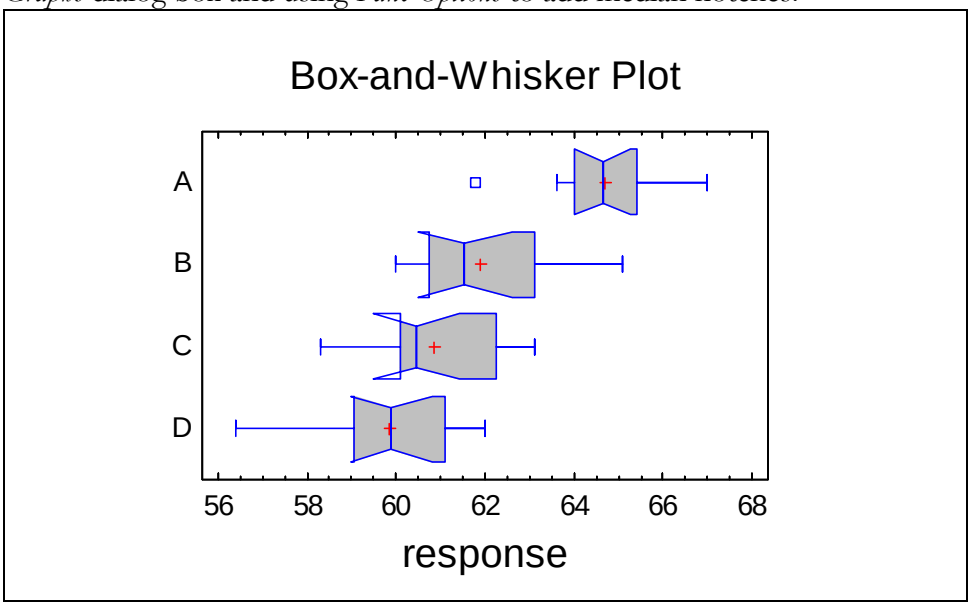


Figure 12-12. Box-and-Whisker Plots with Median Notches

The range covered by each notch shows the uncertainty associated with the estimate of that group's median. The notches are scaled in such a way that any two samples whose notches do not overlap can be declared to have significantly different medians at the default system significance level (usually 5%). In the above plot, the notches for samples B, C and D all overlap, but the median for sample A is significantly higher than that of the other 3 samples.

NOTE: the folding back behavior observed in Figure 12-12 occurs when a notch extends beyond the edge of the box.

12.5 Comparing Standard Deviations

It is also possible to test the hypothesis of equal standard deviations:

Null hypothesis: $\sigma_A = \sigma_B = \sigma_C = \sigma_D$

Alternative hypothesis: the standard deviations are not all equal

This is done by selecting *Variance Check* from the *Tables and Graphs* dialog box:

Variance Check

	Test	*P-Value*
Levene's	0.143286	0.933432

Figure 12-13. Comparison of Sample Variances

One of four tests will be displayed, depending on the settings for *Pane Options*. Three of the available tests, including Levene's test, display P-values. A *P*-value less than 0.05 leads to rejection of the hypothesis of equal sigmas at the 5% significance level. In this case, the standard deviations are not significantly different from one another, since the *P*-value is well above 0.05.

In summary, it appears that the mean strength is different for different materials. However, the variability amongst widgets made of the same material is about the same across all four materials.

12.6 Residual Plots

Whenever a statistical model is fit to data, it is important to examine the residuals from the fitted model. In this analysis, there is a residual corresponding to each of the $n = 48$ widgets, defined as the difference between a widget's strength and the average strength of all widgets made of that same material.

The *Graphs* dialog box contains an entry for automatically generating plots of the residuals. Depending on the selection in *Pane Options*, you may plot the residuals by group, versus predicted values, or in row order as found in the datasheet. The plot below shows the residuals plotted versus predicted *strength*:

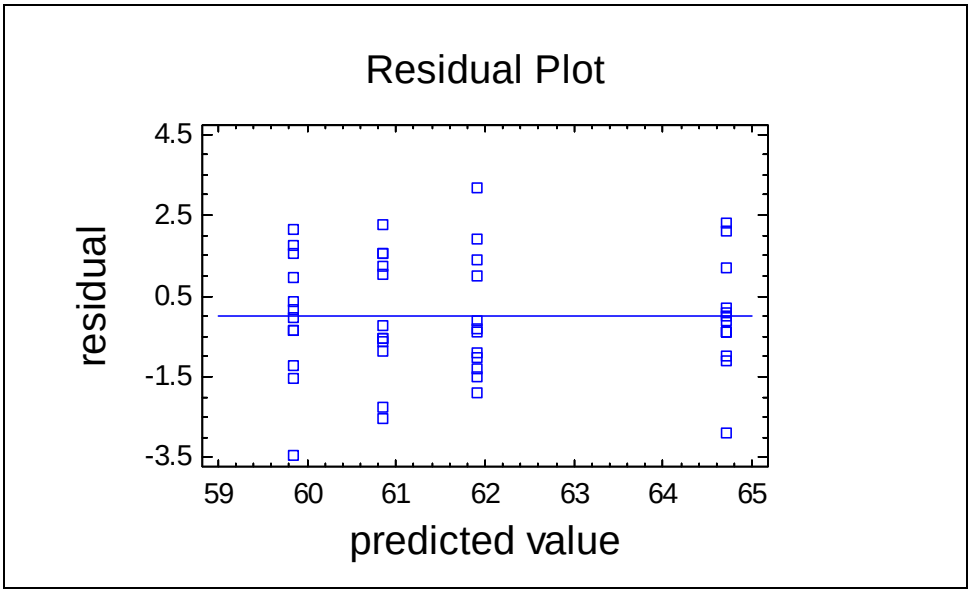


Figure 12-14. Plot of residuals Versus Predicted Strength

In these types of plots, you should look for:

1. *Outliers* – isolated residuals far away from all of the others. Such points would need further investigation to determine whether an assignable cause exists that explains their unusual behavior.

2. *Heteroscedasticity* – a systematic change in the variance as the predicted values increase or decrease. This condition typically results in a funnel-like appearance in the plot and might necessitate transforming the original observations by taking logarithms of the data before performing the analysis. Procedures such as the *Multiple Range Tests* will not work properly when the within-group variability differs significantly amongst the groups.

If desired, the residuals may be saved to a column of any datasheet by pressing the *Save Results* button  on the analysis toolbar.

12.7 Analysis of Means Plot (ANOM)

A somewhat different way to compare several means is by using an *Analysis of Means Plot*, also available on the *Tables and Graphs* dialog box:

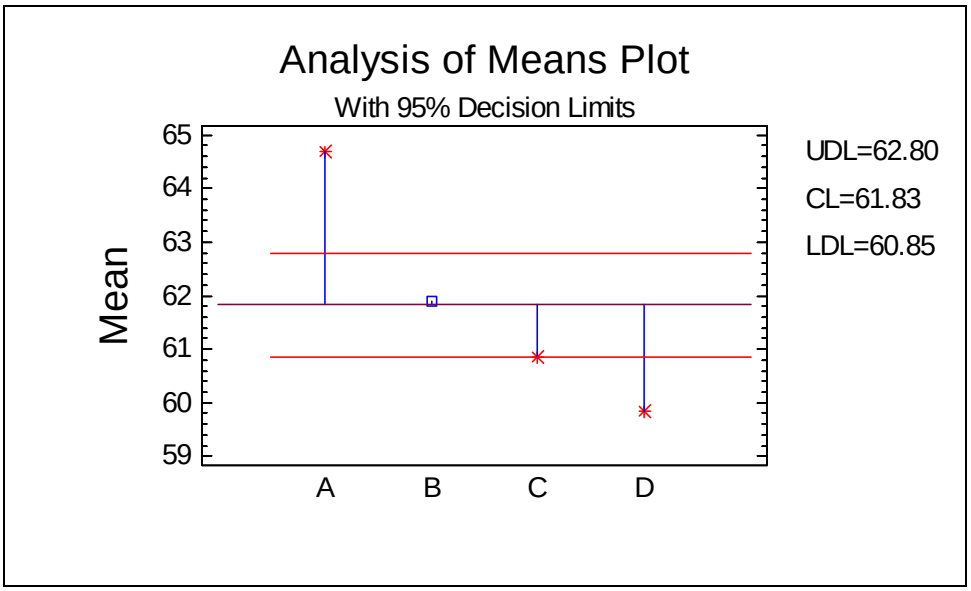


Figure 12-15. Analysis of Means Plot

Designed to be similar to a control chart, this plot displays each sample mean together with a vertical line drawn to the grand mean of all the observations. Decision limits are included above and below the grand mean. Any sample means that fall outside the limits may be declared to be significantly different than the grand mean.

In this case, the interpretation is that widgets from sample A are significantly stronger than average, while widgets from samples C and D are significantly weaker than average. This type of interpretation can sometimes be quite useful.

Chapter

13

Tutorial #4: Regression Analysis

Fitting linear and nonlinear models, selecting the best model, plotting residuals, and displaying results.

One of the most heavily used sections of STATGRAPHICS Centurion XVI is the set of procedures that fit statistical regression models. In a regression model, a response variable Y is expressed as a function of one or more predictor variables X, plus noise. In many (but not all) cases, the functional form is linear in the unknown coefficients, so that the model can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 x_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i$$

where the subscript *i* represents the i^{th} observation in the data sample, the β's are unknown model coefficients, and ε is a random deviation, usually assumed to come from a normal distribution with mean 0 and standard deviation σ.

Given a set of data with a response variable Y and one or more possible predictor variables, the goal of regression analysis is to construct a model that:

1. Describes the relationships that exist between the variables in a manner that permits Y to be predicted well given known values of the X's.

2. Contains no more X variables than it necessary to generate a good prediction.

The latter consideration is sometimes referred to as *parsimony*. Typically, models involving a small set of well-selected predictors perform best in practice.

This chapter considers several types of regression models. As an example, the miles per gallon in city driving for the automobiles in the *93cars.sgd* file will serve as the response variable Y. The goal is to build a model from the other columns in that file that can successfully predict the miles per gallon of an automobile.

13.1 Correlation Analysis

A useful place to start when beginning to build a regression model is with the *Multiple Variable Analysis* procedure. This analysis may be found on the main menu under:

1. If using the Classic menu, select *Describe – Numeric Data – Multiple-Variable Analysis*.

2. If using the Six Sigma menu, select *Analyze – Variable Data – Multivariate Methods – Multiple-Variable Analysis*.

The analysis begins by displaying the following data input dialog box:

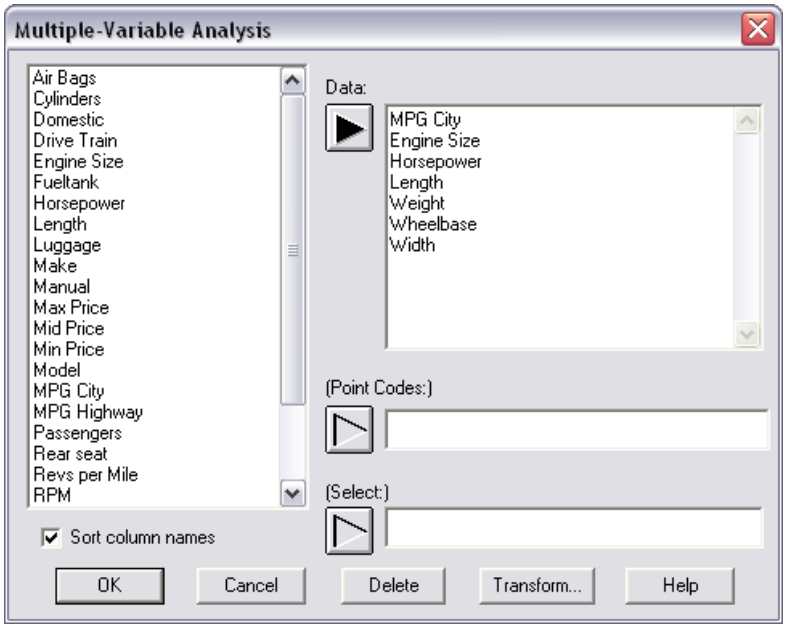


Figure 13-1. Multiple Variable Analysis Data Input Dialog Box

Six possible predictor variables have been selected, in addition to *MPG City*. The potential predictors are:

X_1: *Engine Size* (liters)
X_2: *Horsepower* (maximum)
X_3: *Length* (inches)
X_4: *Weight* (pounds)
X_5: *Wheelbase* (inches)
X_6: *Width* (inches)

Pressing *OK* displays the *Options* menu, then the *Tables and Graphs* dialog box, then the analysis window:

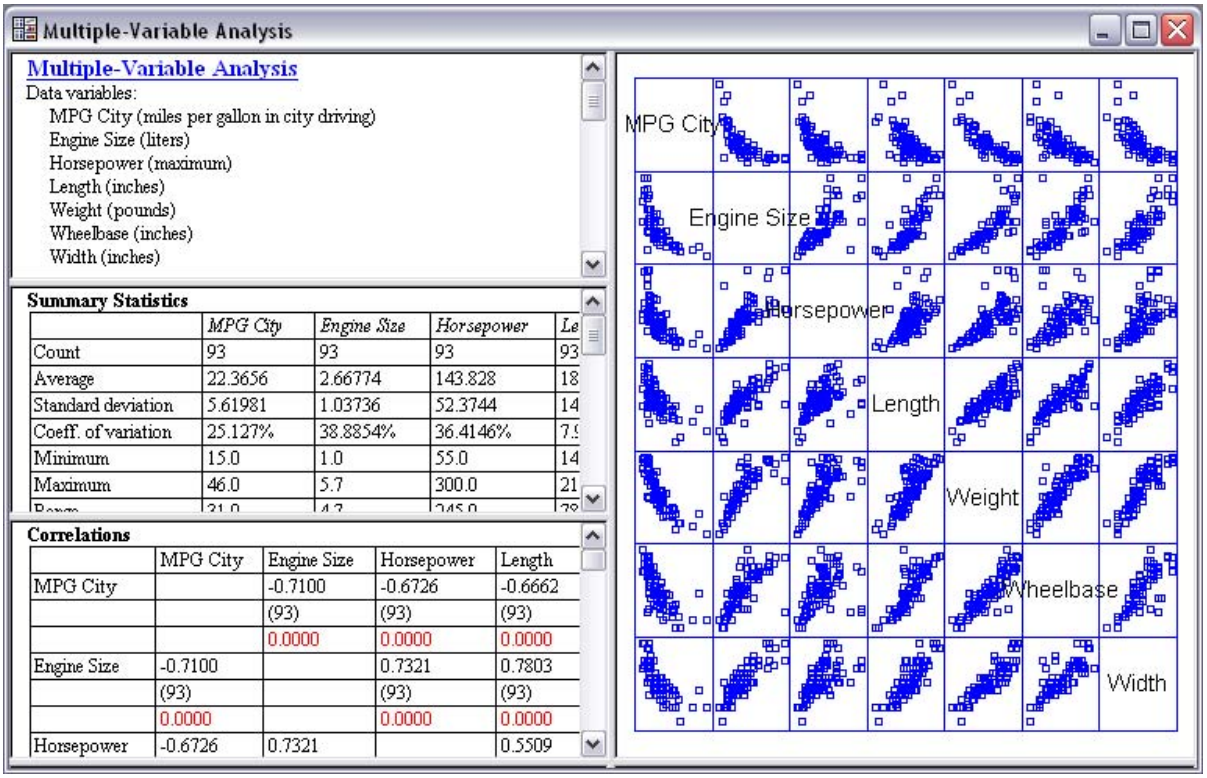


Figure 13-2. Multiple Variable Analysis Window

The upper left pane lists the input variables, while the center left pane displays summary statistics. There are a total of 93 rows in the data file that have complete information on all of the variables to be analyzed.

The matrix plot on the right displays X-Y plots for each pair of variables:

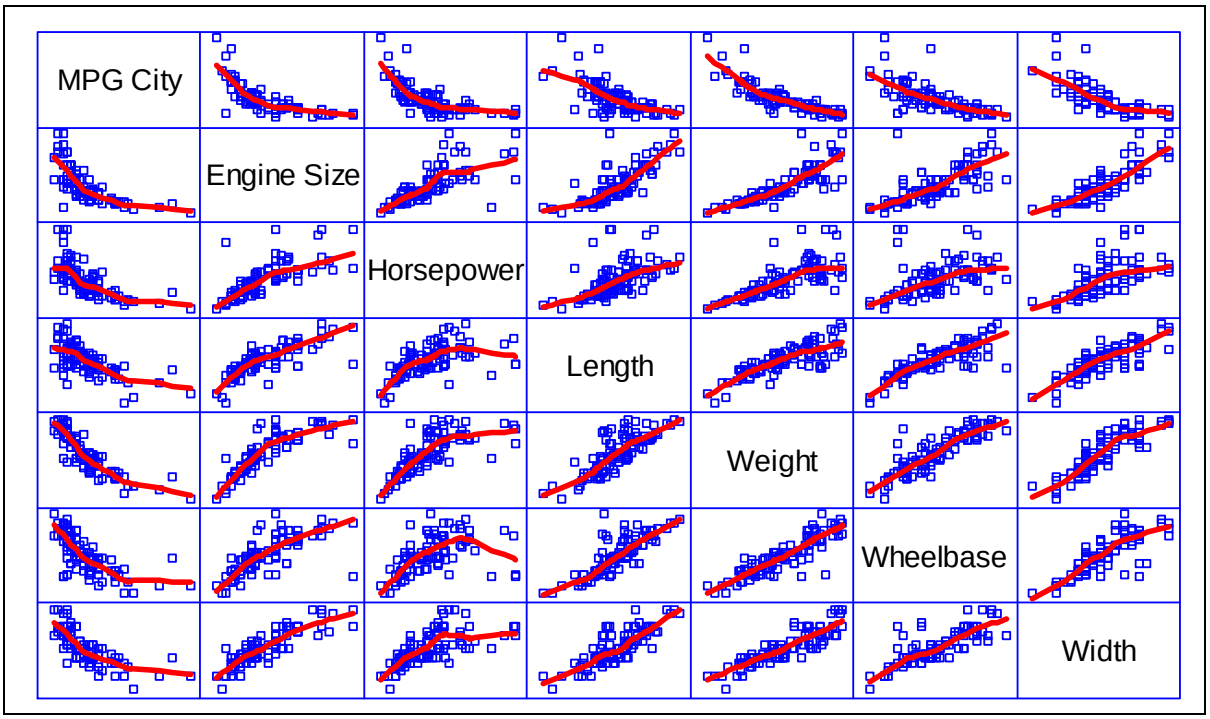


Figure 13-3. Matrix Plot with Added Smooth

To interpret the plot, find a variable's label, such as *MPG City*. The indicated variable is displayed on the vertical axis of every plot in that row and on the horizontal axis of every plot in that column. Each pair of variables is thus shown twice, once above the diagonal and once below it.

Robust LOWESS smoothers have been added in the above figure by maximizing the pane and selecting the *Smooth/Rotate* button on the analysis toolbar. Of most interest is the top row of plots, which show *MPG City* plotted versus each of the 6 potential predictor variables. All of the variables are clearly correlated with miles per gallon, some in a nonlinear manner. There is also a great deal of multicollinearity present (correlation amongst the predictor variables), which suggests that many different combinations of variables may be equally good at predicting Y.

The table at the bottom left shows a matrix of estimated correlation coefficients for every pair of variables in the analysis:

Correlations

	MPG City	Engine Size	Horsepower	Length	Weight	Wheelbase	Width
MPG City		-0.7100	-0.6726	-0.6662	-0.8431	-0.6671	-0.720
		(93)	(93)	(93)	(93)	(93)	(93)
		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Engine Size	-0.7100		0.7321	0.7803	0.8451	0.7325	0.8671
	(93)		(93)	(93)	(93)	(93)	(93)
	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000
Horsepower	-0.6726	0.7321		0.5509	0.7388	0.4869	0.6444
	(93)	(93)		(93)	(93)	(93)	(93)
	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000
Length	-0.6662	0.7803	0.5509		0.8063	0.8237	0.8221
	(93)	(93)	(93)		(93)	(93)	(93)
	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
Weight	-0.8431	0.8451	0.7388	0.8063		0.8719	0.8750
	(93)	(93)	(93)	(93)		(93)	(93)
	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000
Wheelbase	-0.6671	0.7325	0.4869	0.8237	0.8719		0.8072
	(93)	(93)	(93)	(93)	(93)		(93)
	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000
Width	-0.7205	0.8671	0.6444	0.8221	0.8750	0.8072	
	(93)	(93)	(93)	(93)	(93)	(93)	
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	

Correlation
(Sample Size)
P-Value

Figure 13-4. Correlation Matrix

The table shows the correlation coefficient for each pair of variables, the number of observations used to obtain the estimate, and a *P*-value. A correlation coefficient *r* is a number between -1 and +1, which measures the strength of the linear relationship between two variables. The closer the correlation is to -1 or +1, the stronger the relationship. The sign of the correlation indicates the direction of the relationship. A positive value means that Y goes up as X goes up. A negative value means that Y goes down as X goes down.

To determine whether or not two variables are significantly related to each other, a *P*-value is calculated for each correlation coefficient. Any pair of variables for which the *P*-value is less than 0.05 exhibits a statistically significant linear correlation at the 5% significance level.

The top row shows the correlations between *MPG City* and the 6 predictors. The strongest correlation is with *Weight*, at -0.8431. The negative sign implies that as *Weight* increases, *MPG City* decreases, which is not at all surprising.

13.2 Simple Regression

The first statistical model that will be fit is a straight line of the form:

$$MPG\ City = \beta_0 + \beta_1 Weight + \varepsilon$$

In the above equation, β_1 is the slope of the line in units of miles per gallon per pound, while β_o is the Y-intercept. To fit this model:

1. If using the Classic menu, select *Relate – One Factor – Simple Regression.*

2. If using the Six Sigma menu, select *Improve – Regression Analysis – One Factor – Simple Regression.*

The data input dialog box should be completed as follows:

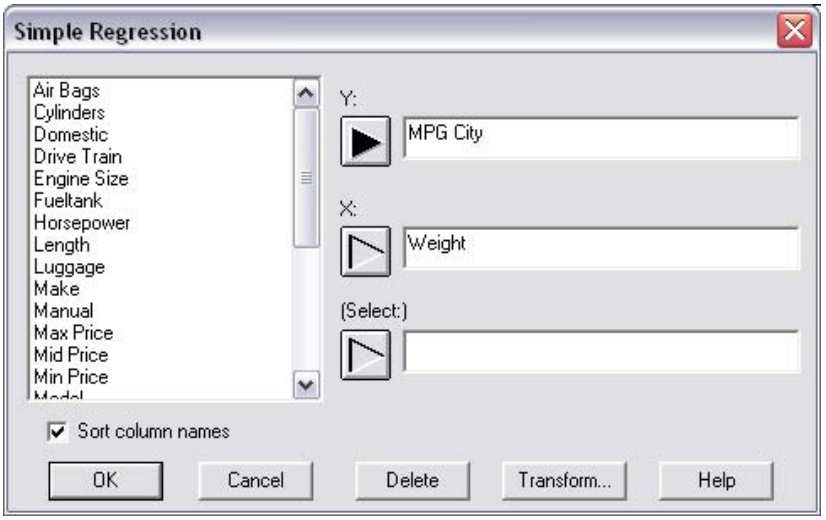


Figure 13-5. Simple Regression Data Input Dialog Box

After the *Options* menu and the *Tables and Graphs* dialog box, the initial analysis window has four panes providing information about the fitted model and the residuals:

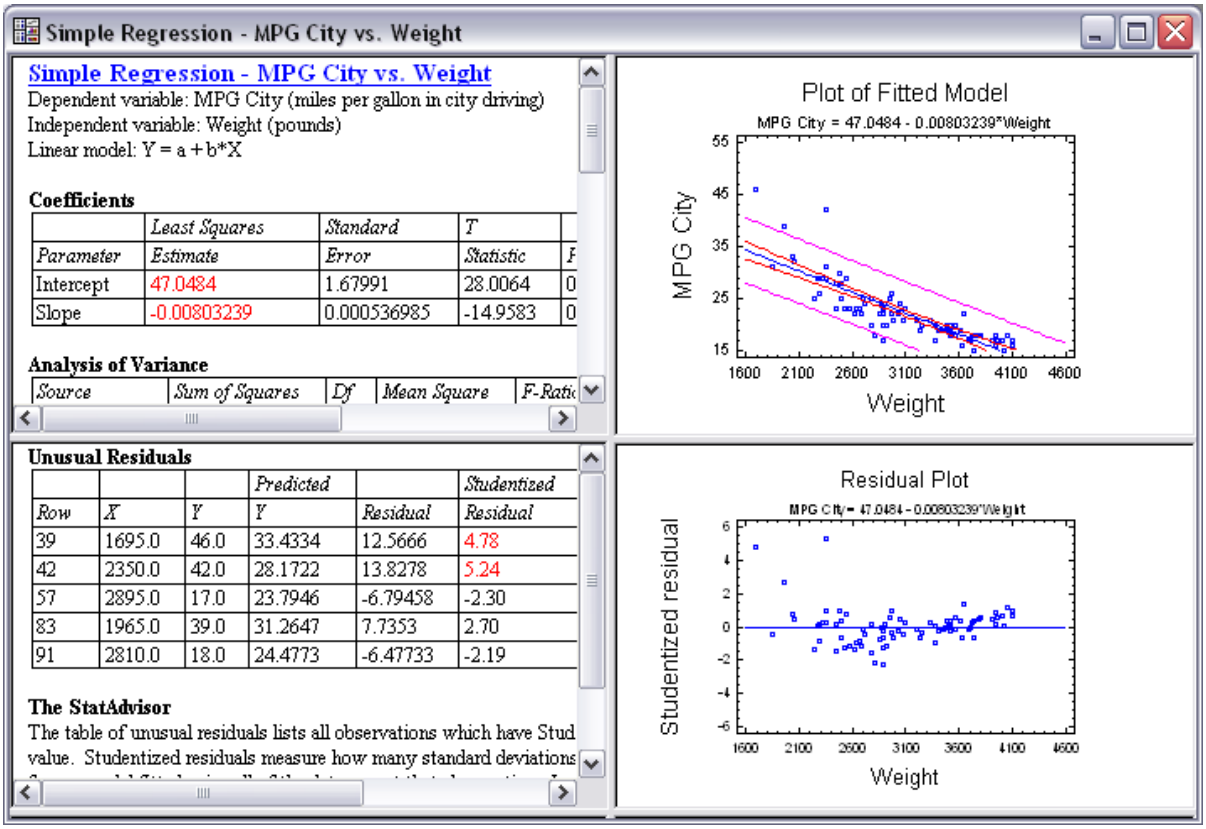


Figure 13-6. Simple Regression Analysis Window

The *Analysis Summary* in the top left pane summarizes the fit:

Simple Regression - MPG City vs. Weight

Dependent variable: MPG City (miles per gallon in city driving)
Independent variable: Weight (pounds)
Linear model: Y = a + b*X

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	47.0484	1.67991	28.0064	0.0000
Slope	-0.00803239	0.000536985	-14.9583	0.0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	2065.52	1	2065.52	223.75	0.0000
Residual	840.051	91	9.23133		
Total (Corr.)	2905.57	92			

Correlation Coefficient = -0.843139
R-squared = 71.0883 percent
R-squared (adjusted for d.f.) = 70.7705 percent
Standard Error of Est. = 3.03831
Mean absolute error = 1.99274
Durbin-Watson statistic = 1.64586 (P=0.0405)
Lag 1 residual autocorrelation = 0.176433

Figure 13-7. Simple Regression Analysis Summary

Of the many statistics in the above table, the following are the most important:

1. **Coefficients**: the estimated model coefficients. The fitted model that would be used to make predictions is:

$$MPG\ City = 47.0484 - 0.00803239 Weight$$

2. **R-squared**: the percentage of the variability in Y that has been explained by the model. In this case, a linear regression against *Weight* explains about 71.1% of the variability in *MPG City*.

3. **Model *P*-Value**: tests the null hypothesis that the fitted model is no better than a model that does not include *Weight*. A *P*-value below 0.05, as in the current example, indicates that *Weight* is a useful predictor of *MPG City*.

The plot in the top right pane displays the fitted model:

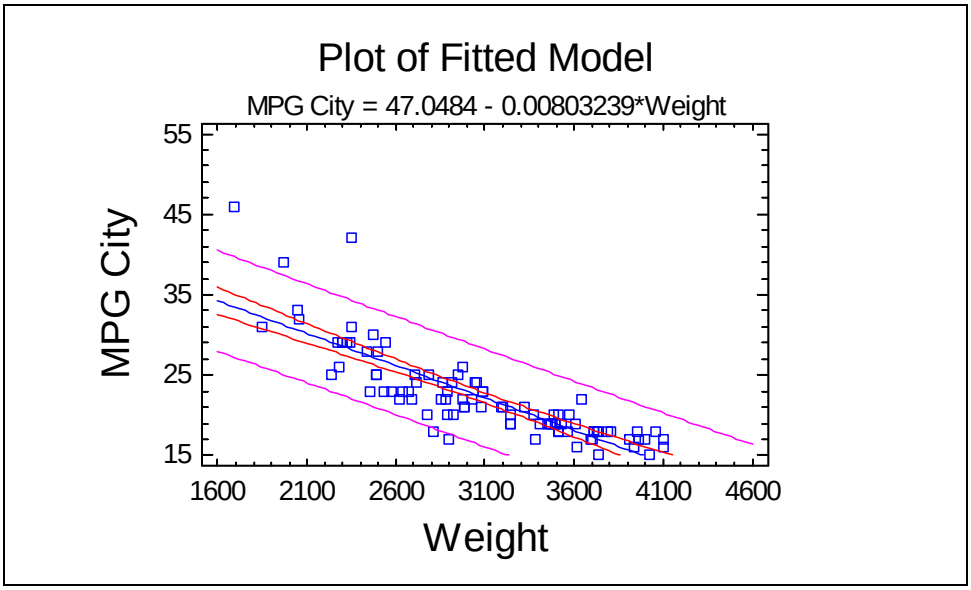


Figure 13-8. Plot of Fitted Linear Model

The plot shows the least squares regression line and two sets of limits. The inner limits provide 95% confidence intervals for the mean value of Y at any selected X. This indicates how well the location of the line has been estimated, given that the relationship is linear. The larger the sample, the tighter the limits. The outer lines are 95% prediction limits for new observations. It is estimated that 95% of additional observations, similar to those in the sample, would fall within those bounds.

It is worthy of note that 3 observations at low values of *Weight* fall fairly far beyond the 95% prediction limits. This may be indicative either of outliers or of the failure of the model to account for the nonlinearity of the actual relationship between *MPG City* and *Weight*.

13.3 Fitting a Nonlinear Model

The *Simple Regression* procedure includes the ability to fit a wide variety of nonlinear models. To assess the relative improvement that various models could make, select *Comparison of Alternative Models* from the *Tables and Graphs* dialog box. This will fit all of the possible models and list them in decreasing order of R-squared:

Comparison of Alternative Models

Model	*Correlation*	*R-Squared*
S-curve model	0.9016	81.29%
Reciprocal-Y square root-X	0.8995	80.92%
Reciprocal-Y logarithmic-X	0.8995	80.90%
Square root-Y reciprocal-X	0.8988	80.78%
Multiplicative	-0.8981	80.65%
Reciprocal-Y	0.8969	80.44%
Logarithmic-Y square root-X	-0.8919	79.54%
Double reciprocal	-0.8896	79.14%
Reciprocal-X	0.8888	79.00%
Square root-Y logarithmic-X	-0.8879	78.83%
Reciprocal-Y squared-X	0.8852	78.35%
Exponential	-0.8833	78.03%
Double square root	-0.8784	77.16%
Logarithmic-X	-0.8705	75.78%
Square root-Y	-0.8668	75.14%
Logarithmic-Y squared-X	-0.8611	74.15%
Square root-X	-0.8577	73.56%
Squared-Y reciprocal-X	0.8472	71.77%
Linear	-0.8431	71.09%
Square root-Y squared-X	-0.8393	70.44%
Squared-Y logarithmic-X	-0.8146	66.35%
Squared-X	-0.8106	65.71%
Squared-Y square root-X	-0.7957	63.31%
Squared-Y	-0.7758	60.18%
Double squared	-0.7346	53.96%
Logistic	<no fit>	
Log probit	<no fit>	

Figure 13-9. Alternative Nonlinear Models

The models at the top of the list explain the largest percentage of the variation in the response variable. R-squared is only one criteria that can be used to help pick a model. Models with somewhat lower R-squared values than the model at the top of the list may be preferable if they make more sense in the context of the data.

In the current example, an attractive model near the top of the list is the *Reciprocal-Y* model. This model takes the form:

$$\frac{1}{MPGCity} = \beta_0 + \beta_1 Weight + \varepsilon$$

In it, the reciprocal of miles per gallon (gallons per mile) is expressed as a linear function of weight. It is not uncommon that transformations of Y, X, or both may lead to better models. To fit a *Reciprocal-Y* model, press the *Analysis Options* button and select *Reciprocal-Y* on the dialog box. The resulting fit is shown below:

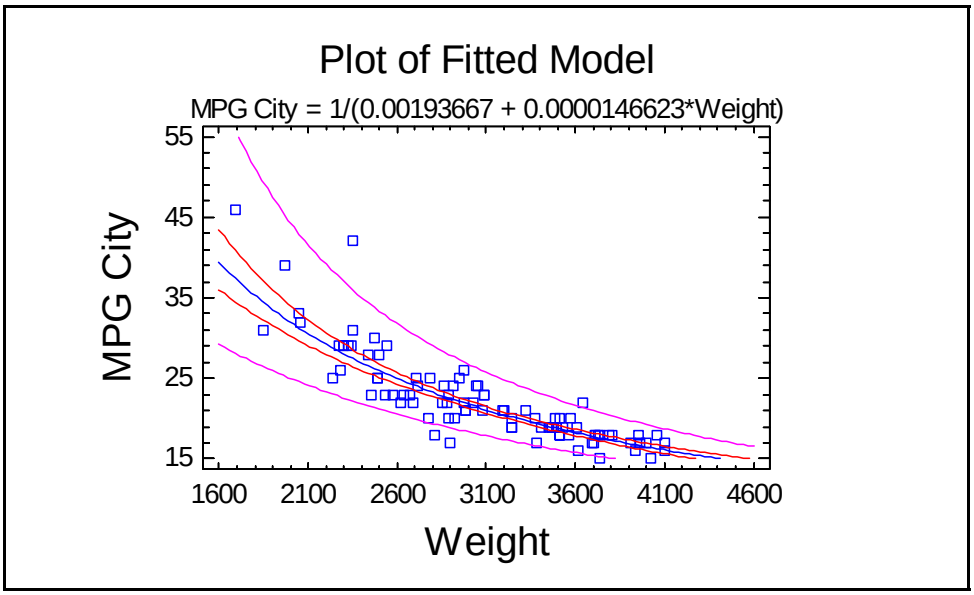


Figure 13-10. Fitted Reciprocal-Y Model

While linear in the reciprocal of *MPG City*, the model is nonlinear in the original metric. Note also that the prediction limits for *Weight* become larger as the predicted values become larger. This makes sense in the context of the data, since it implies that there is more variability amongst the lighter cars than amongst the heavier cars.

13.4 Examining the Residuals

Once a reasonable model has been fit, the residuals from the fit should be examined. In general, a residual may be thought of as the difference between the observed value of Y and the value predicted by the model:

residual = observed Y – predicted Y

The *Simple Regression* analysis automatically plots residuals versus the X variable:

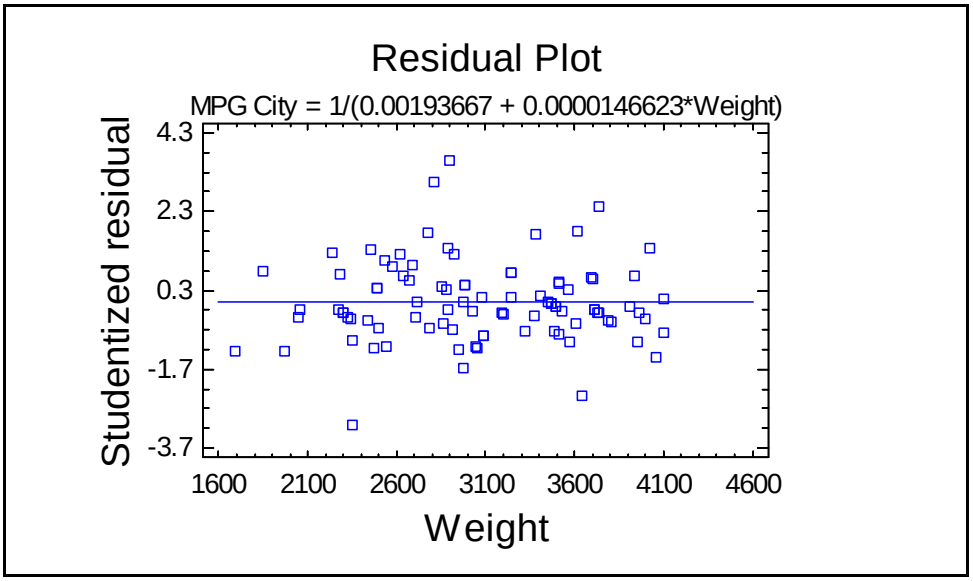


Figure 13-11. Plot of Studentized Residuals

Using *Pane Options*, you may elect to plot either simple residuals or Studentized residuals. Studentized residuals reexpress the ordinary residuals defined above by dividing them by their estimated standard errors. A Studentized residual thus indicates how many standard errors the data value is from the fitted model.

STATGRAPHICS Centurion XVI actually calculates Studentized *deleted* residuals. Deleted residuals are calculated by withholding one observation at a time, refitting the model, and determining the number of standard errors that the withheld observation lies from the newly fitted model. This keeps outliers from having a large impact on the model when its residual is calculated.

The *Unusual Residuals* selection on the *Tables and Graphs* dialog box lists all Studentized residuals that are greater than 2 in absolute value:

Unusual Residuals

			Predicted		*Studentized*
Row	*X*	*Y*	*Y*	*Residual*	*Residual*
5	3640.0	22.0	18.0808	3.91924	-2.38
36	3735.0	15.0	17.6366	-2.63658	2.41
42	2350.0	42.0	27.4778	14.5222	-3.11
57	2895.0	17.0	22.5306	-5.53064	3.60
91	2810.0	18.0	23.1816	-5.18157	3.04

Figure 13-12. Table of Unusual Residuals

Studentized residuals greater than 3, such as row #57, are potential outliers that do not appear to belong with the rest of the data. Row #57 corresponds to a Mazda RX-7 which is recorded as achieving only 17 miles per gallon in city driving, although the model predicts 22.5 mpg. Since the next section adds additional variables to the model, which may help improve its predictive ability for such sports cars, row #57 will not be excluded from the fit, although careful attention should be paid to it.

13.5 Multiple Regression

To improve the model, other predictor variables need to be added. This is most easily accomplished using the *Multiple Regression* analysis, which may be found on the main menu under:

1. If using the Classic menu, select *Relate – Multiple Factors – Multiple Regression.*

2. If using the Six Sigma menu, select *Improve – Regression Analysis – Multiple Factors – Multiple Regression.*

The data input dialog box takes the following form:

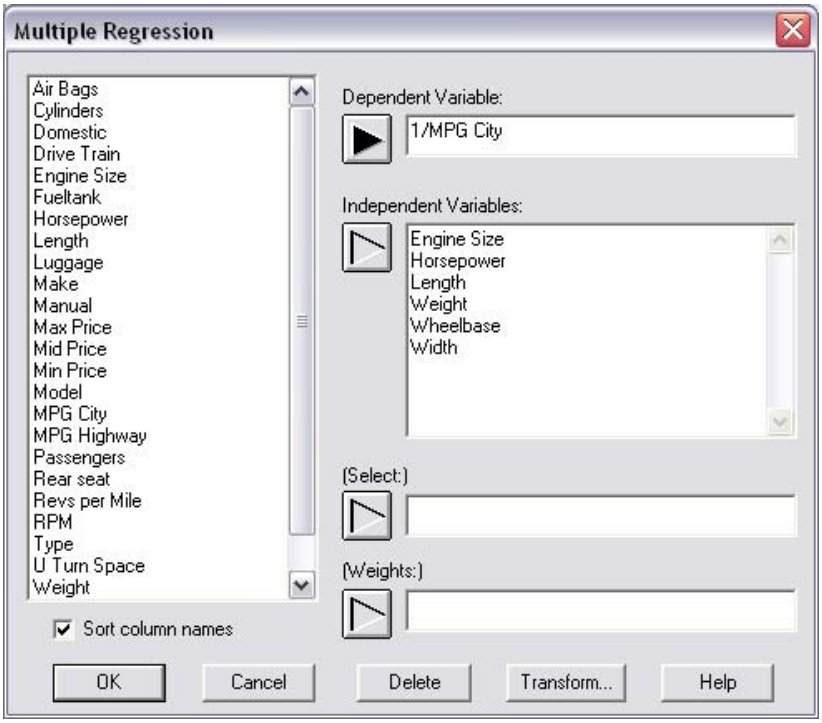


Figure 13-13. Multiple Regression Data Input Dialog Box

To begin, all 6 predictors considered in the *Multiple-Variable Analysis* procedure discussed earlier will be entered as independent variables. The dependent variable is the reciprocal of *MPG City*, which equates to gallons per mile. The *Options* menu is then shown, and then the *Tables and Graphs* dialog box is shown. The resulting analysis summary is shown below:

Multiple Regression - 1/MPG City

Dependent variable: 1/MPG City
Independent variables:
Engine Size (liters)
Horsepower (maximum)
Length (inches)
Weight (pounds)
Wheelbase (inches)
Width (inches)

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	0.0155897	0.0177088	0.880334	0.3811
Engine Size	0.00072849	0.000980504	0.742974	0.4595
Horsepower	0.0000132632	0.000014911	0.889485	0.3762
Length	-0.000101355	0.0000608857	-1.66468	0.0996
Weight	0.0000149727	0.00000242804	6.1666	0.0000
Wheelbase	-0.000148122	0.000163073	-0.908321	0.3662
Width	0.000223526	0.00028967	0.771658	0.4424

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	0.00705967	6	0.00117661	67.64	0.0000
Residual	0.001496	86	0.0000173954		
Total (Corr.)	0.00855567	92			

R-squared = 82.5145 percent
R-squared (adjusted for d.f.) = 81.2946 percent
Standard Error of Est. = 0.00417077
Mean absolute error = 0.00304978
Durbin-Watson statistic = 1.6264 (P=0.0306)
Lag 1 residual autocorrelation = 0.186005

The StatAdvisor
The output shows the results of fitting a multiple linear regression model to describe the relationship between 1/MPG City and 6 independent variables. The equation of the fitted model is

1/MPG City = 0.0155897 + 0.00072849*Engine Size + 0.0000132632*Horsepower - 0.000101355*Length + 0.0000149727*Weight - 0.000148122*Wheelbase + 0.000223526*Width

Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between the variables at the 95% confidence level.

Figure 13-14. Multiple Regression Analysis Summary with 6 Predictor Variables

Notice that the R-squared statistic has risen to 82.5%. However, the model is unnecessarily complicated. Near the top of the output is a column of *P*-values. These *P*-values test the hypothesis that the coefficient corresponding to a selected variable equals 0, given that all of the other variables remain in the model. *P*-values greater than 0.05 indicate that a variable does not contribute significantly to the fit, in the presence of all of the other variables.

Except for *Weight*, all predictors have *P*-values above 0.05. This implies that at least one of those predictor variables could be removed without hurting the model significantly.

> NOTE: It would be wrong at this point to assume that all 5 predictor variables with *P*-values above 0.05 could be removed. Due to the high multicollinearity in the data, all *P*-values may change dramatically if even one variable is removed from the model.

A useful method for simplifying the model is to perform a stepwise regression. In a stepwise regression, variables are added or removed from a regression model one at a time, with the goal of obtaining a model that contains only significant predictors but does not exclude any useful variables. Stepwise regression is available as an option on the *Analysis Options* dialog box:

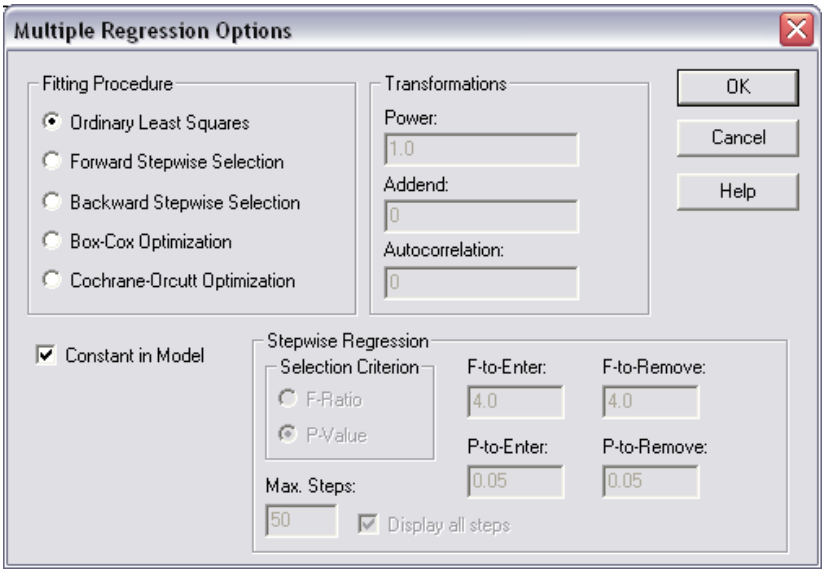


Figure 13-15. Multiple Regression Analysis Options Dialog Box

Two stepwise options are provided:

1. *Forward Selection* – starts with a model containing only a constant and brings variables in one at a time if they improve the fit significantly.

2. *Backward Selection* – starts with a model containing all of the variables and removes them one at a time until all remaining variables are statistically significant.

In both methods, removed variables may be reentered at a later step if they later appear to be useful predictors, or variables entered early may later be removed if they are no longer significant.

Performing a backward selection results in the following model:

Multiple Regression - 1/MPG City

Dependent variable: 1/MPG City

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	0.0034427	0.00243602	1.41325	0.1610
Horsepower	0.0000260839	0.0000124356	2.09752	0.0388
Weight	0.0000129513	0.0000011041	11.7302	0.0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	0.00696044	2	0.00348022	196.35	0.0000
Residual	0.00159524	90	0.0000177249		
Total (Corr.)	0.00855567	92			

R-squared = 81.3546 percent
R-squared (adjusted for d.f.) = 80.9403 percent
Standard Error of Est. = 0.00421009
Mean absolute error = 0.00313061
Durbin-Watson statistic = 1.62892 (P=0.0338)
Lag 1 residual autocorrelation = 0.184113

The StatAdvisor

The output shows the results of fitting a multiple linear regression model to describe the relationship between 1/MPG City and 6 independent variables. The equation of the fitted model is

1/MPG City = 0.0034427 + 0.0000260839*Horsepower + 0.0000129513*Weight

Since the P-value in the ANOVA table is less than 0.05, there is a statistically significant relationship between the variables at the 95% confidence level.

Figure 13-16. Multiple Regression Analysis Summary after Backward Selection

Only two variables remain in the model: *Horsepower* and *Weight*. Both variables have *P*-values below 0.05.

Once a mathematical equation has been found, it is informative to plot that equation. When the model contains 2 predictor variables, the equation represents a surface in 3 dimensions, usually referred to as a *response surface*. In this case, the fitted equation corresponds to a plane, since *Horsepower* and *Weight* enter the model in a linear manner.

To plot the model, you can:

Use the *Surface and Contour Plots* procedure by copying the function to be plotted and define your own titles and scaling by -

1. If using the Classic menu, select *Plot – Surface and Contour Plots*.

2. If using the Six Sigma menu, select *Tools – Surface and Contour Plots*.

On the data input dialog box, enter the model, expressing the two predictor variables as *X* and *Y*. The easiest way to do this is to paste in the equation generated by the *Multiple Regression* procedure, changing *Horsepower* to X and *Weight* to Y:

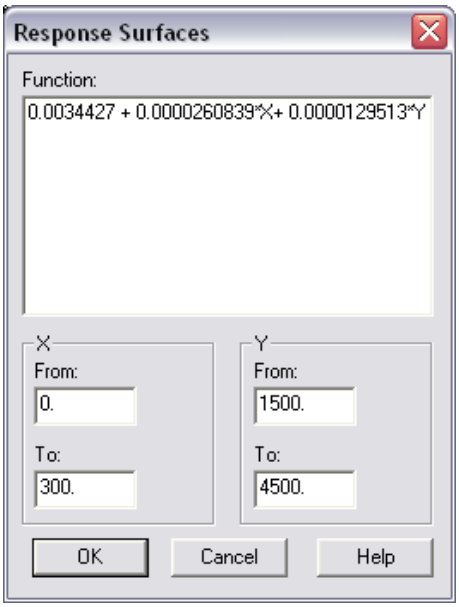


Figure 13-17 Response Surface and Contour Plot Data Input Dialog Box

The scaling of X and Y should also be changed to be representative of the data used to fit the model.

When you press *OK*, the *Tables and Graphs* dialog box will appear. When *OK* is pressed, then a surface plot will be generated. The initial plot takes the form of a wire frame surface:

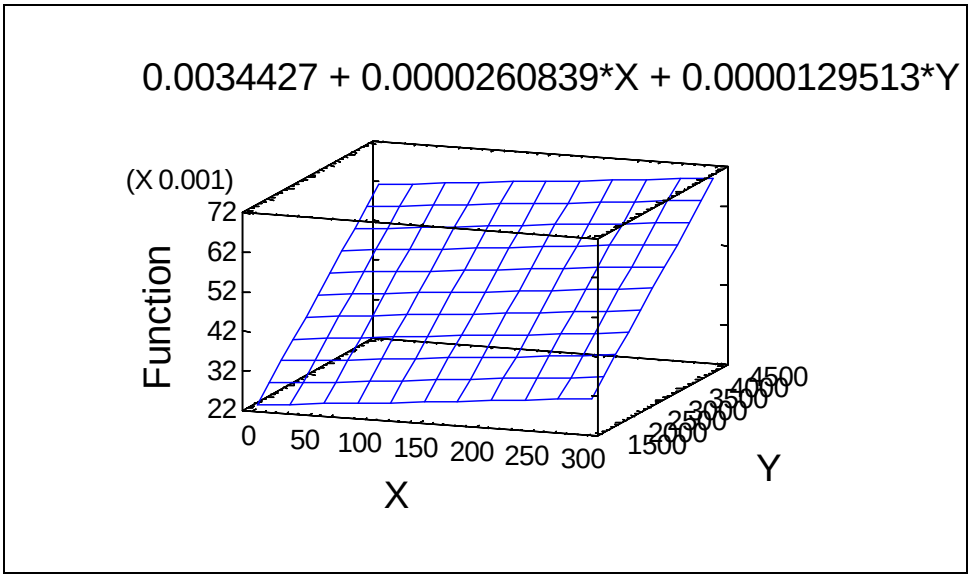


Figure 13-18. Surface Plot with Default Labels and Scaling

You can improve the plot greatly by:

Selecting *Graphics options* from the analysis toolbar and changing the labels and scaling on the *Top Title*, *X-Axis*, *Y-Axis*, and *Z-Axis* tabs. In particular:

- Change the X-axis title to *Horsepower*.
- Change the Y-axis title to *Weight*.
- Change the Y-axis scaling to run from 1500 to 4500 by 1000.
- Change the Z-axis title to *1/MPG City*.

Selecting *Pane Options* and changing the type of plot displayed:

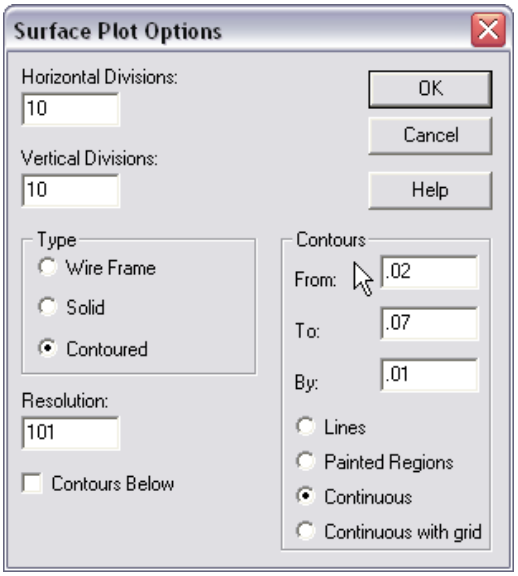


Figure 13-19. Surface Plot Pane Options

In the dialog box above, *Type* has been set to *Contoured* and the *Contour* field to *Continuous*. The final plot is shown below:

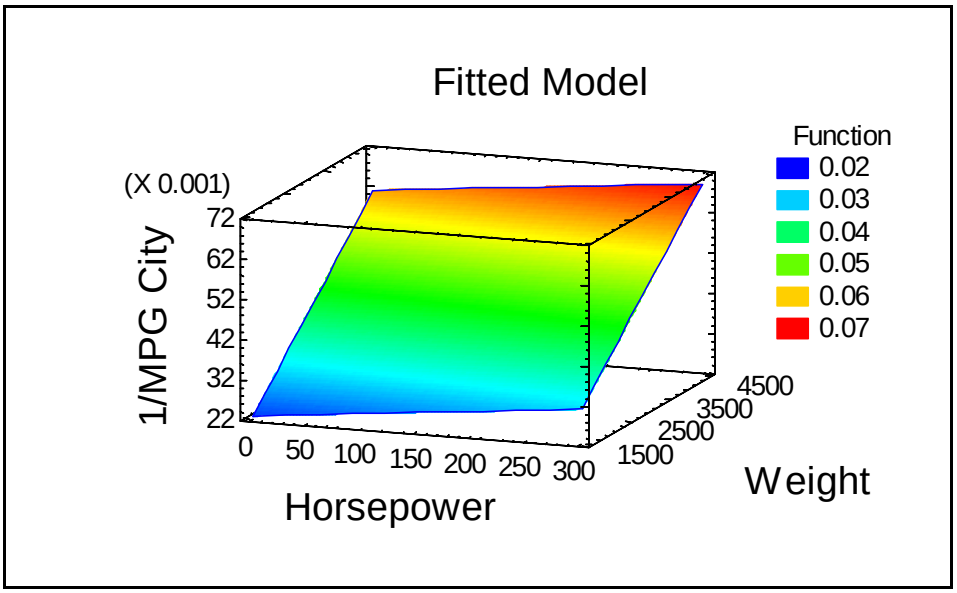


Figure 13-20. Plot of Fitted Model

The cars that use the most fuel are in the back right corner of the plot: big cars with big engines.

Chapter

14

Tutorial #5: Analyzing Attribute Data

Frequency tabulation, contingency tables, and a Pareto analysis.

Each of the first four tutorials deals with variable data, where observations are represented as numbers along a continuous scale. This tutorial examines a set of attribute data, in which each observation represents a category into which an attribute has been classified, rather than a measurement.

As an example, consider the data contained in the file *defects.sgd*. A portion of that file is shown below:

Defect	*Facility*
Misaligned	Virginia
Contaminated	Texas
Contaminated	Virginia
Contaminated	Texas
Missing parts	Texas
Misaligned	Virginia
Contaminated	Texas
Leaking	Texas
Damaged	Virginia
Contaminated	Texas

The data consist of $n = 120$ rows, each corresponding to a defect that was observed in a manufactured item. The file also indicates the type of defect and the facility in which the item was produced.

14.1 Summarizing Attribute Data

Ignoring for a moment the facility in which each item was produced, the data on defect type may be summarized by:

1. If using the Classic menu, select *Describe – Categorical Data – Tabulation*.

2. If using the Six Sigma menu, select *Analyze – Attribute Data – One Factor - Tabulation*.

The data input dialog box expects a single column containing the attribute data:

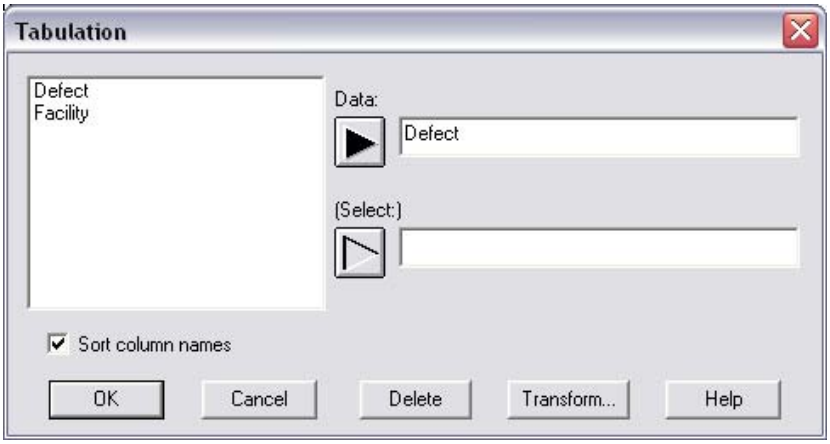
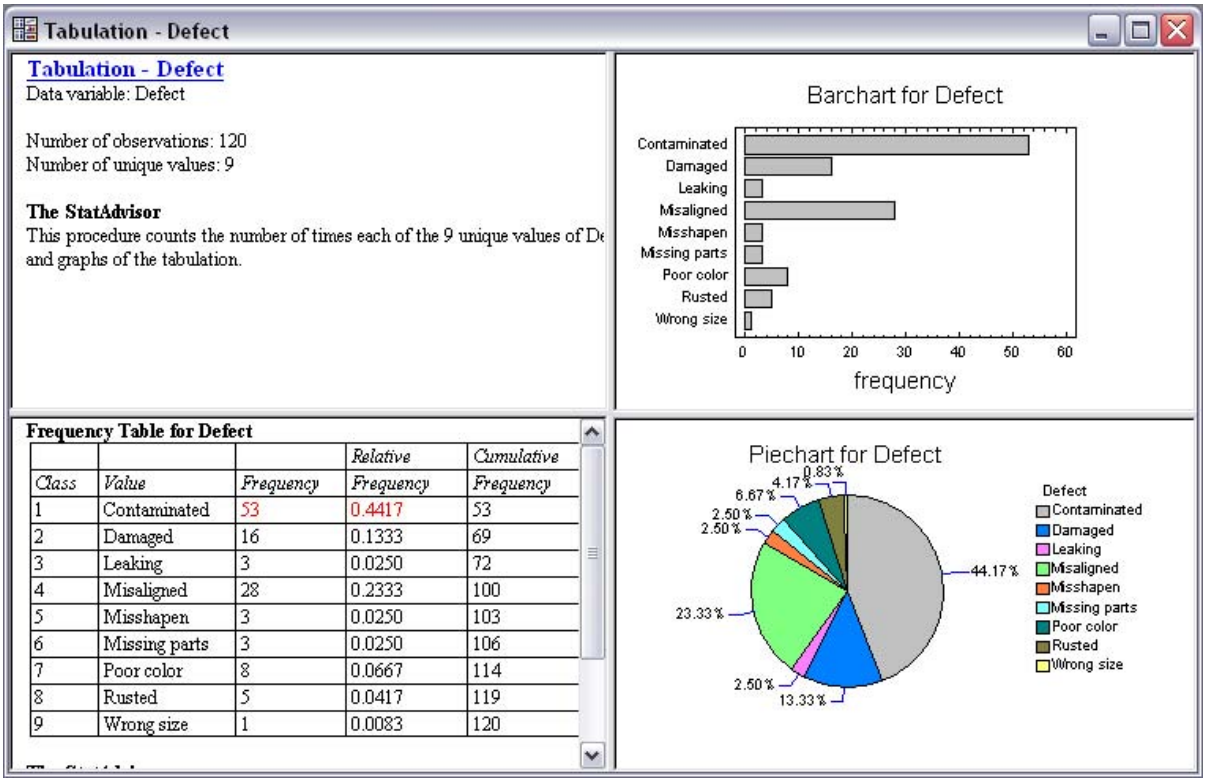


Figure 14-1. Tabulation Data Input Dialog Box

The procedure scans the column, identifying each unique value. The *Tables and Graphs* dialog box will appear, then an analysis window will be generated similar to that shown below:



Class	Value	Frequency	Relative Frequency	Cumulative Frequency
1	Contaminated	53	0.4417	53
2	Damaged	16	0.1333	69
3	Leaking	3	0.0250	72
4	Misaligned	28	0.2333	100
5	Misshapen	3	0.0250	103
6	Missing parts	3	0.0250	106
7	Poor color	8	0.0667	114
8	Rusted	5	0.0417	119
9	Wrong size	1	0.0083	120

Figure 14-2. Tabulation Analysis Window

The upper left pane shows that 9 unique values were found in the $n = 120$ rows. The barchart and piechart on the right illustrate the observed frequency of each type of defect, which is also tabulated in the bottom left pane. The most common type of defect is "Contaminated", which represents about 44% of all defects.

14.2 Pareto Analysis

The *Frequency Tabulation* procedure orders the types of defects in alphabetical order. To order the types from most frequent to least frequent, use the *Pareto Analysis* procedure instead. The Pareto analysis is accessed by:

1. If using the Classic menu, select *SPC – Quality Assessment - Pareto Analysis*.

2. If using the Six Sigma menu, select *Analyze – Attribute Data – One Factor – Pareto Analysis*.

The data input dialog box should be completed as shown below:

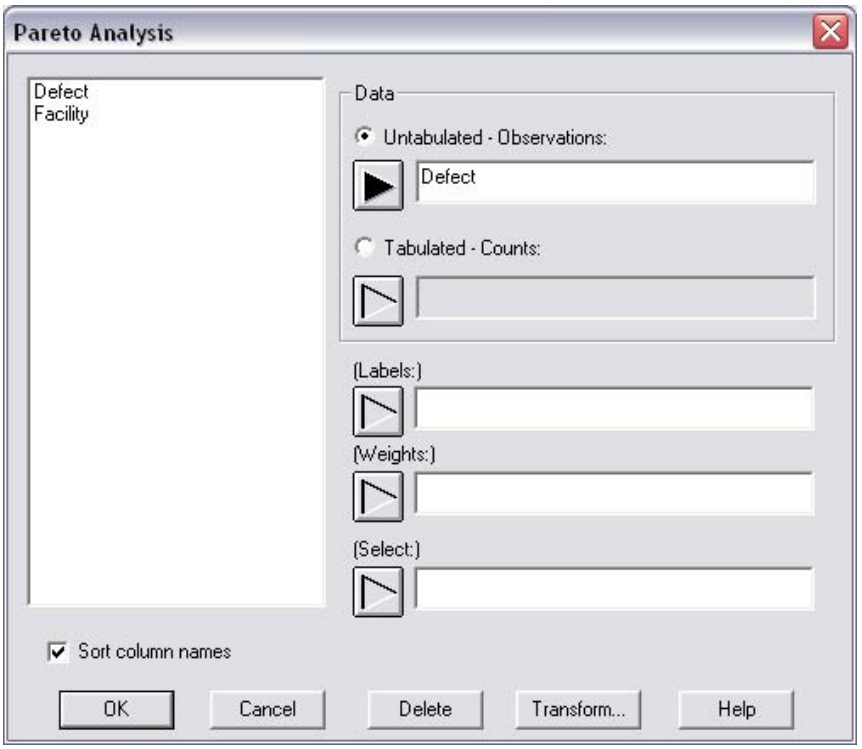


Figure 14-3. Pareto Analysis Data Input Dialog Box

The *Pareto Analysis* accepts data in two formats:

1. Untabulated data that need to be counted, as in the current example.

2. Counts for data that have already been grouped by defect type. This is applicable if you have two columns, one identifying the types of defects and a second containing the number of times each defect type occurred.

The analysis window displays both a summary table and a Pareto chart:

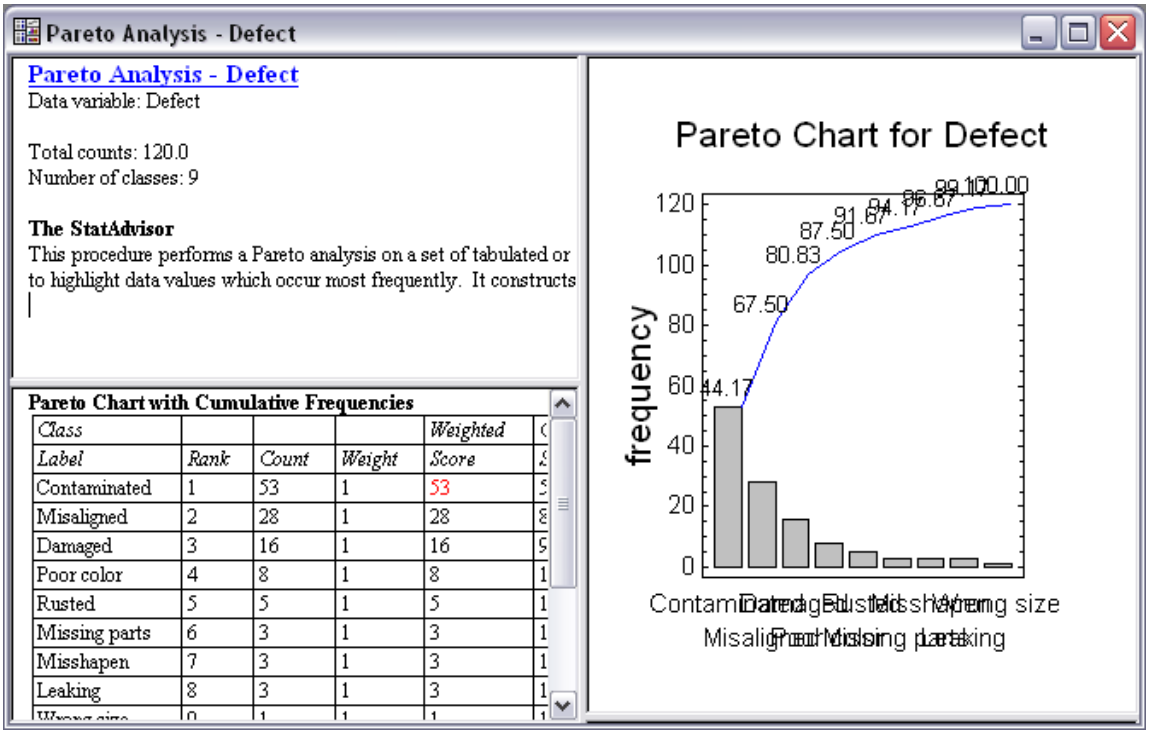


Figure 14-4. Pareto Analysis Window

Of particular interest is the Pareto chart on the right, which plots the frequencies of each type of defect from most common to least common. Initially, the bar labels overlap badly due to their number and length. This may be fixed by:

1. Double-clicking on the graph with your mouse to maximize the pane within the analysis window.

2. Pressing the *Graphics options* button on the analysis toolbar, clicking on the *X-Axis* tab, and checking the *Rotate Axis Labels* box.

3. After exiting the *Graphics options* dialog box, the labels may not fit completely on the screen. If not, you can hold your mouse button down within the main part of the graph and drag it higher, or you can drag the X-axis up to reduce the size of the vertical axis.

When finished, the Pareto chart should look like that shown below:

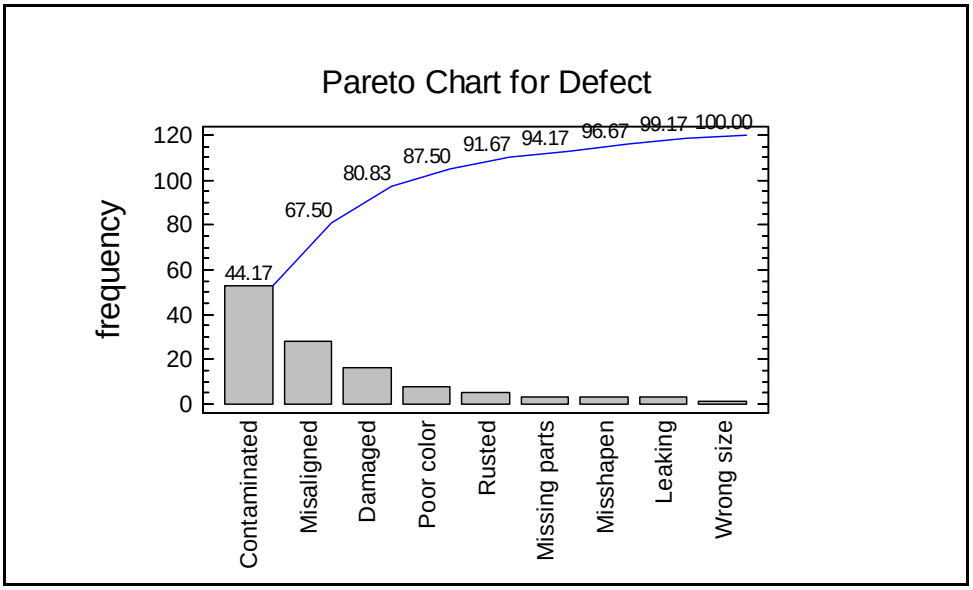


Figure 14-5. Enlarged Pareto Chart

The vertical bars in the Pareto chart are drawn with height proportional to the number of times each defect type occurred. The line above the bars is a cumulative count from left to right. Shown above each bar is the percentage of defects occurring in a particular class or classes farther to the left.

The basic Pareto principle states that a large majority of defects are usually due to a small number of possible causes. In this case, the 3 most frequent defect types account for over 80% of all the defects.

14.3 Crosstabulation

The *defects.sgd* data file also contains an identification of which facility produced each defective item. To summarize the data by both defect type and facility:

1. If using the Classic menu, select *Describe – Categorical Data – Crosstabulation.*

2. If using the Six Sigma menu, select *Analyze – Attribute Data – Multiple Factors - Crosstabulation.*

The data input dialog box expects two columns, one defining the rows of a two-way frequency or *contingency table* and the other defining the columns:

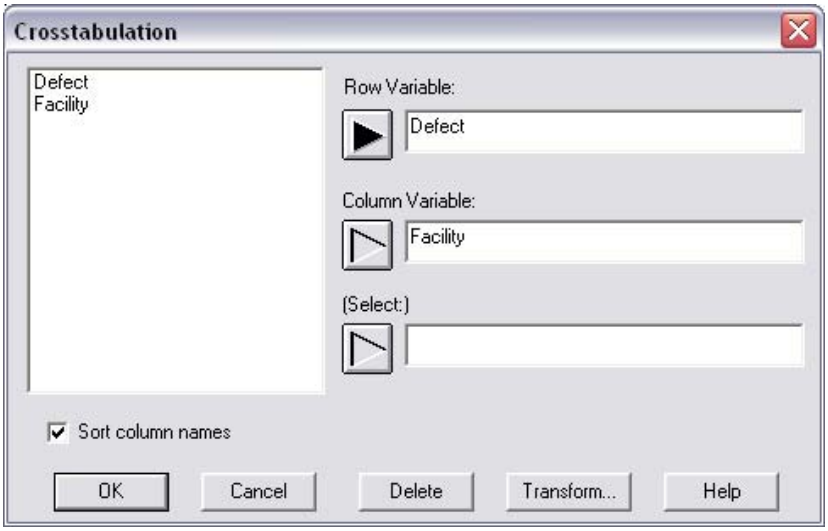
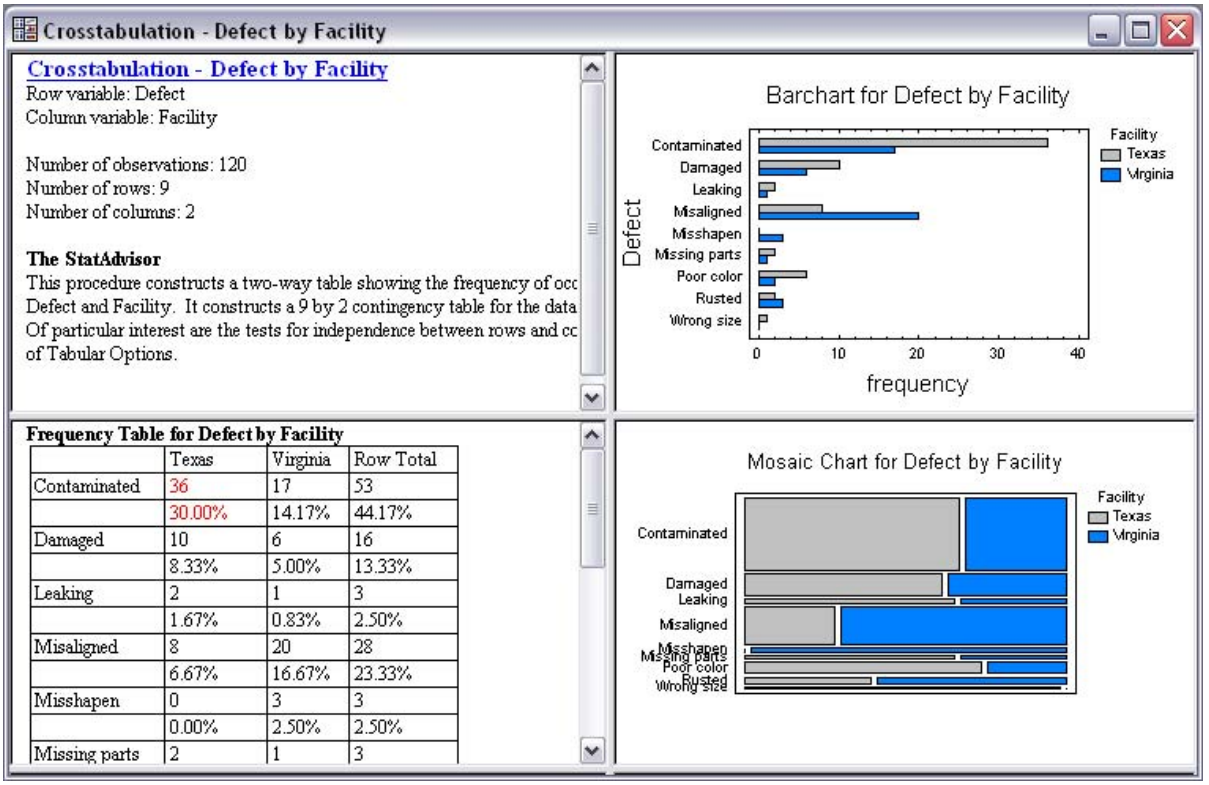


Figure 14-6. Crosstabulation Data Input Dialog Box

After the *Options* and *Tables and Graphs* dialog boxes, the following analysis window is generated:



	Texas	Virginia	Row Total
Contaminated	36	17	53
	30.00%	14.17%	44.17%
Damaged	10	6	16
	8.33%	5.00%	13.33%
Leaking	2	1	3
	1.67%	0.83%	2.50%
Misaligned	8	20	28
	6.67%	16.67%	23.33%
Misshapen	0	3	3
	0.00%	2.50%	2.50%
Missing parts	2	1	3

Figure 14-7. Crosstabulation Analysis Window

The table in the bottom left pane tabulates the data by both defect type and facility:

Frequency Table for Defect by Facility

	Texas	Virginia	Row Total
Contaminated	36	17	53
	30.00%	14.17%	44.17%
Damaged	10	6	16
	8.33%	5.00%	13.33%
Leaking	2	1	3
	1.67%	0.83%	2.50%
Misaligned	8	20	28
	6.67%	16.67%	23.33%
Misshapen	0	3	3
	0.00%	2.50%	2.50%
Missing parts	2	1	3
	1.67%	0.83%	2.50%
Poor color	6	2	8
	5.00%	1.67%	6.67%
Rusted	2	3	5
	1.67%	2.50%	4.17%
Wrong size	1	0	1
	0.83%	0.00%	0.83%
Column Total	67	53	120
	55.83%	44.17%	100.00%

Cell contents:
Observed frequency
Percentage of table

Figure 14-8 Two-Way Table with Table Percentages

As initially displayed, each cell of the table displays the number of rows in the data file corresponding to a particular row-column combination. It also indicates the percentage of the entire table represented by that cell. For example, there were 36 contaminated items produced in the Texas facility, representing 30 percent of all defective items in the sample.

Pane Options allows you to select other items to display in each cell:

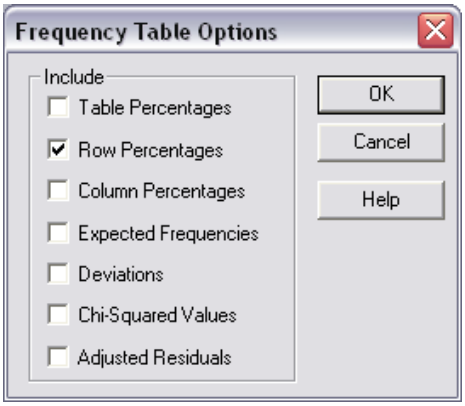


Figure 14-9 Pane Options Dialog Box for Crosstabulation

An interesting choice for the current data is to display *Row Percentages* rather than *Table Percentages*:

Frequency Table for Defect by Facility

	Texas	Virginia	Row Total
Contaminated	36	17	53
	67.92%	32.08%	44.17%
Damaged	10	6	16
	62.50%	37.50%	13.33%
Leaking	2	1	3
	66.67%	33.33%	2.50%
Misaligned	8	20	28
	28.57%	71.43%	23.33%
Misshapen	0	3	3
	0.00%	100.00%	2.50%
Missing parts	2	1	3
	66.67%	33.33%	2.50%
Poor color	6	2	8
	75.00%	25.00%	6.67%
Rusted	2	3	5
	40.00%	60.00%	4.17%
Wrong size	1	0	1
	100.00%	0.00%	0.83%
Column Total	67	53	120
	55.83%	44.17%	100.00%

Cell contents:
Observed frequency
Percentage of row

Figure 14-10 Two-Way Table with Row Percentages

The tabled percentage now indicates the percentage that each cell represents of its row. For example, 67.92% of all contaminated items were produced in Texas, while 71.43% of all

misaligned items were produced in Virginia. This suggests that some defect types may occur more frequently in one facility than another, a hypothesis that will be tested formally in the following section.

Various graphical displays are also helpful. For example, the barchart shows the data by both defect and facility:

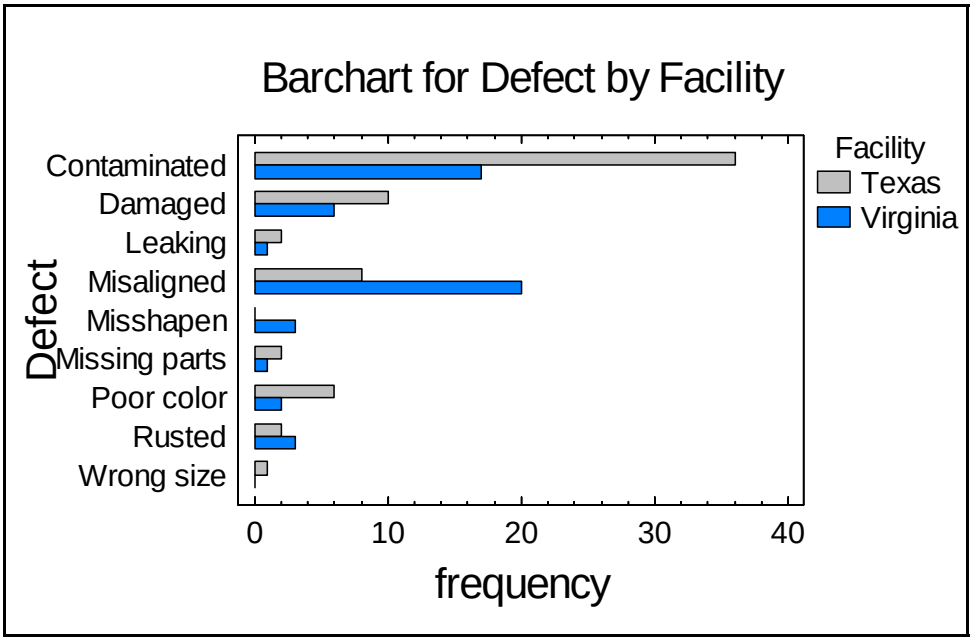


Figure 14-11. Clustered Barchart

The difference between the two facilities is quite apparent. A related plot, called a *Mosaic Plot*, is also quite informative:

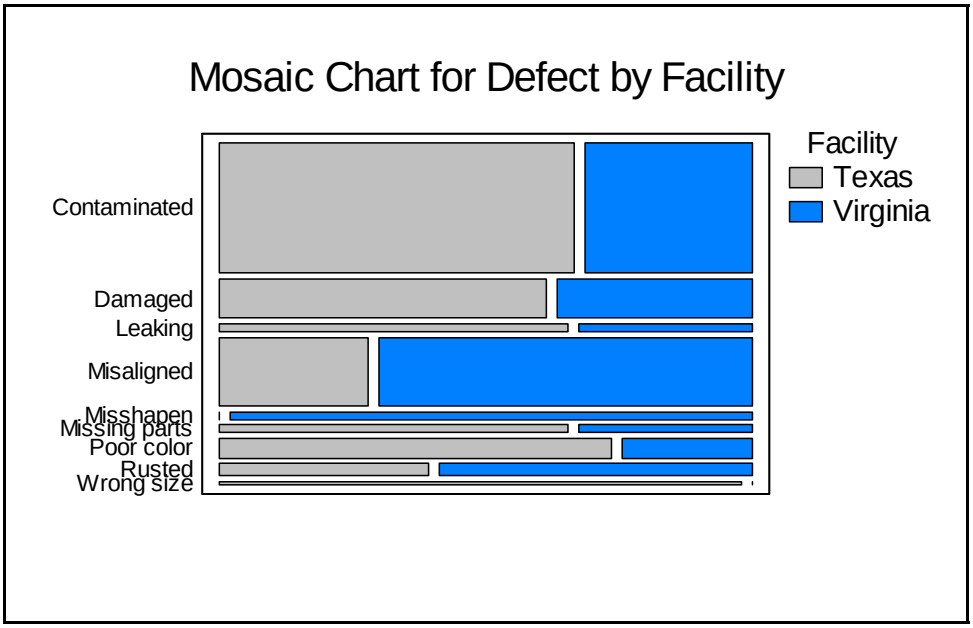


Figure 14-12. Mosaic Plot

In this chart, the height of each bar is proportional to the total number of defects of each type. The width of the bars is proportional to the relative percentage of each defect type at each location. Consequently, the overall area of each rectangle is proportional to the frequency of the corresponding cell in the two-way table.

If desired, the cell frequencies may also be displayed in three dimensions by selecting *Skychart* from the *Tables and Graphs* dialog box:

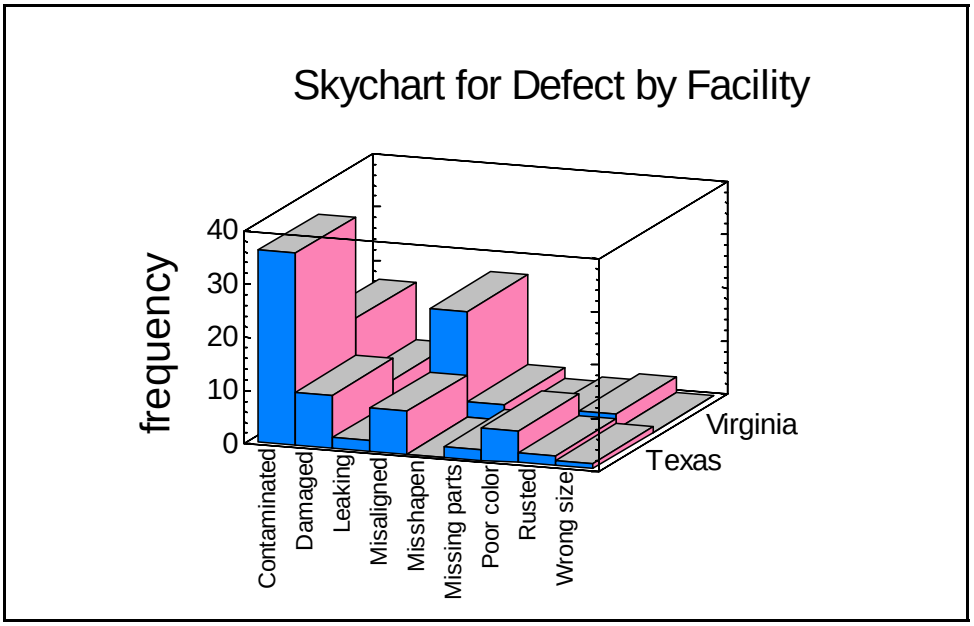


Figure 14-13. Three-Dimensional Skychart

In a *Skychart*, the height of each bar represents the frequency of a cell in the contingency table.

14.4 Comparing Two or More Samples

To determine whether or not the apparent differences between the Texas and Virginia facilities are statistically significant, select *Tests of Independence* from the *Tables and Graphs* dialog box. For a table of this size, the procedure displays the results of a chi-square test:

Tests of Independence

Test	*Statistic*	*Df*	*P-Value*
Chi-Square	18.438	8	0.0182

Warning: some cell counts < 5.

Figure 14-14. Chi-Square Test of Independence

The chi-square test is used to decide between two hypotheses:

Null hypothesis: row and column classifications are independent.

Alternative hypothesis: row and column classifications are not independent.

Independence would imply that the type of defect found in an item had nothing to do with the facility in which that item was manufactured.

For the chi-square test, a small *P*-value indicates that the row and column classifications are not independent. In this case, the *P*-value is less than 0.05, indicating at the 5% level of significance that the distribution of defect types is different in the Texas facility than in the Virginia facility.

A warning is also displayed; however, since some cell counts in the two-way table are less than 5. (Technically, the warning occurs if the expected count in any cell is less than 5 assuming that the null hypothesis is true). With small cell counts, the *P*-value may be unreliable. One solution to this problem is to group all infrequent defect types into a single class and rerun the test. This is easily done in STATGRAPHICS Centurion XVI in the following way:

1. Return to the datasheet and click on the header of the *Defects* column to select it.

2. Press the alternate mouse button and select *Recode Data* from the popup menu.

3. Complete the *Recode Data* dialog box as shown below to combine the less common defect types into a single class labeled "Other":

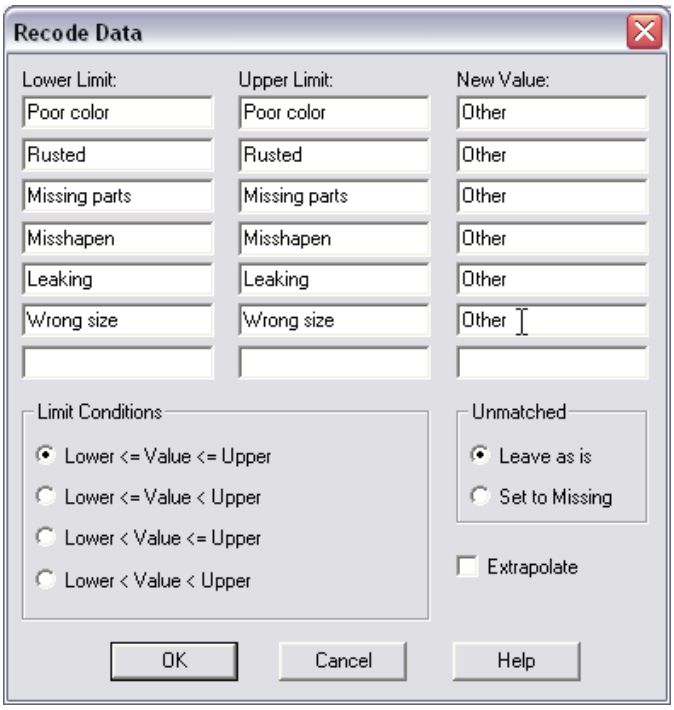


Figure 14-15. Recoding Least Frequent Defect Types

The entries on the *Recode Data* dialog box instruct the program to search for values in the *Defects* column falling within each defined interval. Any label falling alphabetically between the limits shown in a given row is recoded to the value specified in the *New Value* column.

After performing the recode operation, return to the *Crosstabulation* analysis window. In response to the change in the datasheet, the analysis will have automatically been updated. The new *Other* class now has a reasonably high frequency, as shown in the revised *Mosaic Plot*:

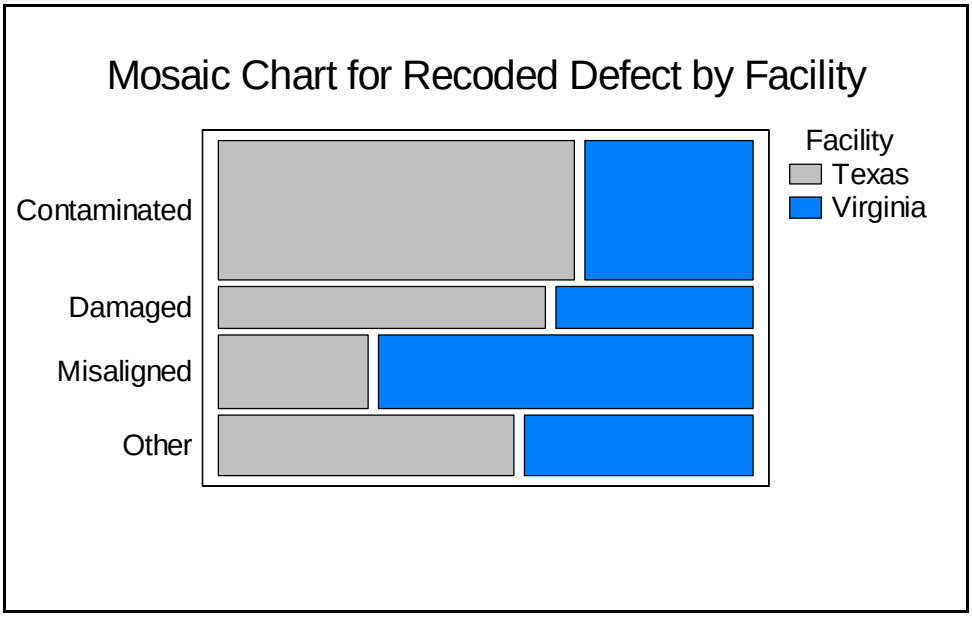


Figure 14-16. Mosaic Plot for Recoded Data

After recoding, the chi-square test still shows a statistically significant difference between the Texas and Virginia facilities:

Tests of Independence

Test	*Statistic*	*Df*	*P-Value*
Chi-Squared	11.874	3	0.0078

The StatAdvisor

This table shows the results of a hypothesis test run to determine whether or not to reject the idea that the row and column classifications are independent. Since the P-value is less than 0.05, we can reject the hypothesis that rows and columns are independent at the 95% confidence level. Therefore, the observed value of Recoded Defect for a particular case is related to its value for Facility.

Figure 14-17. Chi-Square Test after Recoding Data

It thus appears that defect type is indeed related to the facility in which an item was produced.

It should be noted that the above test compares the *distribution* of defect types between the two facilities. It does not compare the numbers or percentages of defective items at each location. Such a comparison requires a different test, as explained in the next section.

14.5 Contingency Tables

To determine whether one facility produces more defective items than another, we need to know the total production at each facility. Suppose the following describes one month's production:

Facility	*Number of Defects*	*Number of Items Produced*
Texas	67	6,237
Virginia	53	7,343

Let θ_1 be the proportion of defective items produced in Texas. Let θ_2 be the proportion of defective items produced in Virginia. The estimated proportions are given by:

$$\hat{\theta}_1 = \frac{67}{6237} = 0.0107 \qquad \hat{\theta}_2 = \frac{53}{7343} = 0.0072$$

Based on this data, it appears that the percentage of defective items produced in Texas may be greater than the percentage of defective items produced in Virginia. To determine whether this apparent difference is statistically significant, create a datasheet as shown below:

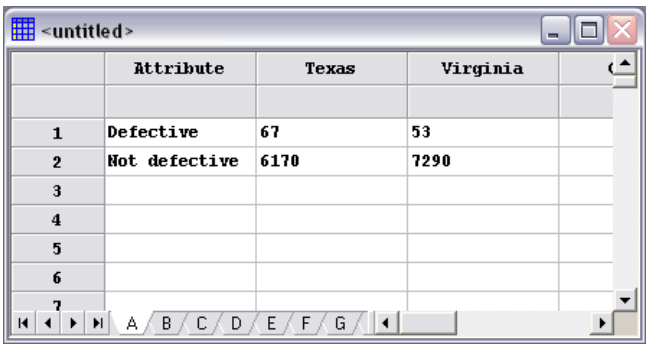


Figure 14-18. Datasheet for Comparing Two Proportions

The rows hold counts of defective and non-defective items. Then select *Contingency Tables* from the same menu as *Crosstabulation*. Enter:

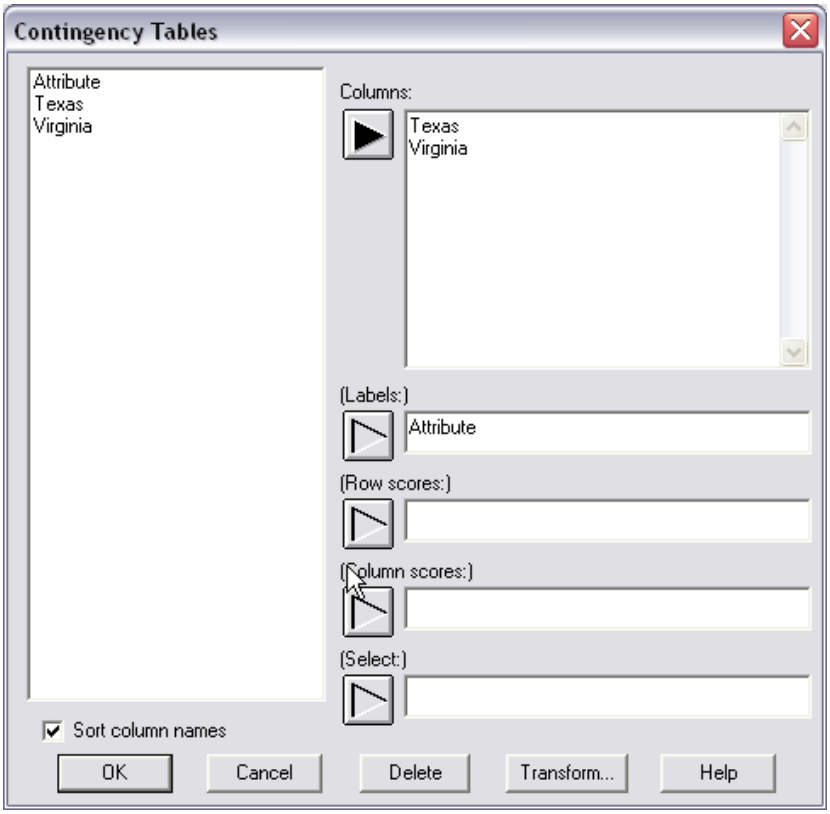


Figure 14-19. Contingency Tables Data Input Dialog Box

The analysis will display a chi-square test of the 2 by 2 table:

Tests of Independence

Test	*Statistic*	*Df*	*P-Value*
Chi-Squared	4.783	1	0.0287

Figure 14-20. Chi-Square test of the 2 by 2 Table

Recall that the chi-square test determines whether or not row and column classifications are independent. In this case, independence would imply than whether an item was defective or not had nothing to do with the facility in which it was produced.

Since the *P*-value in the above table is less than 0.05, the hypothesis of independence is rejected at the 5% significance level. We can thus conclude that the proportions of defectives at the two facilities are significantly different.

Chapter

15

Tutorial #6: Process Capability Analysis

Determining the Defects Per Million or percent beyond the specification limits.

STATGRAPHICS Centurion XVI is widely used by individuals whose job it is to ensure that the products and services they provide are of the highest quality. A common task in such a job is to collect data from the process and compare it to established specification limits. The output from this type of *capability analysis* is an estimate of how capable the process is of meeting those specifications. Six Sigma, which is a widely practiced methodology for achieving world class quality, targets a defect rate of 3.4 defects per million opportunities.

As an example, consider a product whose strength is required to fall between 190 and 230 psi. Suppose that $n = 100$ samples are taken from the manufacturing process and their strength measured, as shown in the following table:

213.5	203.3	191.3	197.1	205.7	215.6	193.7	201.7	201.5	207.1
207.0	200.4	197.2	202.4	205.2	211.0	214.5	201.5	200.9	206.8
205.8	200.3	196.1	205.9	195.1	203.9	192.9	199.0	195.5	203.1
197.4	194.8	201.0	202.5	199.0	200.7	197.6	198.5	205.3	197.1
202.8	201.6	197.4	200.9	203.3	209.4	201.4	199.5	207.8	204.9
205.5	203.0	208.1	200.2	218.2	202.0	209.3	201.2	200.4	201.0
195.7	229.5	199.9	208.1	210.3	202.0	202.6	213.6	198.0	197.8
196.7	216.0	211.6	208.7	199.4	200.8	201.1	195.3	206.8	211.3
201.5	200.0	211.8	195.6	201.9	199.0	200.3	197.8	200.8	194.8
199.5	195.5	201.0	206.0	215.3	202.6	199.9	200.6	197.6	207.4

This chapter describes how to conduct a typical capability analysis for this type of variable data.

15.1 Plotting the Data

The first step in examining any new set of data is to plot it. For a set of data such as that shown above, the *One-Variable Analysis* described in Chapter 10 provides several useful tools. To analyze this data:

1. Open the file called *items.sgd*.

2. Execute the *One-Variable Analysis* procedure using the column named *Strength*.

The initial analysis window is shown below:

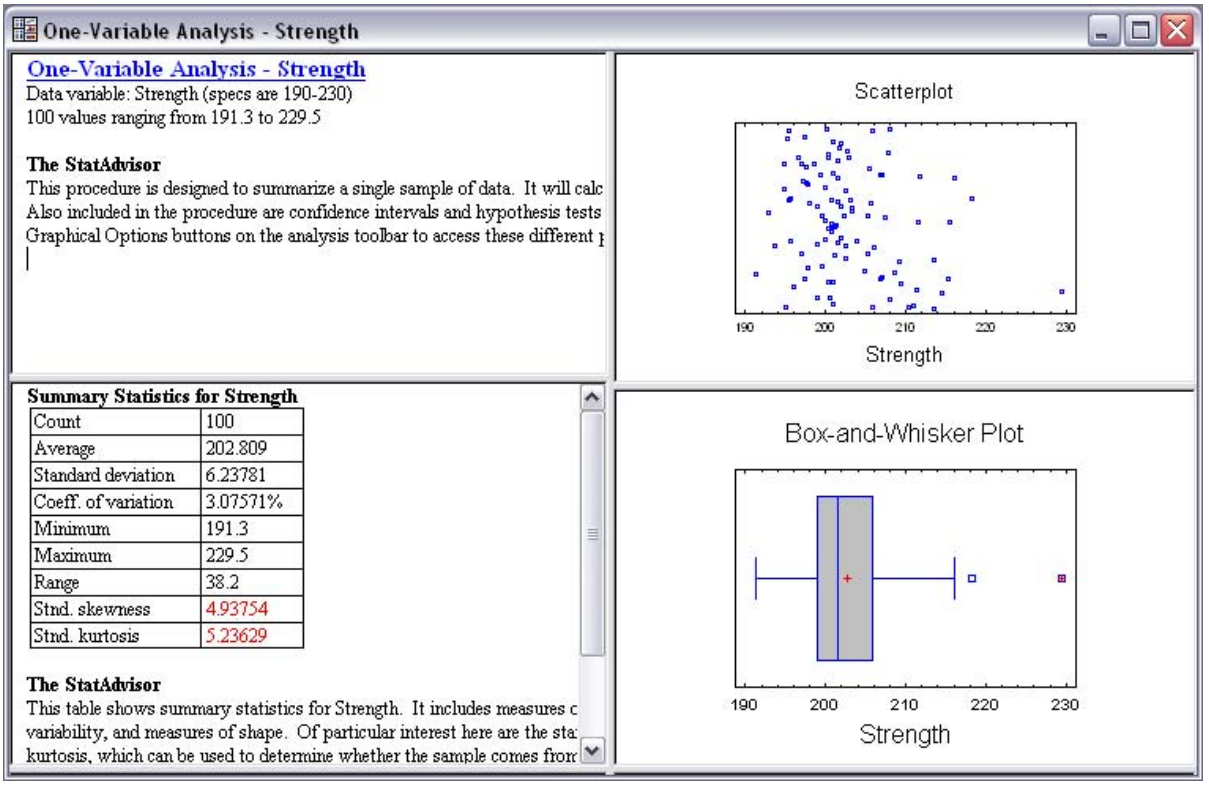


Figure 15-1. One-Variable Analysis Window

Several interesting factors are immediately evident:

1. The data are all within the specification limits, but just barely, ranging from 191.3 to 229.5.

2. The box-and-whisker plot shows a far outside point (a small square with a red plus sign drawn through it). Such points are often considered to be outliers, if the rest of the data appear to come from a normal distribution. In this case, however, even discounting that apparent outlier, the shape of the box is not very symmetric. The upper whisker is longer than the lower whisker, and the box extends farther above the median (the vertical line within the box) than it does below.

3. If you expand the *Summary Statistics* pane, you will see that the standardized skewness equals 4.94. If the data came from a normal distribution, this value should lie between -2 and +2. Even removing the largest data value only reduces the standardized skewness to 2.81.

A frequency histogram may also be displayed by pressing the *Tables and Graphs* button on the analysis toolbar and selecting *Frequency Histogram* on the *Graphs* dialog box:

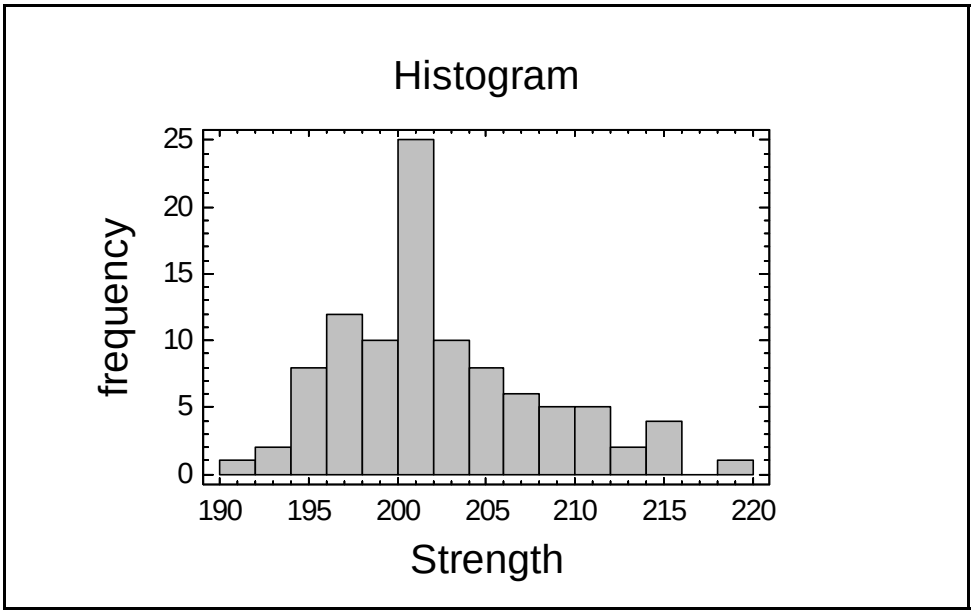


Figure 15-2. Frequency Histogram

The data appear quite clearly to be positively skewed, extending farther to the right of the peak than to the left.

Non-normal data such as that shown above are commonplace. One typical approach to dealing with such data, unfortunately, is to simply ignore the non-normality and calculate indices such as C_{pk} using formulas designed for data from a normal distribution. As will be seen in this tutorial, ignoring non-normality can lead to incorrect results, often significantly overestimating or underestimating the percent of the product that is beyond the specification limits.

15.2 Capability Analysis Procedure

STATGRAPHICS Centurion XVI contains procedures for performing a capability analysis on data collected either one at a time (individuals data) or in subgroups (such as 5 observations every hour). Assuming the sample data are individuals, a process capability analysis may be conducted by:

1. If using the Classic menu, selecting *SPC – Capability Analysis – Variables – Individuals*.

2. If using the Six Sigma menu, selecting *Analyze – Variable Data – Capability Analysis – Individuals*.

The data input dialog box requests the name of a single column containing the data. The sample data may be found in a column called *Strength* in the file named *items.sgd*:

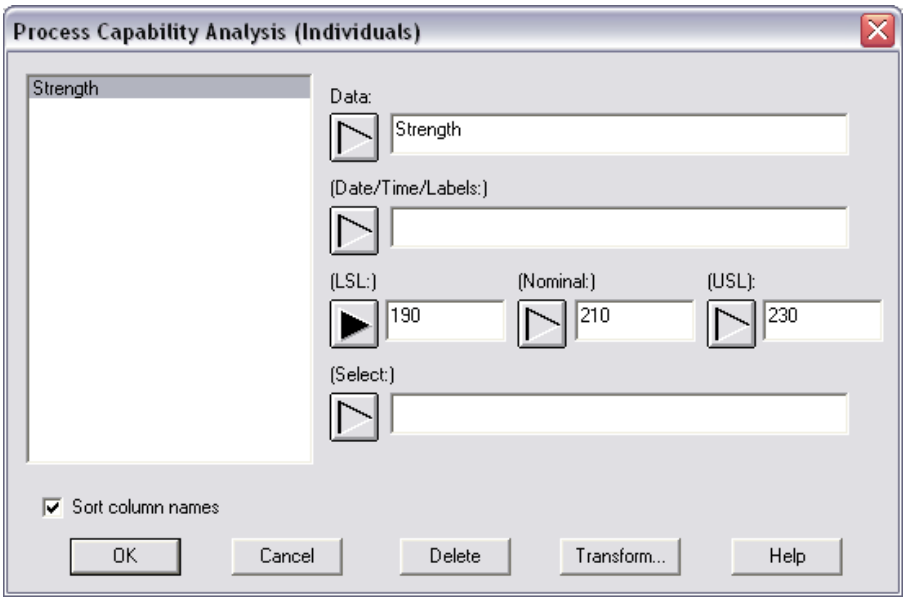


Figure 15-3. Process Capability Analysis Dialog Box

Upper and lower specification limits have also been indicated, as has a nominal or target value.

When *OK* is pressed, the *Options* menu, then the *Tables and Graphs* dialog box will appear. Use the default values for both menus for the sake of this tutorial.

The initial analysis window shows a summary of the data, a table of capability indices, and a capability plot:

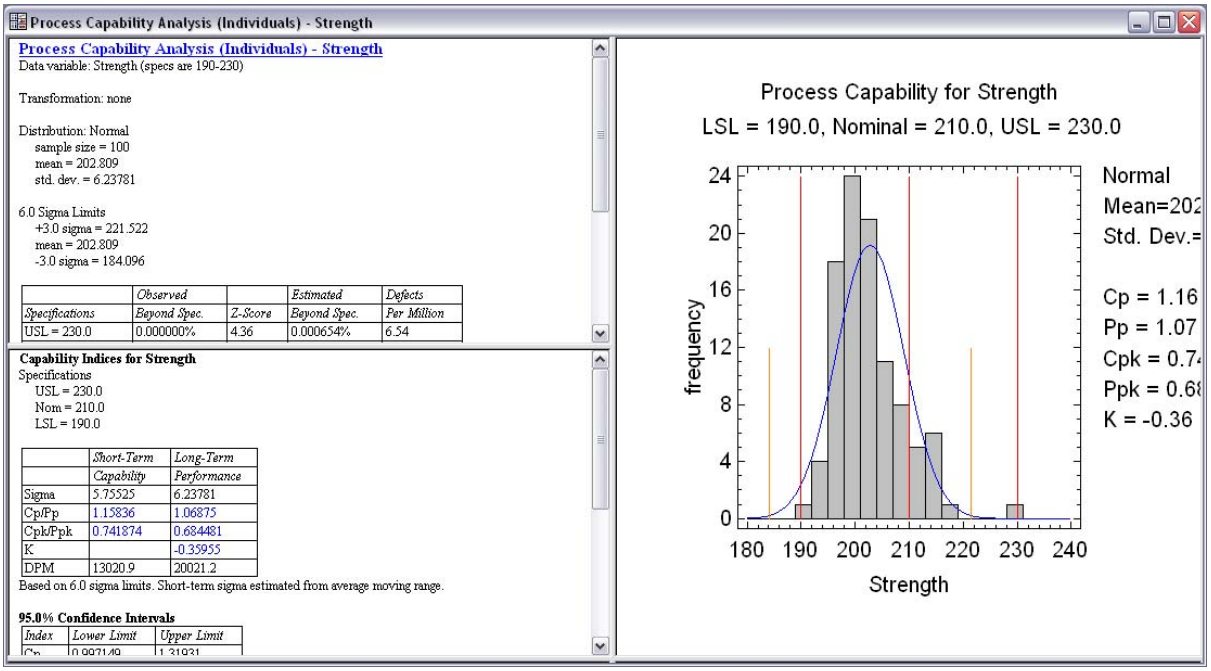


Figure 15-4. Process Capability Analysis Window

When a capability analysis is first run, a normal distribution is fit to the data. The *Capability Plot* shows a histogram of the data, together with the best fitting normal distribution:

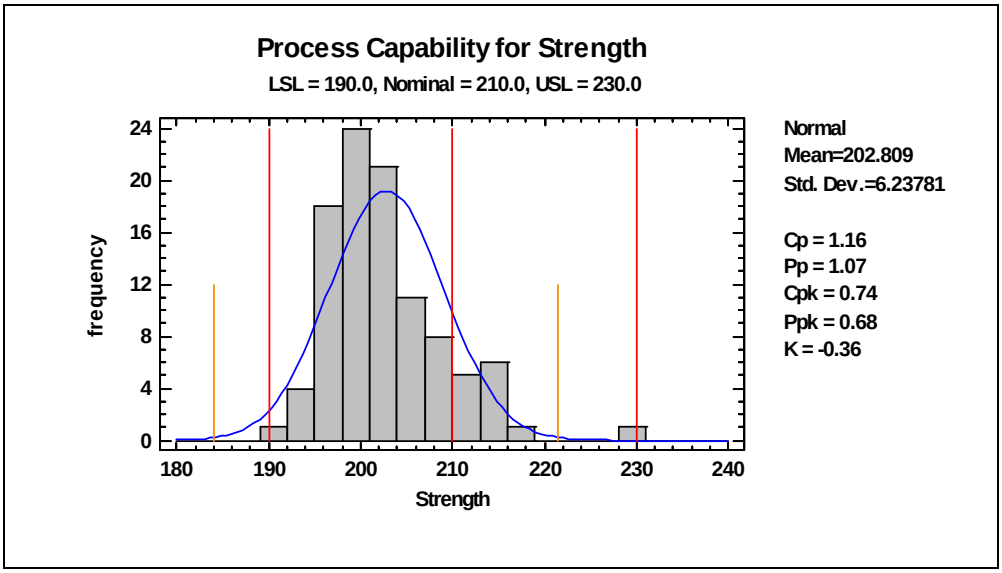


Figure 15-5. Capability Plot with Normal Distribution

The tall vertical lines in the plot show the location of the specification limits and the nominal value. The shorter vertical lines are located at the sample mean plus and minus 3 standard deviations. Particularly notable in the plot above are:

1. The fitted normal distribution does not match the data very well. Although the normal bell-shaped curve has the same mean and standard deviation as the data, the skewness in the data cause the curve to do a poor job in matching the bars of the histogram.

2. The sample mean is located at 202.8, which is considerably less than the nominal value of 210.

3. Although none of the observations are less than the lower specification limit, a fair amount of the lower tail of the normal distribution is below that limit.

4. The lines at plus and minus 3 sigma are tight enough to fit within the specs. However, they are shifted to the left.

The *Analysis Summary* in the upper left pane quantifies the fit:

Process Capability Analysis (Individuals) - Strength
Data variable: Strength (specs are 190-230)
Transformation: none

Distribution: Normal
sample size = 100
mean = 202.809
std. dev. = 6.23781

6.0 Sigma Limits
+3.0 sigma = 221.522
mean = 202.809
-3.0 sigma = 184.096

	Observed		*Estimated*	*Defects*
Specifications	*Beyond Spec.*	*Z-Score*	*Beyond Spec.*	*Per Million*
USL = 230.0	0.000000%	4.36	0.000654%	6.54
Nominal = 210.0		1.15		
LSL = 190.0	0.000000%	-2.05	2.001465%	20014.65
Total	0.000000%		2.002119%	20021.19

Figure 15-6. Capability Analysis Summary

Of primary interest is the lower table, which estimates the percent of the product that is likely to be out of specification Based on the fitted normal distribution, the estimated percent of the product beyond the specification limits is about 2%, equal to 20,021 defects per million (DPM).

15.3 Dealing with Non-Normal Data

The estimated DPM calculated above relies heavily on the assumption that the data come from a normal distribution. A formal check of that hypothesis may be conducted by selecting *Tests for Normality* from the *Tables and Graphs* dialog box:

Tests for Normality for Strength

Test	*Statistic*	*P-Value*
Shapiro-Wilks W	0.931784	0.0000321356

Figure 15-7. Tests for Normality

Depending on your system preferences, one or more tests for normality will be displayed. Each of the available tests is based on the following set of hypotheses:

Null hypothesis: the data come from a normal distribution.

Alternative hypothesis: the data do not come from a normal distribution.

A *P*-value below 0.05 leads to rejection of the hypothesis of normality at the 5% significance level.

In the table above, the Shapiro-Wilks test soundly rejects the hypothesis that the data come from a normal distribution. Therefore, any estimated DPM values or capability indices based on the assumption of normality are not valid.

When the data are non-normal, one of two possible approaches may be followed:

1. Select a distribution other than the normal on which to base the analysis.

2. Transform the data so that it follows a normal distribution in the transformed metric.

To assist in selecting a different distribution, STATGRAPHICS Centurion XVI provides an option called *Comparison of Alternative Distributions* on the *Tables and Graphs* dialog box. This option fits several other distributions and lists them in order of their goodness-of-fit. Using the default selection of distributions yields the following output:

Comparison of Alternative Distributions

Distribution	*Est. Parameters*	*KS D*	*A^2*
Largest Extreme Value	2	0.0675422	0.372613
Loglogistic	2	0.0913779	1.15081
Laplace	2	0.0920985	1.68399
Logistic	2	0.0941708	1.27599
Lognormal	2	0.13213	1.66564
Gamma	2	0.134136	
Normal	2	0.138628	1.90094
Weibull	2	0.177886	5.67166
Smallest Extreme Value	2	0.189989	6.28546
Exponential	1	0.61064	43.3327
Pareto	1	0.628084	45.3859

Figure 15-8. Fitted Distributions in Order of Goodness-of-Fit

The distributions have been listed according to the value of the Kolmogorov-Smirnov goodness-of-fit statistic, which measures the maximum distance between the cumulative distribution of the data and that of the fitted distribution. In this case, the best fitting distribution is the *largest extreme value* distribution.

You can switch to the largest extreme value distribution by accessing *Analysis Options*:

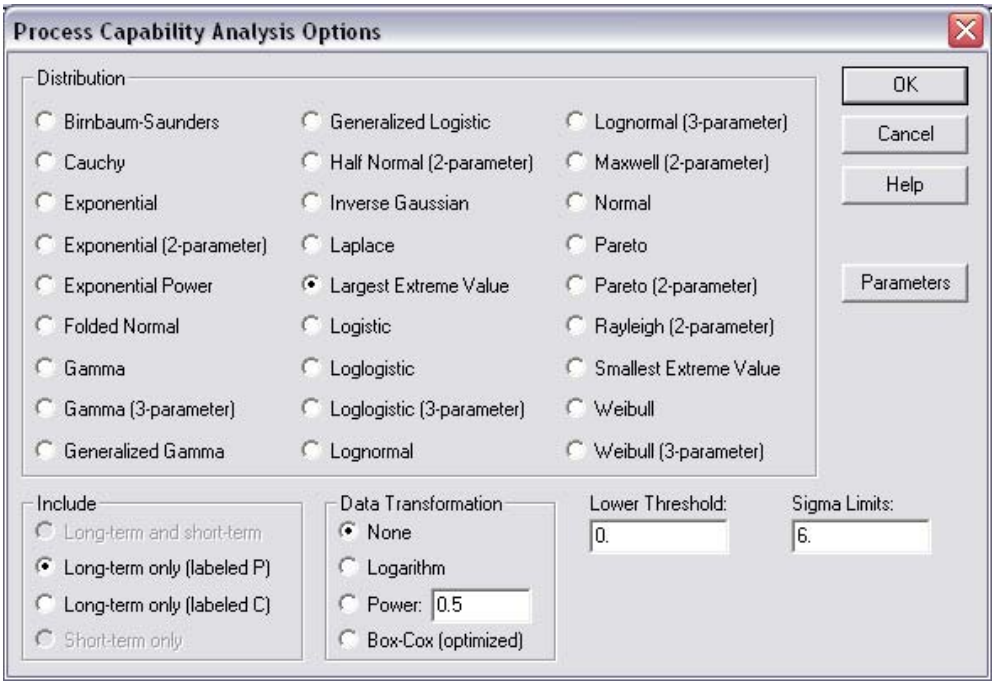


Figure 15-9. Process Capability Analysis Options Dialog Box

The resulting fit is shown below:

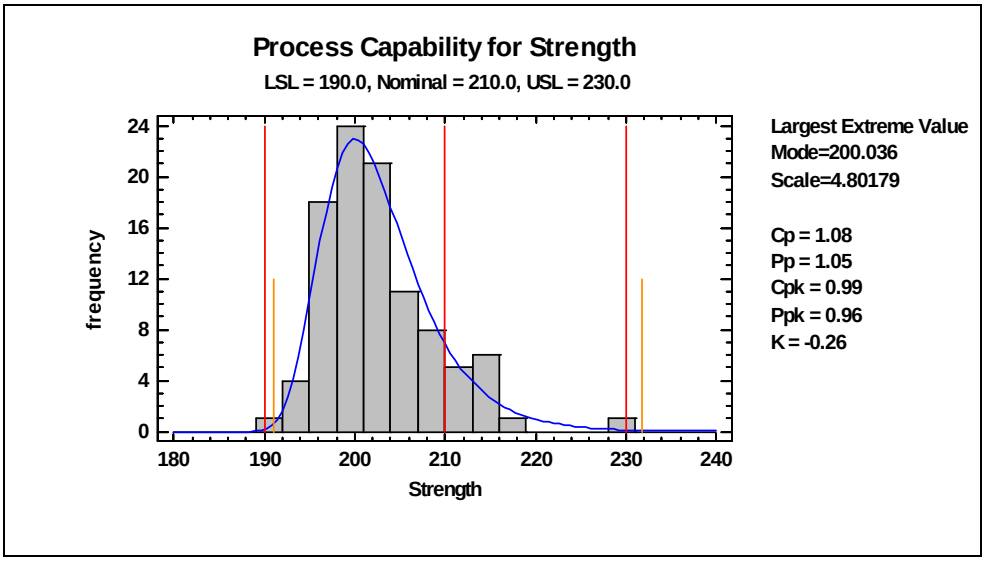


Figure 15-10. Fitted Largest Extreme Value Distribution

Notice that the distribution is skewed to the right, matching the observed data much better than the normal distribution. The short vertical lines have been positioned at "equivalent" 3 sigma limits, i.e., limits within which the same 99.73% of the fitted distribution is located as is the case for the mean plus and minus 3 sigma for a normal distribution. Note that these limits are not symmetrically spaced about the peak of the distribution, due to its positive skewness.

The *Analysis Summary* shows a dramatic difference in the estimated percent of the product that is likely to be out of specification, compared to the earlier fitted normal distribution:

Process Capability Analysis (Individuals) – Strength
Data variable: Strength (specs are 190-230)
Transformation: none

Distribution: Largest Extreme Value
sample size = 100
mode = 200.036
scale = 4.80179
(mean = 202.808)
(sigma = 6.15853)

Equivalent 6.0 Sigma Limits
99.865 percentile = 231.761
median = 201.796
0.134996 percentile = 190.969

	Observed		*Estimated*	*Defects*
Specifications	*Beyond Spec.*	*Z-Score*	*Beyond Spec.*	*Per Million*
USL = 230.0	0.000000%	2.89	0.194758%	1947.58
Nominal = 210.0		1.19		
LSL = 190.0	0.000000%	-3.42	0.030805%	308.05
Total	0.000000%		0.225563%	2255.63

Figure 15-11. Analysis Summary after Fitting Largest Extreme Value Distribution

The estimated percent out of spec is now only 0.23 percent, or 2,256 DPM, one-tenth of what it was using the normal distribution. In this case, incorrectly assuming a normal distribution makes the process appear much worse than it really is.

> NOTE: Depending on the specification limits and the true distribution, incorrectly assuming normality may make the process appear significantly worse or significantly better than when using the proper distribution.

An alternative to selecting a different distribution is to transform the data. The *Analysis Options* dialog box gives several choices for selecting a *Data Transformation*:

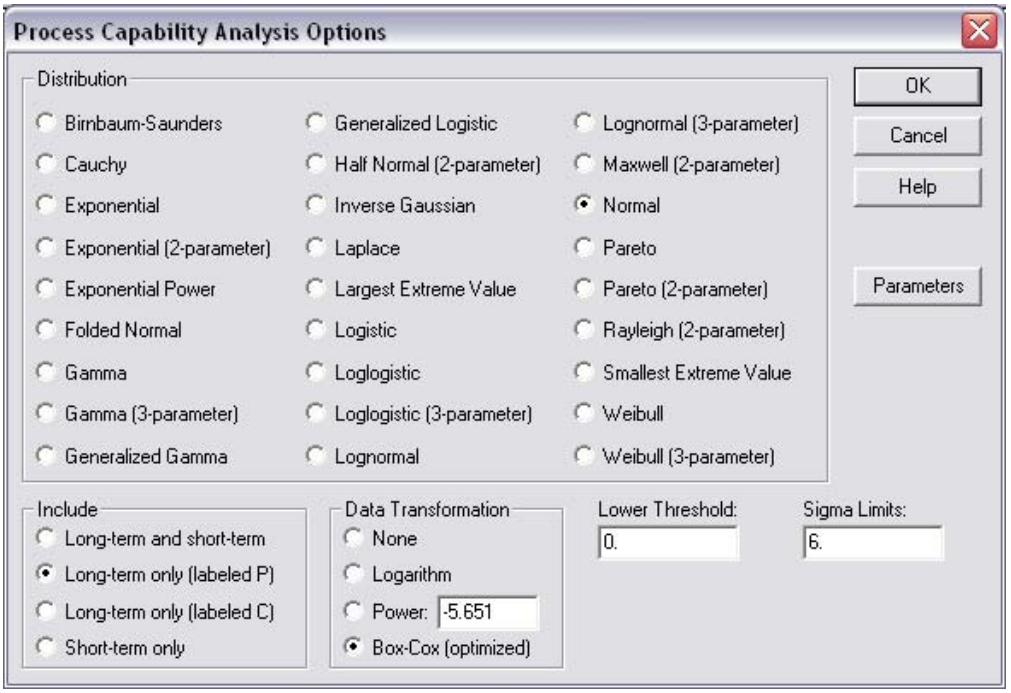


Figure 15-12. Analysis Options Dialog Box for Selecting Transformation

Choices include a natural logarithm, raising each value to a specified power, or selecting a transformation according to the methods of Box and Cox. The latter approach considers a variety of transformations of the form Y^p using the methods of Box and Cox and selects an optimal value for p.

If a transformation is selected, a normal distribution is fit to the transformed data. The plot below shows the results of taking the Box-Cox approach:

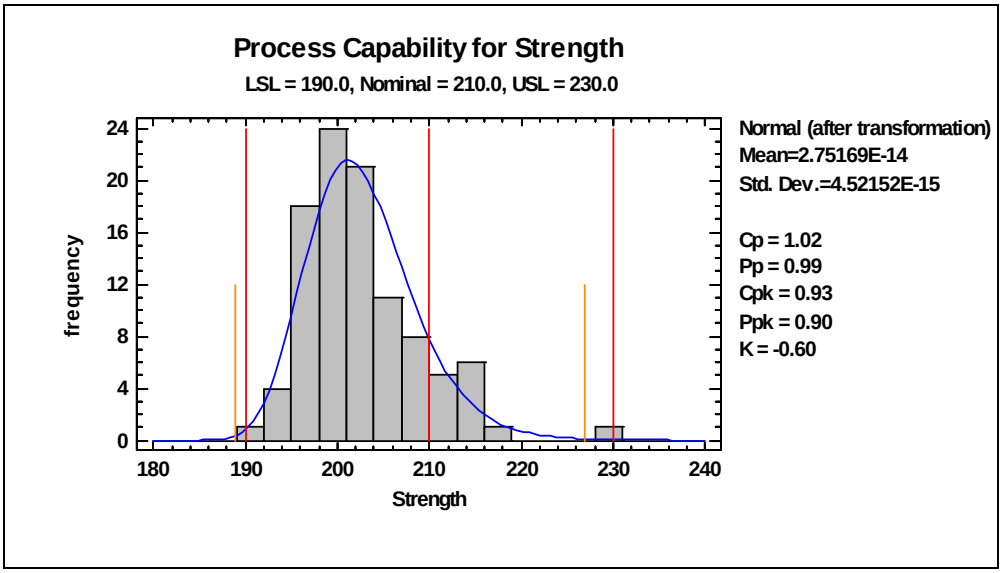


Figure 15-13. Capability Plot after Box-Cox Transformation

For the plot, an inverse transformation has been applied to show the fit in the original metric. The transformation has had a similar effect on the shape of the distribution, although not as strong as assuming a largest extreme value distribution. The estimated DPM is 4,353, which is about twice as large as when using the largest extreme value distribution, but still much smaller than when a normal distribution is assumed.

NOTE: the mean and standard deviation displayed on the plot correspond to the transformed data and are not in general very useful. STATGRAPHICS Centurion XVI converts everything back to the original units automatically for you.

To compare the two approaches, the *Probability Plot* may be selected from the *Tables and Graphs* dialog box for each approach and pasted side-by-side in the StatGallery:

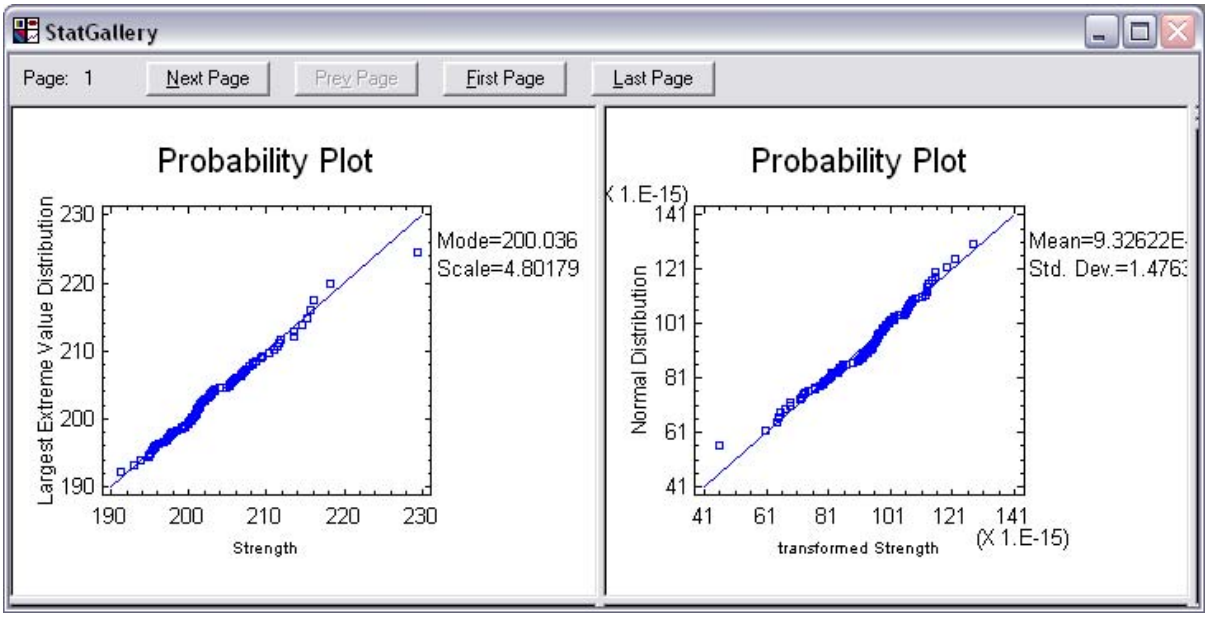


Figure 15-14. Probability Plots in the StatGallery

If the assumed distribution is correct, the points should fall along a diagonal line when displayed on this plot. Both methods appear to have handled the non-normality well, making it difficult to choose between them. Whichever method is used, it is important to establish a protocol for how to handle a particular variable (such as *Strength*) and apply that same protocol every time such data is analyzed. It would be a mistake to do the type of exploratory data analysis described in this chapter every time a set of similar data was collected. Instead, this type of analysis should be done once to determine how a selected variable needs to be handled, and then the selected approach should be applied to that variable whenever it is analyzed.

15.4 Capability Indices

The essence of a capability analysis lies in estimating the percentage of the product that falls outside the specification limits (or equivalently DPM, the defects per million). To summarize process capability, practitioners have also derived various capability indices. The most widely calculated index is C_{pk}, defined as:

$$C_{pk} = \min\left(\frac{\hat{\mu} - LSL}{3\hat{\sigma}}, \frac{USL - \hat{\mu}}{3\hat{\sigma}}\right)$$

Put simply, C_{pk} is the distance from the estimated process mean to the nearer specification limit, divided by 3 times the estimated process sigma.

The *Process Capability Analysis* procedure in STATGRAPHICS Centurion XVI displays capability indices on the *Capability Plot* and also on the *Capability Indices* table. If a normal distribution is assumed, both long-term and short-term indices are calculated:

Capability Indices for Strength

Specifications

USL = 230.0
Nom = 210.0
LSL = 190.0

	Short-Term	*Long-Term*
	Capability	*Performance*
Sigma	5.75525	6.23781
Cp/Pp	1.15836	1.06875
Cpk/Ppk	0.741874	0.684481
Cpk/Ppk (upper)	1.57485	1.45302
Cpk/Ppk (lower)	0.741874	0.684481
K		-0.35955
DPM	13020.9	20021.1
Sigma Quality Level	3.73	3.55

Based on 6 sigma limits. Short-term sigma estimated from average moving range. The Sigma Quality Level includes a 1.5 sigma drift in the mean.

95.0% Confidence Intervals

Index	*Lower Limit*	*Upper Limit*
Cp	0.997149	1.31931
Pp	0.920008	1.21725
Cpk	0.619618	0.864129
Ppk	0.568904	0.800059
Cpm	0.61885	0.777645

Figure 15-15. Table of Capability Indices

The short-term indices, which are calculated using an estimate of sigma obtained from observations close together in time, describe what the process is "capable" of doing if the mean were held constant. The long-term indices, which are calculated using an estimate of sigma obtained from the total variability amongst the observations throughout the sampling period, describe how the process has actually performed. An out of control process in which the mean has moved significantly over the course of the data collection may show considerably worse performance than it is capable of if it can be brought under control. By default, STATGRAPHICS Centurion XVI labels capability indices using the letter "C" and performance indices using the letter "P".

The *Capability* tab of the *Preferences* dialog box, accessible under *Edit* on the main STATGRAPHICS Centurion XVI menu, specifies the indices to be calculated by default, as well as other important options:

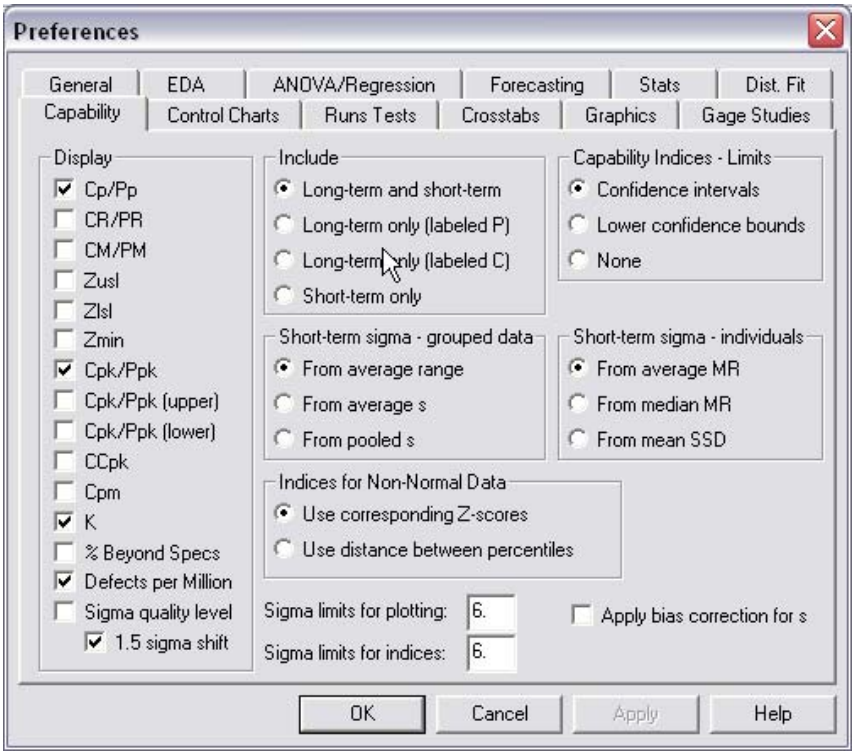


Figure 15-16. System Preferences for Capability Indices

The left-hand side of the dialog box lists the indices that may be calculated. In addition to C_{pk}, the available indices include:

1. C_p – a two-sided capability index calculated from

$$C_p = \frac{USL - LSL}{6\hat{\sigma}}$$

This index measures the distance between the specification limits relative to the distance covered by six standard deviations. C_p is always greater than or equal to C_{pk}. A substantial difference between the two indices indicates that the process is not well centered.

2. K – a measure of how far off center the process is. K is calculated from

$$K = \frac{\hat{\mu} - NOM}{(USL - LSL)/2}$$

where *NOM* is the nominal or target value. A value of K close to 0 is indicative of a well-centered process.

3. *Sigma Quality Level* – an index used in Six Sigma to indicate the level of quality associated with a process. A *Sigma Quality Level* of 6 is usually associated with a defect rate of 3.4 defects per million.

The *Preferences* dialog box also affects which indices are displayed on the *Capability Plot* and how they are labeled. A detailed discussion of the various indices may be found in the PDF document titled *Capability Analysis – Variable Data*.

In addition to the capability indices, the table in Figure 15.15 includes confidence intervals that show the margin of error in estimating those indices. For example, the above table shows a C_{pk} of 0.74. The 95% confidence interval extends from 0.62 to 0.86. This indicates that the true C_{pk} in the process from which the data were sampled may be anywhere between 0.62 and 0.86.

When the data do not follow a normal distribution, the capability indices need to be modified. The default option on the *Preferences* dialog box calculates non-normal indices by first computing equivalent Z-scores for the fitted non-normal distribution. For a normal distribution, the Z-score measures the number of standard deviations from the process mean to a specification limit and is directly related to the probability that an observation is beyond that limit. For a non-normal distribution, an equivalent Z-score is calculated by first determining the probability of exceeding the limit and then finding the Z-score that equates to that probability. After calculating equivalent Z-scores for both the upper and lower specification limits, C_{pk} may be calculated from

$$C_{pk} = \min(Z_{lsl}, Z_{usl})/3$$

NOTE: Although the *Preferences* dialog box gives the option of calculating capability indices from percentiles rather than equivalent Z-scores, doing so destroys the usual relationship between the capability indices and DPM.

15.5 Six Sigma Calculator

As an index, C_{pk} is a useful summary of process capability. Provided it is calculated properly, it can be related to DPM. The STATGRAPHICS Centurion XVI *Tools* menu contains a *Six Sigma Calculator* that will convert between the two, provided that either:

1. The data come from a normal distribution.

2. Equivalent Z-scores are used to calculate the indices.

The data input dialog box for the *Six Sigma Calculator* is shown below:

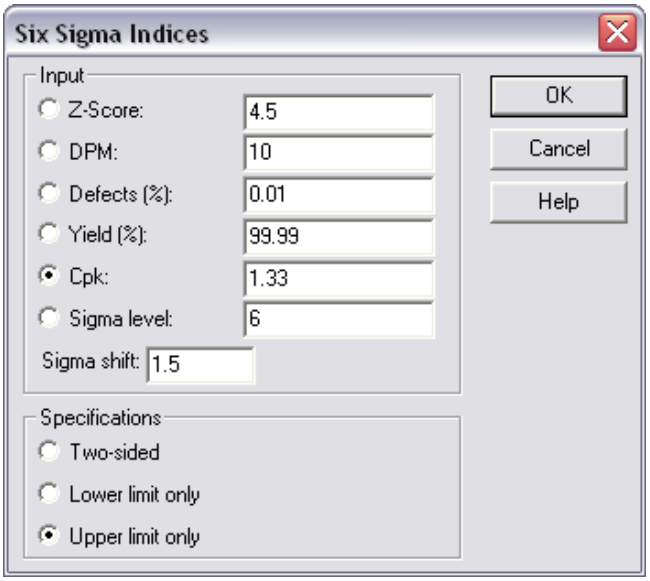


Figure 15-17. Six Sigma Calculator

To use the procedure:

1. Select any of the input radio buttons and enter a value for the corresponding statistic.

2. If you wish to calculate values based the nearer specification limit only, select either the *lower limit only* or *upper limit only* radio button.

3. Indicate the value you wish to assume for the long-term shift in the process mean. In Six Sigma, it is often assumed that the process mean will oscillate around its long-term value by 1.5 sigma.

4. Press the *Calculate* button to display the associated values of the other statistics.

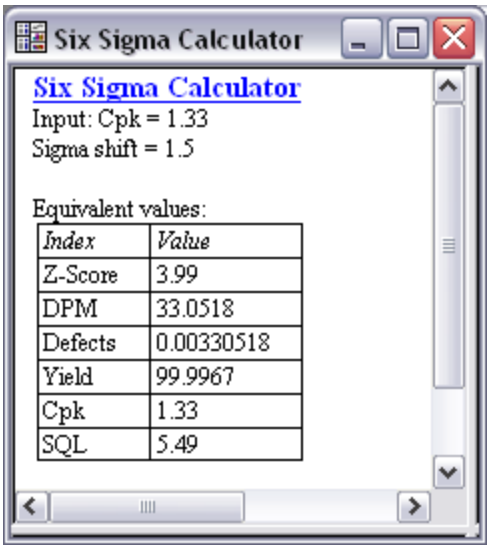


Figure 15-18. Equivalent Values of Quality Indices

Assuming that the process mean does not shift, a C_{pk} of 1.33 equates to about 33 defects per million beyond the nearer spec.

Chapter

16

Tutorial #7: Design of Experiments (DOE)

Designing an experiment to help improve a process.

All data are not created equal. Often, a small but properly planned study provides more information than a large, badly designed study. This final tutorial examines some of the capabilities of STATGRAPHICS Centurion XVI for creating and analyzing designed experiments.

Consider the case of an engineer who wishes to determine which of many process variables have the greatest impact on the final product. She intends to investigate the impact of changing 5 factors: input temperature, flow rate, concentration, agitation rate, and percent of catalyst. In practice, this problem could be approached in several ways, including:

1. *Trial and error*: arbitrarily selecting a different combination of the factors each time she runs an experiment. Such an approach rarely yields useful information.

2. *One factor at a time experimentation*: holding all but one factor constant to determine the effect of that factor. This approach is extremely inefficient and can be misleading if any of the factors interact.

3. *Using a statistically designed experiment*: setting out a sequence of experiments to perform that will yield the most information about the factors and their interactions in as few experiments as possible.

This tutorial will describe how an experimental design could be constructed using the third approach and how the resulting data would be analyzed.

16.1 Creating the Design

STATGRAPHICS Centurion XVI contains an Experimental Design Wizard that guides users through the construction and analysis of a designed experiment. To access the DOE wizard:

1. If using the Classic menu, select *DOE – Experimental Design Wizard.*

2. If using the Six Sigma menu, select *Improve – Experimental Design Wizard.*

A new window will be created containing a toolbar which guides you through a sequence of 12 steps:

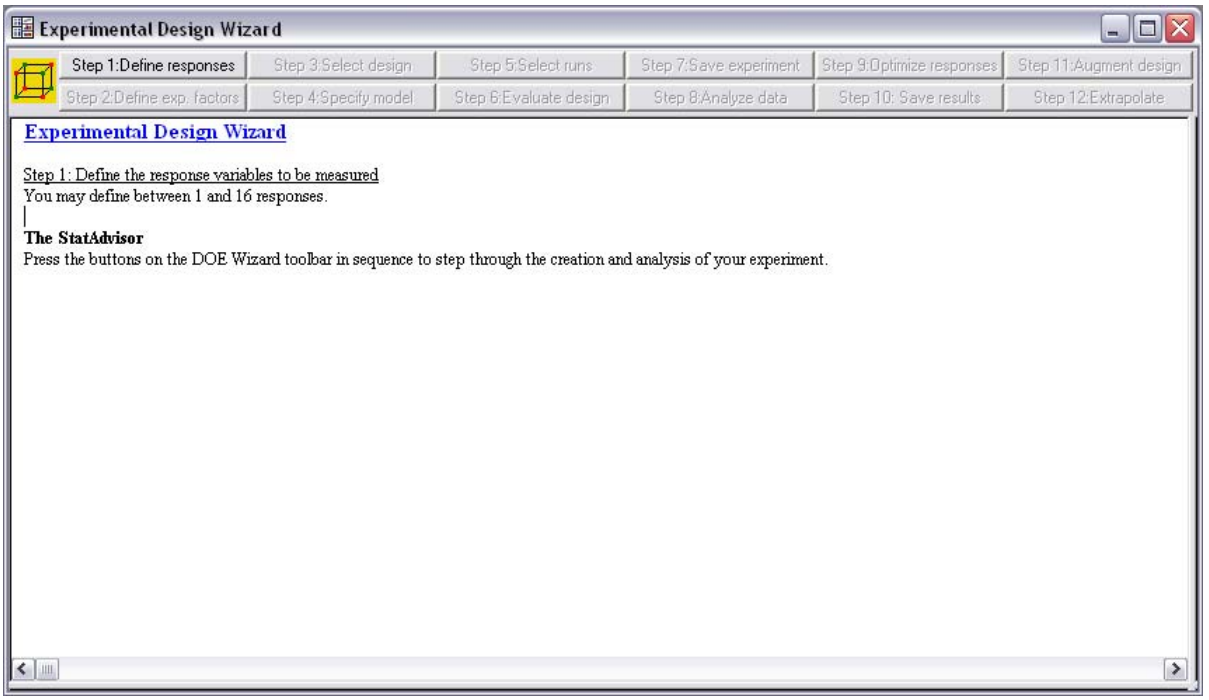
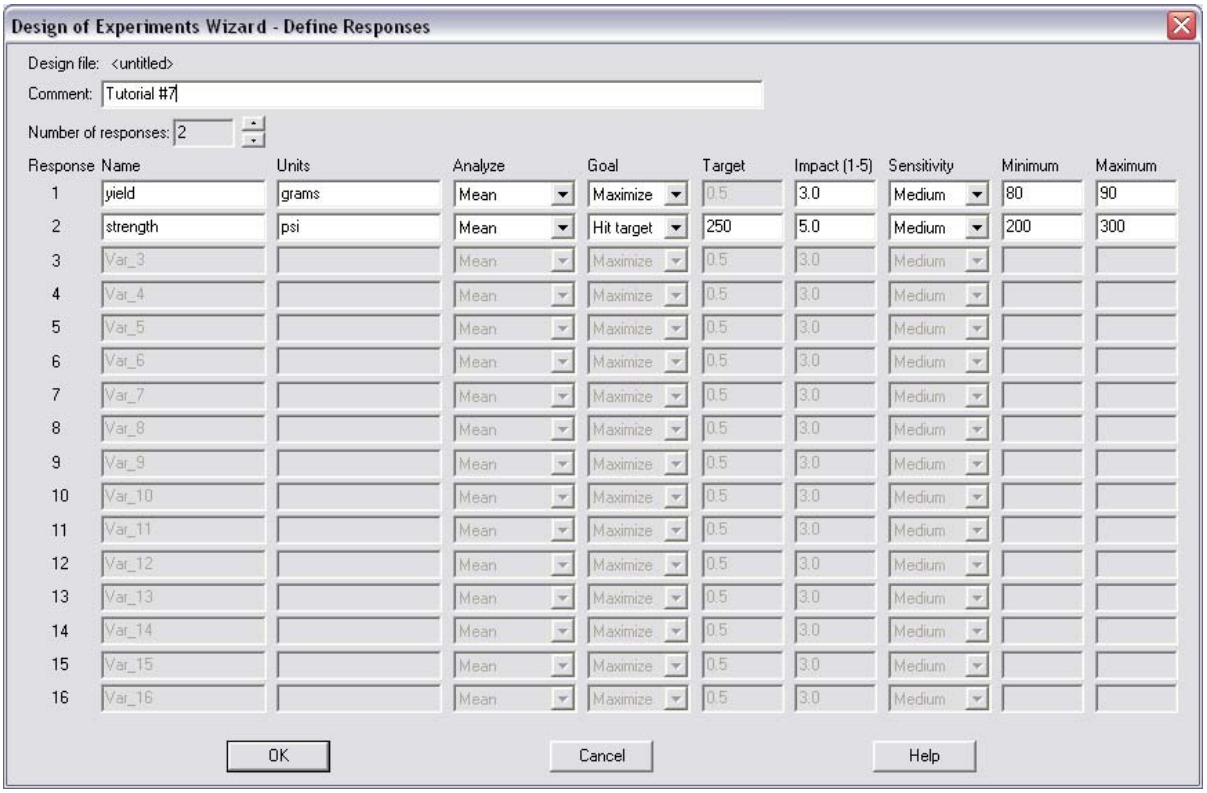


Figure 16-1. Experimental Design Wizard's Main Window with 12-Step Toolbar

The first 7 steps of the sequence construct the experimental design and are executed before the experiment is performed. The last 5 steps are executed after the experiment has been completed and deal with analysis of the resulting data.

Step 1: Define responses

The first step in creating a designed experiment is to specify the response variables that will be measured during each experimental run. Pressing the *Step 1* button displays the following dialog box:



Response	Name	Units	Analyze	Goal	Target	Impact (1-5)	Sensitivity	Minimum	Maximum
1	yield	grams	Mean	Maximize	0.5	3.0	Medium	80	90
2	strength	psi	Mean	Hit target	250	5.0	Medium	200	300
3	Var_3		Mean	Maximize	0.5	3.0	Medium		
4	Var_4		Mean	Maximize	0.5	3.0	Medium		
5	Var_5		Mean	Maximize	0.5	3.0	Medium		
6	Var_6		Mean	Maximize	0.5	3.0	Medium		
7	Var_7		Mean	Maximize	0.5	3.0	Medium		
8	Var_8		Mean	Maximize	0.5	3.0	Medium		
9	Var_9		Mean	Maximize	0.5	3.0	Medium		
10	Var_10		Mean	Maximize	0.5	3.0	Medium		
11	Var_11		Mean	Maximize	0.5	3.0	Medium		
12	Var_12		Mean	Maximize	0.5	3.0	Medium		
13	Var_13		Mean	Maximize	0.5	3.0	Medium		
14	Var_14		Mean	Maximize	0.5	3.0	Medium		
15	Var_15		Mean	Maximize	0.5	3.0	Medium		
16	Var_16		Mean	Maximize	0.5	3.0	Medium		

Figure 16.2. Definition of Response Variables

In the example, there are two response variables: *yield* in grams and *strength* in pounds per square inch. The goal of the experiment is to maximize the *yield* while keeping *strength* as close to 250 as possible. The four rightmost columns are used to balance the requirements of the two responses, which could conflict. *Impact* specifies the importance of each response on a scale of 1 to 5, with 5 meaning most important. The *minimum* and *maximum* values specify the desirable range for each response, with *sensitivity* indicating how important it is to be close to the best position within that range. In the current example, *strength* is more important than *yield* and thus is assigned a higher impact. The sensitivity of both responses is set to "Medium", which means that the desirability of each response increases in a linear fashion throughout the specified range.

Step 2: Define experimental factors

The *Step 2* button is used to enter information about the experimental factors that will be changed during the course of the experiment. It displays the dialog box shown below:

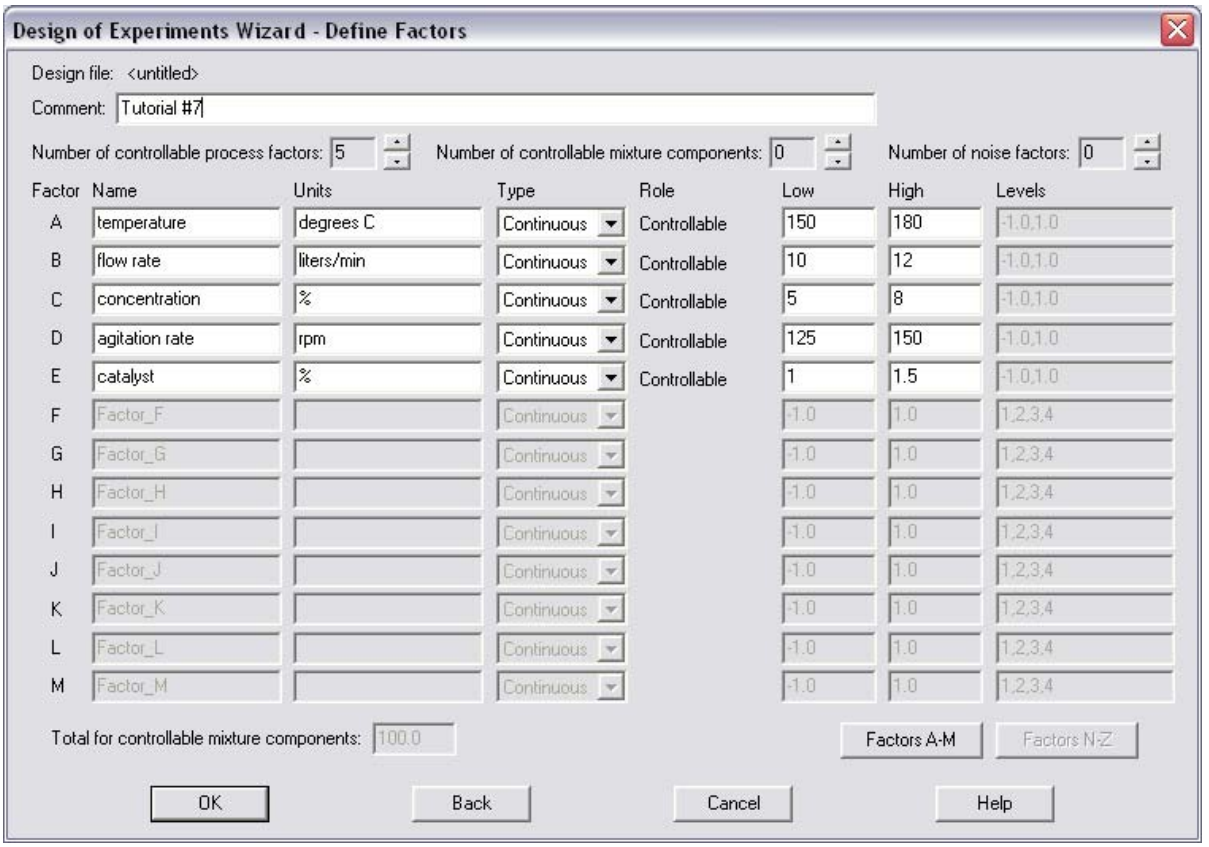


Figure 16-3. Definition of Experimental Factors

In the example, 5 controllable process factors are changed. Enter the name of each factor, its units, and the range over which it will be changed. All factors are *continuous*, since they can be set to any value between the indicated low and high levels.

Step 3: Select design

The third step in creating an experiment is to select the type of design to be performed. When the *Step 3* button is pressed, the first dialog box displayed is shown below:

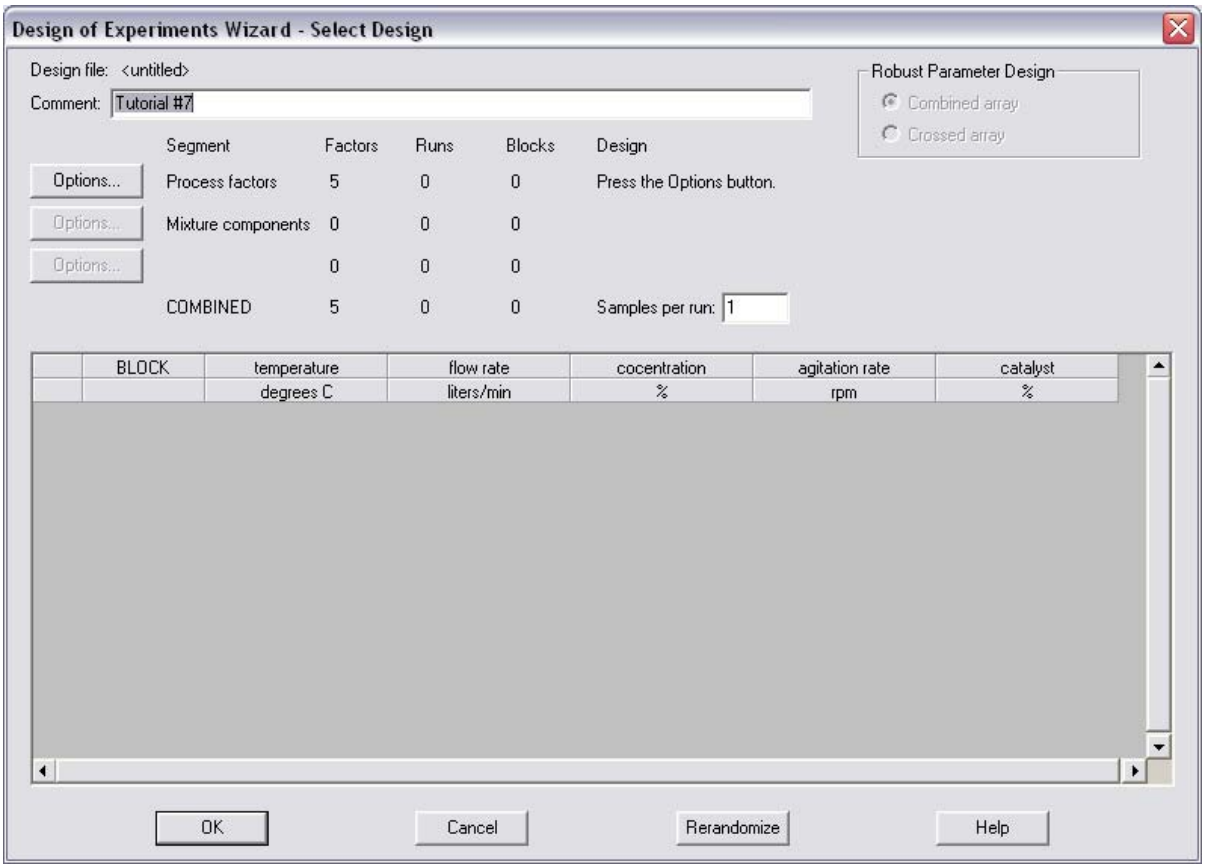


Figure 16-4. Select Design Dialog Box

To create a design for the 5 process factors, press the *Options* button. This displays a list of types of designs that would be appropriate for 5 continuous factors:

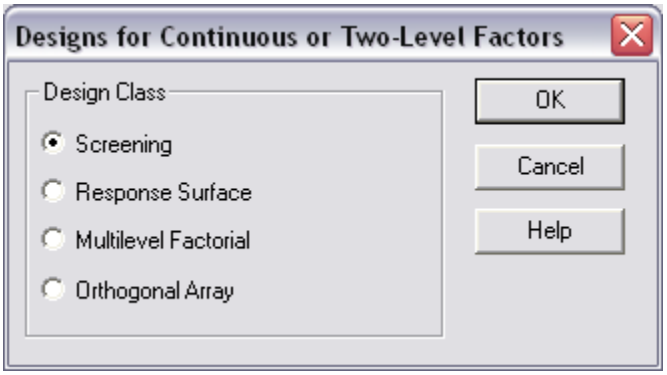


Figure 16.5. Dialog Box Showing Available Types of Designs

Since we wish to create a screening design, just press *OK*.

The next dialog box is used to select the desired design from a catalog of screening designs appropriate for 5 factors:

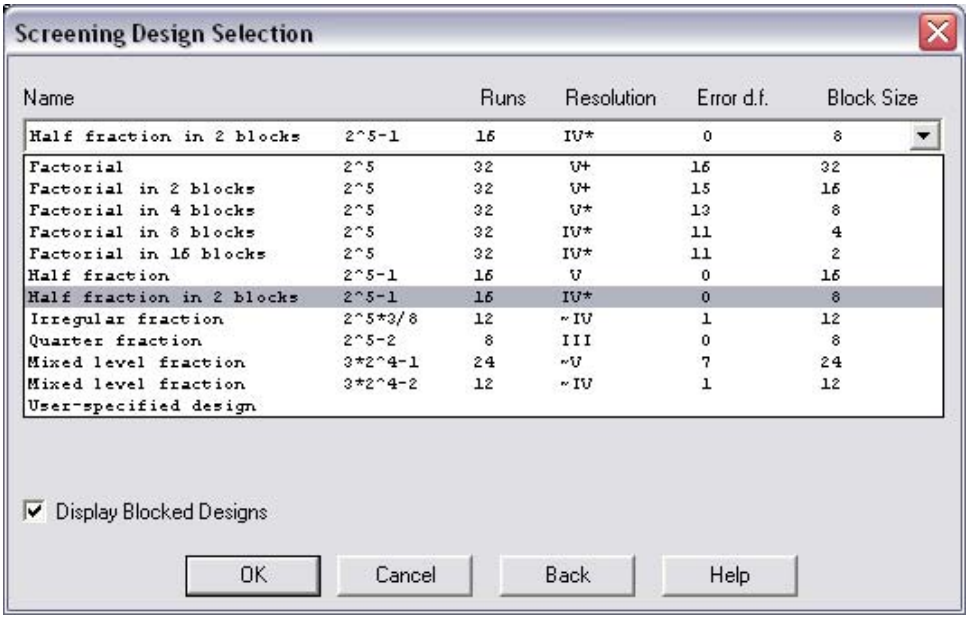


Figure 16-6. Design Selection

To see the list of screening designs available for five factors, click on the arrow to pull down the list. The list shows:

1. ***Name***: the name of each available design.

2. ***Runs***: the number of runs in the base design, before any centerpoints or replicate runs are added.

3. ***Resolution***: the resolution of the design. Resolution V designs can estimate all main effects and all two-factor interactions. Resolution IV designs can estimate all main effects, but two-factor interactions are confounded amongst themselves or with block effects. Resolution III designs confound two-factor interactions with main effects.

4. ***Error d.f.***: the number of degrees of freedom available to estimate the experimental error. The power of the statistical tests is related to the number of degrees of freedom, as well as the total number of runs in the experiment. Normally, at least 3 degrees of freedom should be available, although more is preferable.

5. ***Block size***: the number of runs in the largest block.

In this case, the engineer selected a half-fraction in two blocks of 8 runs each.

The final dialog box is used to add centerpoints or replicate runs:

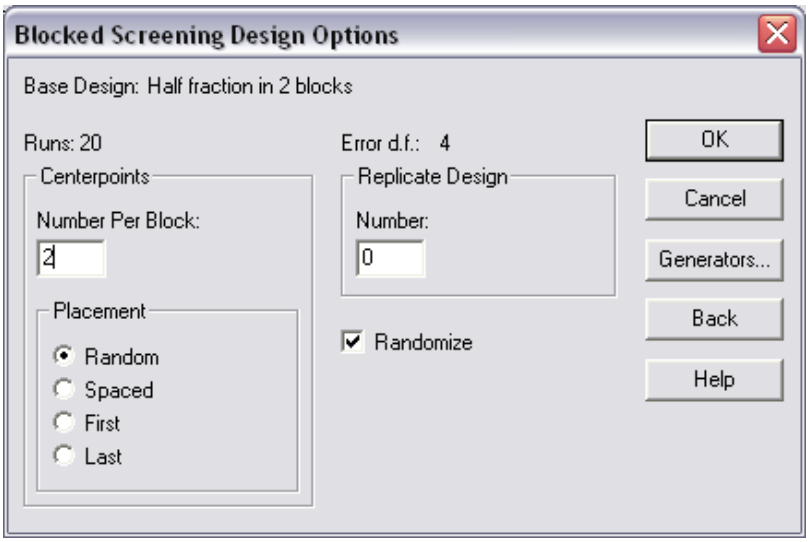


Figure 16-7. Blocked Screening Design Options

The input fields specify:

1. *Centerpoints*: the number of runs to be performed in the center of the experimental region. Adding centerpoints is a good way to add degrees of freedom for the experimental error.

2. *Placement*: the placement of the centerpoints. The most common choices are *Random*, which spreads the centerpoints randomly throughout the other runs, and *Spaced*, which spaces the centerpoints evenly throughout the design.

3. *Replicate design*: the number of additional times each set of experimental conditions is to be run. Replicating the entire design this way can increase the number of runs to be done very quickly.

4. *Randomize*: whether the runs should be listed in random order. Randomization should be done whenever possible to prevent external lurking variables (such as changes in the process over time) to bias the results.

For the current experiment, four centerpoints have been requested, bringing the final design up to 20 runs. It has also been requested that the design be done in random order, which means that the order of the 10 runs within each block will be randomly generated.

After the final dialog box, the *Select Design* window is filled with the experimental runs to be performed:

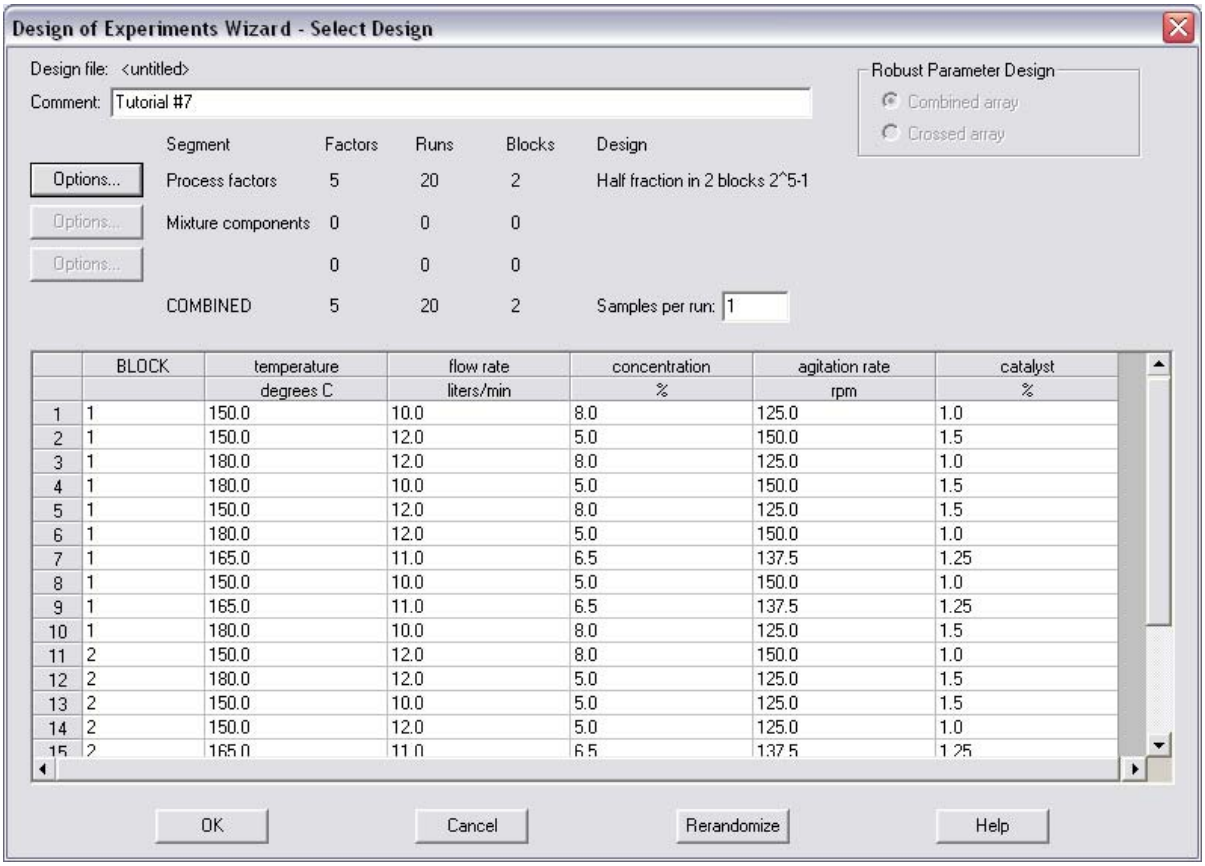


Figure 16-8. Select Design Window with Runs to be Performed

If you are satisfied with the design, press *OK* once more to return to the *Experimental Design Wizard* window which will summarize the selections made so far:

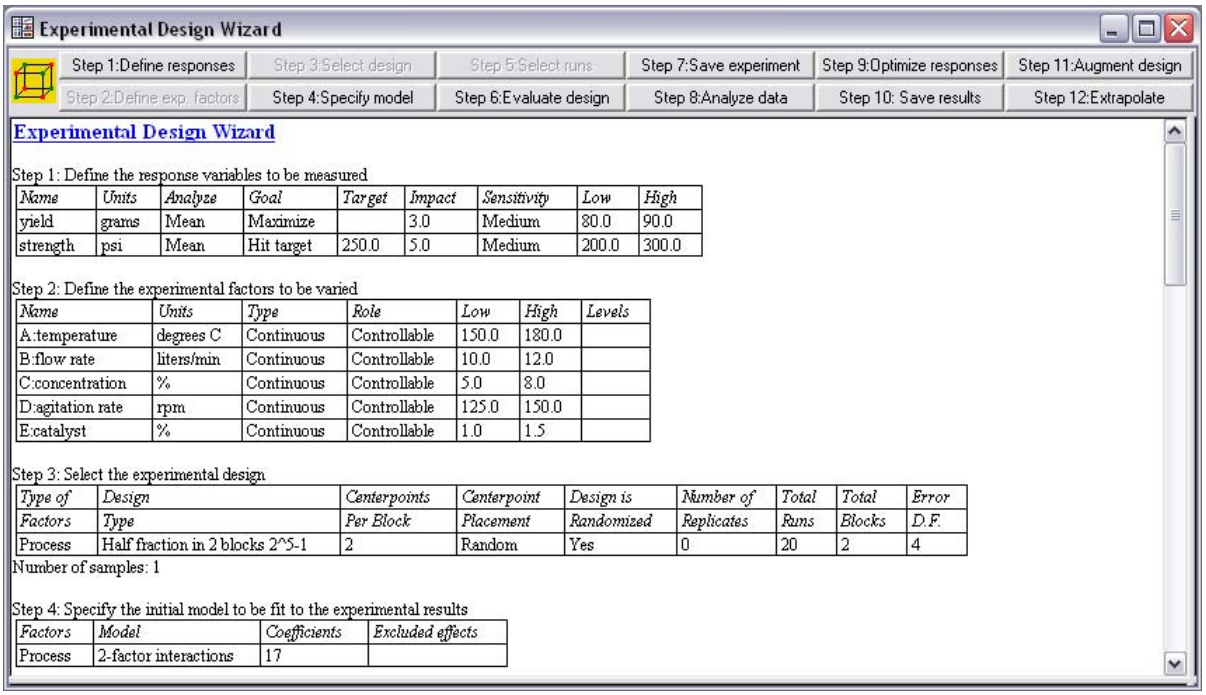
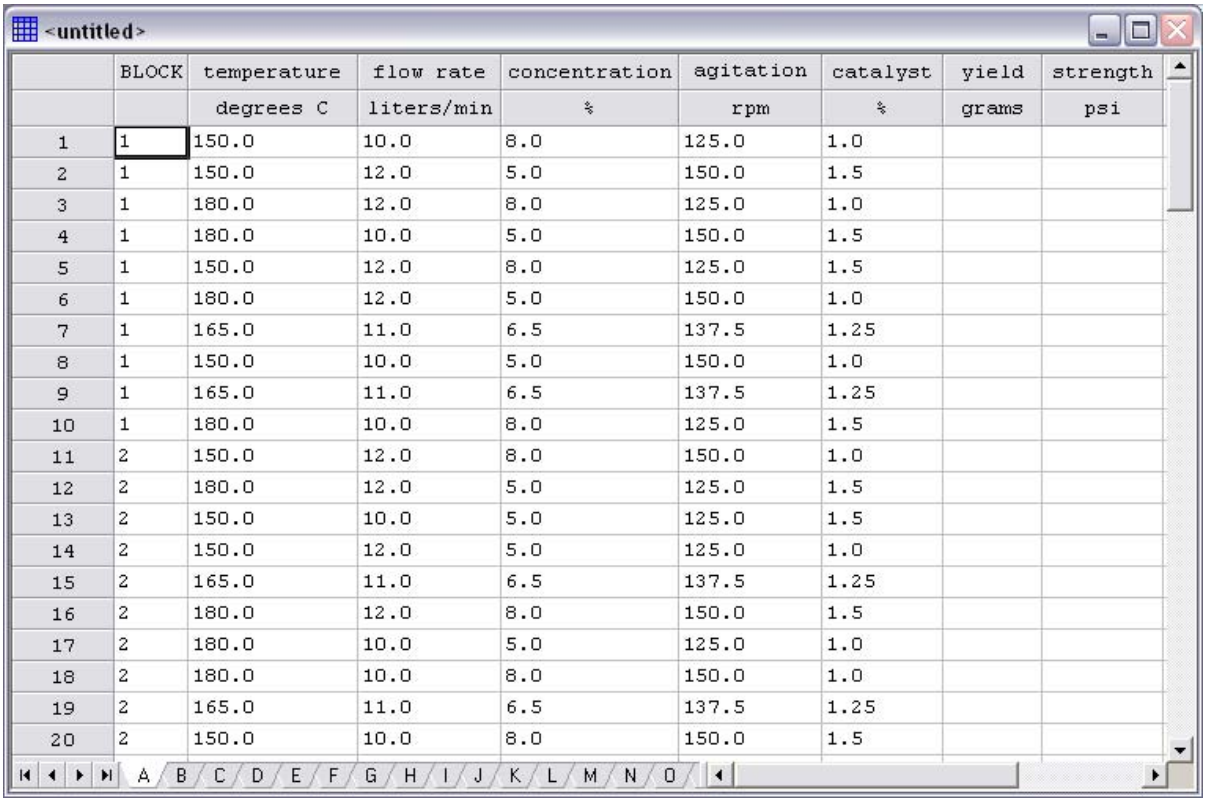


Figure 16-9. Experimental Design Wizard Window After Selecting a Design

At the same time, the design has been loaded into datasheet A of the STATGRAPHICS Centurion XVI DataBook:



	BLOCK	temperature	flow rate	concentration	agitation	catalyst	yield	strength
		degrees C	liters/min	%	rpm	%	grams	psi
1	1	150.0	10.0	8.0	125.0	1.0		
2	1	150.0	12.0	5.0	150.0	1.5		
3	1	180.0	12.0	8.0	125.0	1.0		
4	1	180.0	10.0	5.0	150.0	1.5		
5	1	150.0	12.0	8.0	125.0	1.5		
6	1	180.0	12.0	5.0	150.0	1.0		
7	1	165.0	11.0	6.5	137.5	1.25		
8	1	150.0	10.0	5.0	150.0	1.0		
9	1	165.0	11.0	6.5	137.5	1.25		
10	1	180.0	10.0	8.0	125.0	1.5		
11	2	150.0	12.0	8.0	150.0	1.0		
12	2	180.0	12.0	5.0	125.0	1.5		
13	2	150.0	10.0	5.0	125.0	1.5		
14	2	150.0	12.0	5.0	125.0	1.0		
15	2	165.0	11.0	6.5	137.5	1.25		
16	2	180.0	12.0	8.0	150.0	1.5		
17	2	180.0	10.0	5.0	125.0	1.0		
18	2	180.0	10.0	8.0	150.0	1.0		
19	2	165.0	11.0	6.5	137.5	1.25		
20	2	150.0	10.0	8.0	150.0	1.5		

Figure 16-10. Final Design

The datasheet contains a column with block numbers, 5 columns with the settings of the experimental factors, and 2 columns for entering the responses once the experimental runs have been performed.

Step 4: Specify model

The Experimental Design Wizard will evaluate the design you have created with respect to a specific statistical model. If you press the button labeled *Step 4*, the following dialog box will be displayed:

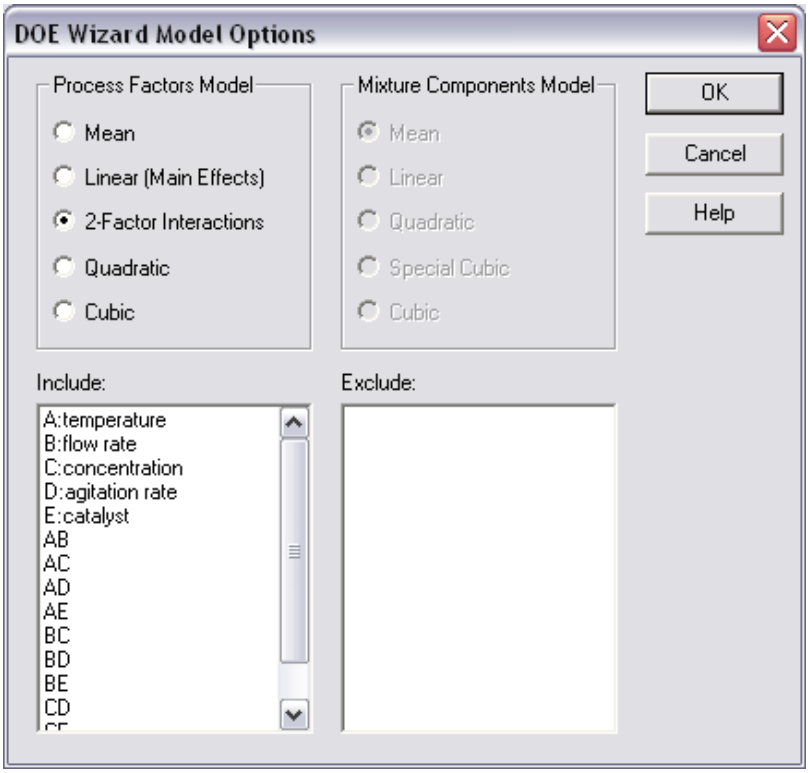


Figure 16-11. Model Selection Dialog Box

You should select the most complicated model you wish to consider for your data. In the case of a two-level factorial, the most complicated model that can be fit is the two-factor interaction model defined by:

$$Y = \begin{aligned} &\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 \\ &+ \beta_{15} x_1 x_5 + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{25} x_2 x_5 + \beta_{34} x_3 x_4 + \beta_{35} x_3 x_5 + \beta_{45} x_4 x_5 \end{aligned}$$

It consists of each experimental factor by itself (the main effects) and terms involving each pair of factors (two-factor interactions). Individual terms can be excluded from the selected model by

double-clicking on them with the mouse, which moves them to the excluded field on the dialog box. In this case, we will select the full 2-factor interaction model.

Step 5: Select runs

For more complicated designs, it may be desirable to run only a subset of the runs that were created in Step 3. If the *Step 5* button is pressed, a run selection algorithm can be accessed to create a subset of the runs that is *D-optimal*. In this case, all runs will be performed, so Step 5 can be omitted.

Step 6: Evaluate design

If the button labeled *Step 6* is pressed, a dialog box will be displayed showing a list of tables and graphs that can be added to the Experimental Design Wizard's window:

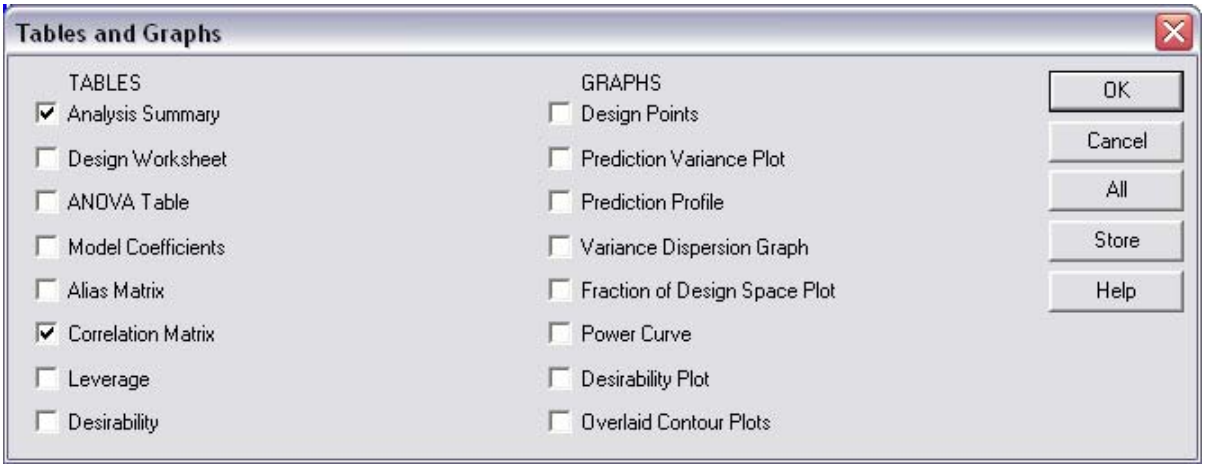


Figure 16-12. Tables and Graphs for Evaluating Selected Experimental Design

A useful option for screening designs is the *Correlation Matrix*, which shows whether there is any confounding amongst the terms in the model that will be fit:

Correlation Matrix

	block	A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE	CD
block	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8944
A	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
B	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
C	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
E	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AD	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
BC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
BD	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
BE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
CD	0.8944	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
CE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

	CE	DE
block	0.0000	0.0000
A	0.0000	0.0000
B	0.0000	0.0000
C	0.0000	0.0000
D	0.0000	0.0000
E	0.0000	0.0000
AB	0.0000	0.0000
AC	0.0000	0.0000
AD	0.0000	0.0000
AE	0.0000	0.0000
BC	0.0000	0.0000
BD	0.0000	0.0000
BE	0.0000	0.0000
CD	0.0000	0.0000
CE	1.0000	0.0000
DE	0.0000	1.0000

Figure 16-13. Correlation Matrix for Selected Design

A non-zero value in any off-diagonal cell of the table indicates that the effects of that row and column are confounded and cannot be separated cleanly. In the current design, the CD interaction has a large correlation with blocks. Note that the design has arbitrarily sacrificed the ability to estimate the interaction between factors C and D, which are *concentration* and *agitation rate*. If this is an interaction that the engineer believes may be important, she should change the order of the variables so that C and D correspond to two variables that are not likely to interact.

Step 7: Save experiment

Pressing the button labeled *Step 7* allows the experiment to be saved in a file. It uses the dialog box shown below:

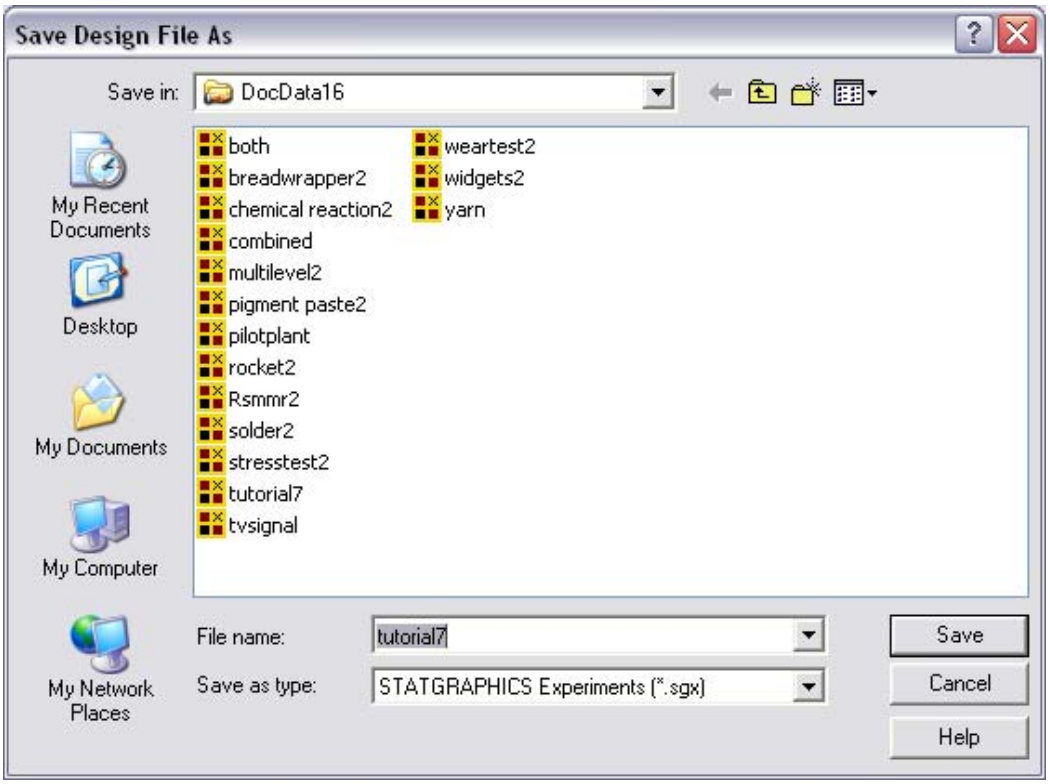


Figure 16-14. Dialog Box for Saving Experiment

Experimental designs created by the DOE Wizard are saved in files with the extension *.sgx*. These are similar to standard data files, with the exception that they contain additional information about the experimental design and the selected statistical model.

16.2 Analyzing the Results

After designing the experiment, the engineer performed the indicated 20 runs. She then restarted the program, opened the stored experiment file, and entered the measured values of *yield* and *strength* into the experiment datasheet. To replicate her analysis, you may load the *tutorial7.sgx* file

in the same manner as you would any STATGRAPHICS data file, by selecting *Open Data Source* from the *File* menu. Opening an experiment file automatically opens the main DOE Wizard window.

Step 8: Analyze data

To analyze the experimental data, press the button labeled *Step 8*. The analysis begins by displaying the following dialog box:

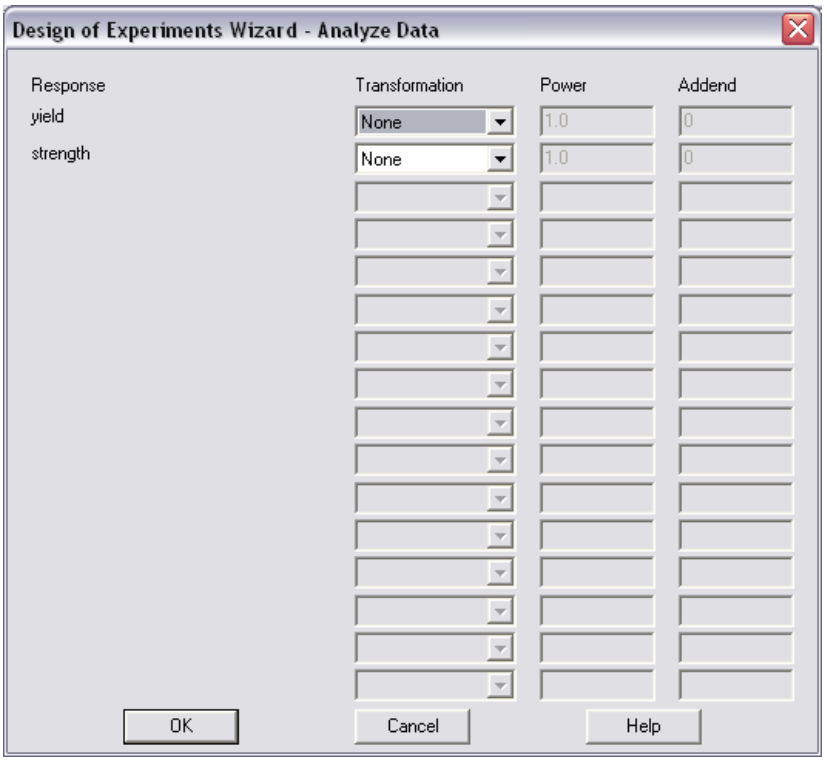


Figure 16-15. Analyze Data Input Dialog Box

If necessary, a transformation may be specified for one or more variables. Since we are analyzing continuous response variables, transformations might be necessary if the variance of the response increases with the mean. In the current example, no transformations are needed.

When the *OK* button is pressed, a new analysis window is opened for each response. The analysis window for *Yield* initially shows the following output:

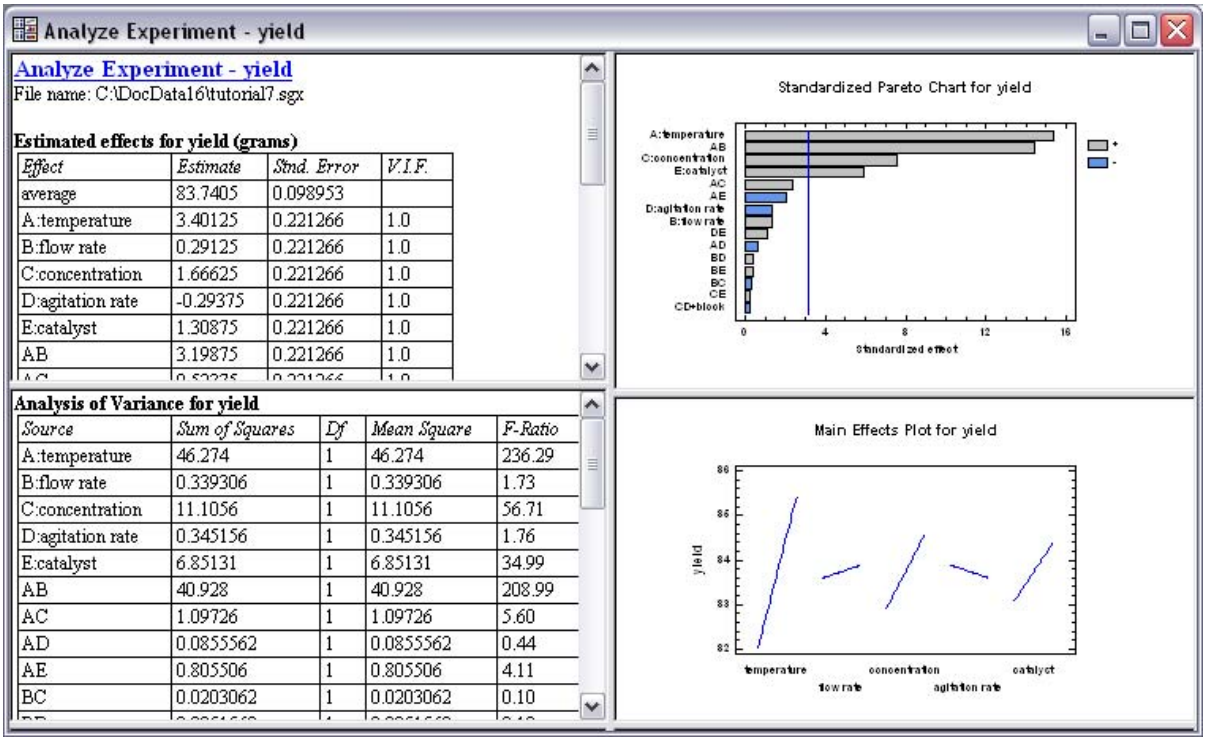


Figure 16-16. Analyze Design Analysis Window for Yield

The window displays four panes:

1. *Analysis Summary*: lists the estimated main effects and interactions.

2. *ANOVA Table*: contains *P*-values that can be used to test the statistical significance of each effect.

3. *Standardized Pareto Chart*: display the effects in decreasing order of significance, with a line by which to determine which effects are statistically significant.

4. *Main Effects Plot*: plots the estimated change in the response when each of the factors is moved from its low level to its high level.

The standardized Pareto chart in the upper right corner can be used to quickly determine which effects are most important:

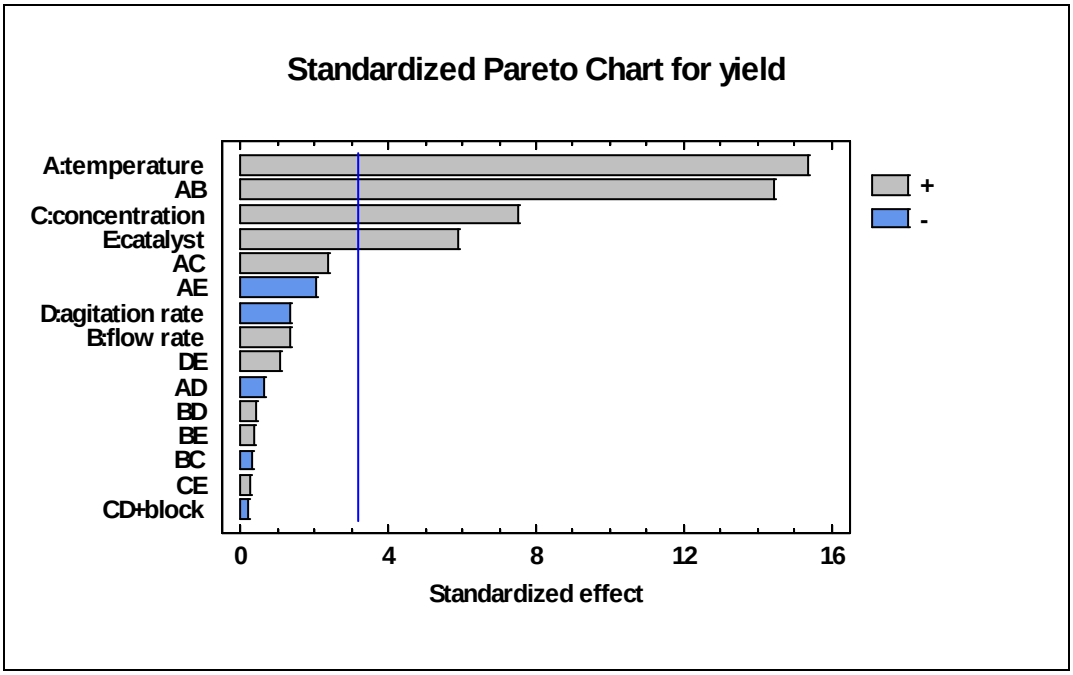


Figure 16-17. Standardized Pareto Chart

The length of each bar is proportional to the value of a t-statistic calculated for the corresponding effect. Any bars beyond the vertical line are statistically significant at the selected significance level, set by default at 5%. In this case, there are 3 significant main effects: *temperature*, *concentration*, and *catalyst*. There is also a significant interaction between *temperature* and *flow rate*.

The *Main Effects Plot* in the bottom right pane shows how each factor affects *yield*:

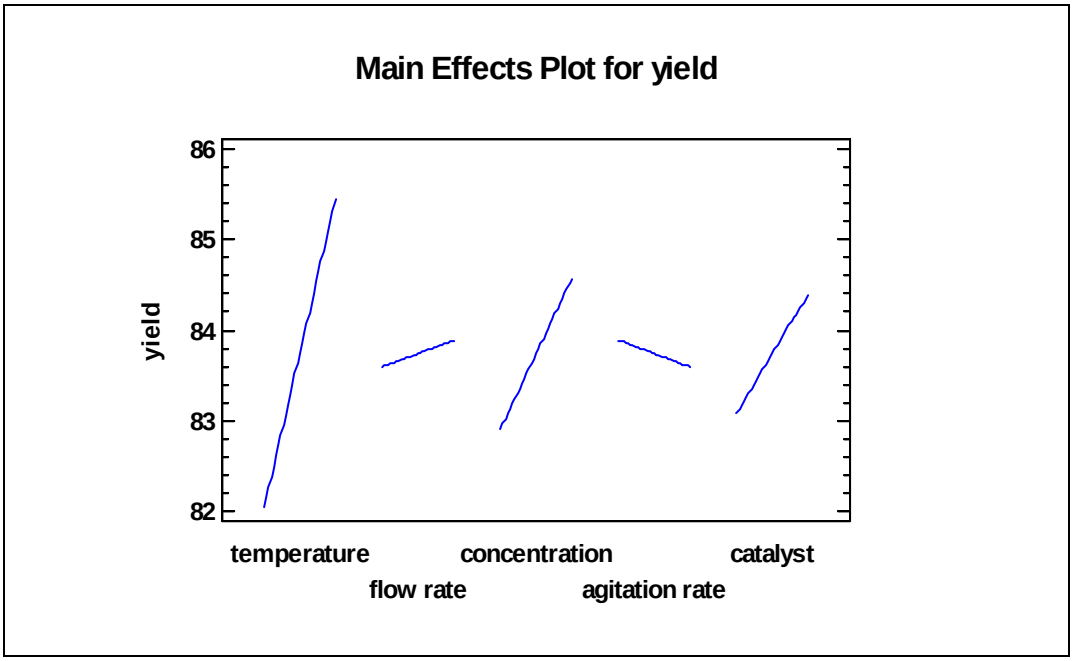


Figure 16-18. Main Effects Plot

The lines indicate the estimated change in *yield* as each factor is moved from its low level to its high level, with all other factors held constant at a value midway between their lows and their highs. Note that the three factors with significant main effects have a bigger impact on the response than the others. For example, the average yield at low temperature is approximately 82, while the average yield at high temperature is approximately 85.4. The difference of 3.4 is called the "main effect" of temperature.

To plot the interaction between *temperature* and *flow rate*, first select *Interaction Plot* from the *Graphs* dialog box. Then use *Pane Options* to select only those two factors:

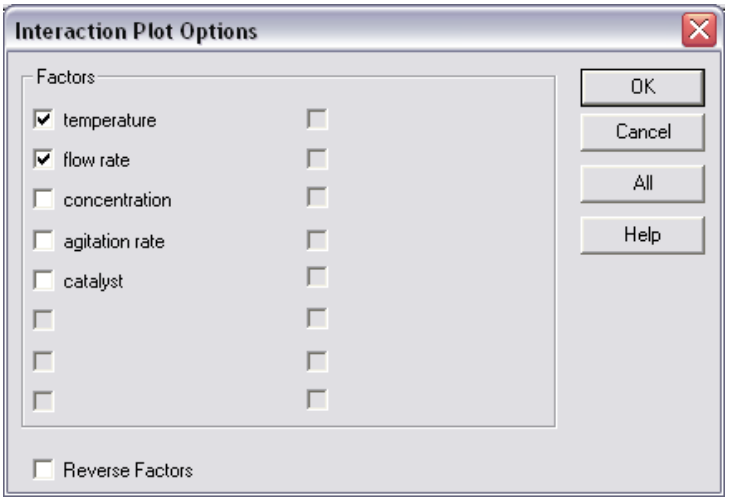


Figure 16-19. Pane Options Dialog Box for Interaction Plot

The resulting plot shows the average *yield* as *temperature* is changed, for each level of *flow rate*:

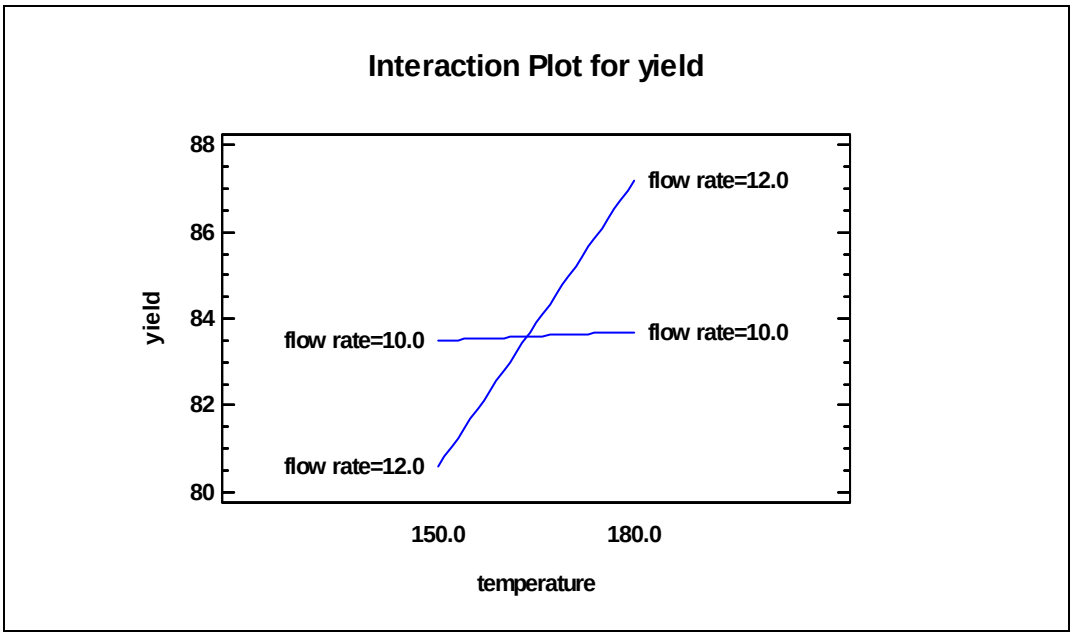


Figure 16-20. Interaction Plot for Flow Rate and Temperature

Notice that at low *flow rate*, *temperature* has little if any effect. At high *flow rate*, *temperature* is a very important factor.

Before using the statistical model underlying this analysis, it is important to remove insignificant effects. To remove effects:

1. Press the *Analysis Options* button on the analysis toolbar.

2. Press the *Exclude* button on the *Analysis Options* dialog box.

3. On the *Exclude Effects Options* dialog box, double-click on any effect you wish to exclude, which moves it from the *Include* column to the *Exclude* column:

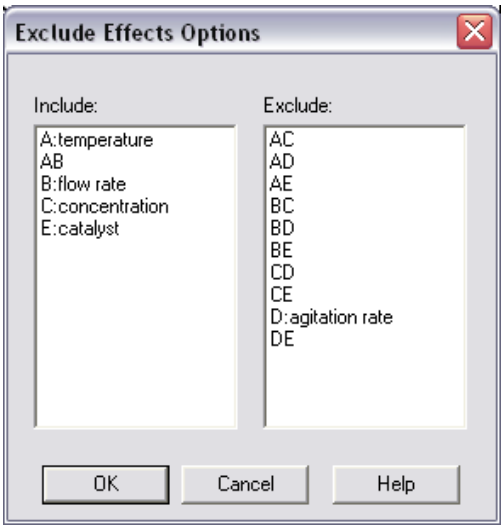


Figure 16-21. Dialog Box for Excluding Effects

The rule to follow in excluding effects is:

1. Exclude any insignificant two-factor interactions.

2. Exclude any insignificant main effects that are not involved in significant interactions.

In this case, that means removing everything that was not significant on the Pareto chart, except for the main effect of B. That main effect is retained because it is involved in a significant interaction with factor A.

Once the effects are removed, the Pareto chart should appear as shown below:

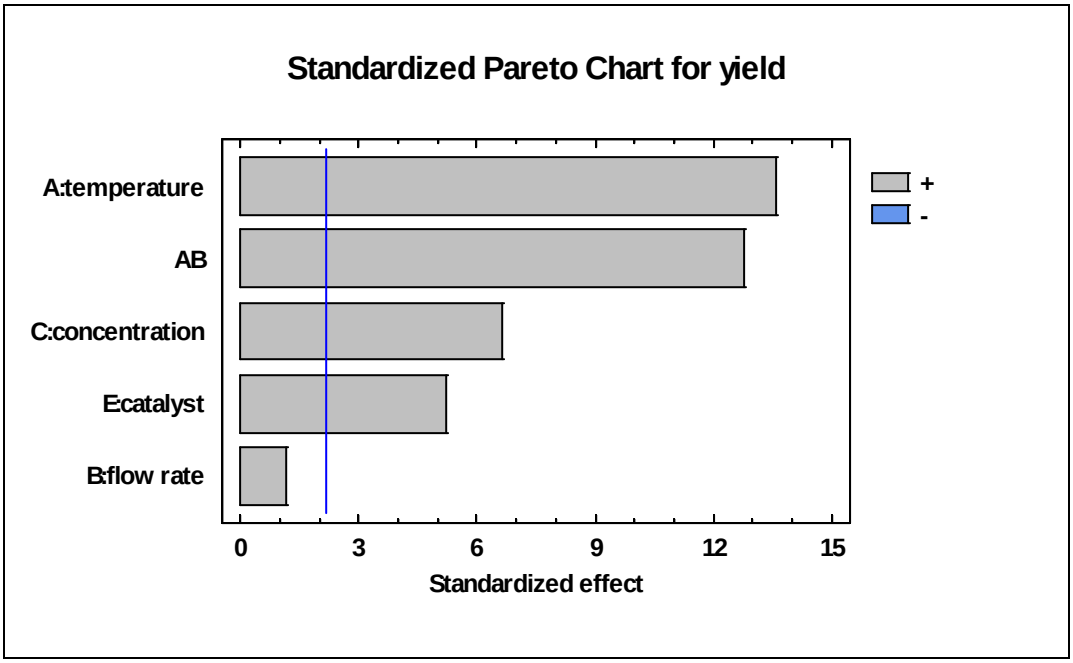


Figure 16-22. Standardized Pareto Chart after Removing Effects

Except for the main effect of factor B, all of the remaining effects are statistically significant. The final model may be viewed by selecting *Regression Coefficients* from the *Tables* dialog box:

Regression coeffs. for yield - Tutorial #7

Coefficient	*Estimate*
constant	250.074
A:temperature	-1.0595
B:flow rate	-17.4475
C:concentration	0.555417
E:catalyst	2.6175
AB	0.106625

The StatAdvisor

This pane displays the regression equation which has been fitted to the data. The equation of the fitted model is

yield = 250.074 - 1.0595*temperature - 17.4475*flow rate + 0.555417*concentration + 2.6175*catalyst + 0.106625*temperature*flow rate

Figure 16-23. Fitted Regression Model for Yield

Note that the underlying model takes the form of a multiple linear regression model. Each retained main effect is included in the model by itself, while the two-factor interaction is represented by a crossproduct of *temperature* and *flow rate*.

To fully understand the fitted model, it is best to plot it. Several types of plots may be created by selecting *Response Plots* from the *Tables and Graphs* dialog box. By default, a wire frame surface plot is displayed:

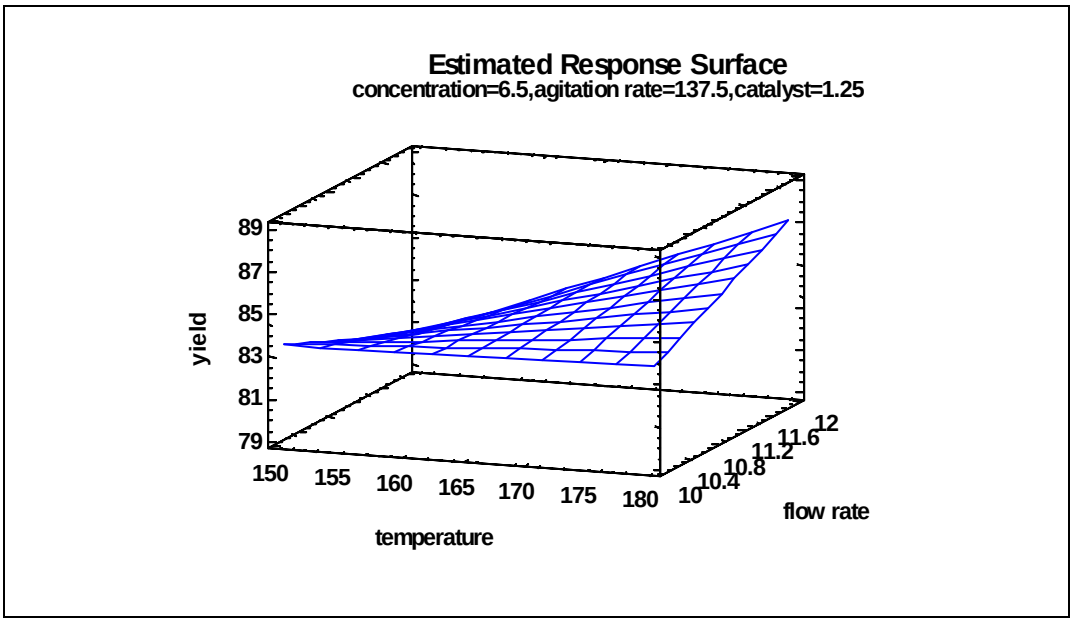


Figure 16-24. Response Surface Plot

In this plot, the height of the surface represents the predicted value of *yield* over the space of *temperature* and *flow rate*, with the other three factors held constant at their middle values. Highest yields are obtained at high temperature and high flow rate.

The type of plot and the factors over which the response is plotted can be changed using *Pane Options*:

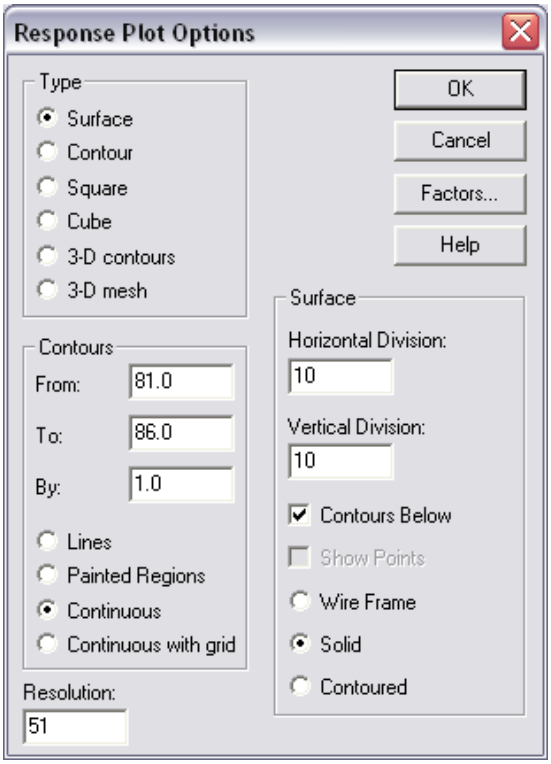


Figure 16-25. Pane Options for Response Plots

The types of plots that may be created are:

1. *Surface*: plots the fitted equation as a 3-D surface with respect to any 2 experimental factors. The surface may be a *wire frame*, a *solid* color, or show *contours* levels for the response. *Contours below* includes contours in the bottom face of the plot.

2. *Contour*: creates a 2-D contour plot with respect to any 2 experimental factors. Contours may be shown as *lines*, as on a topographical map, as *painted regions*, or using a *continuous* color ramp.

3. *Square*: plots the experimental region for any 2 experimental factors and displays the predicted response at each corner of the square.

4. *Cube*: plots the experimental region for any 3 experimental factors and displays the predicted response at each corner of the cube. To create this plot, you must first press the *Factors* button and select a third factor.

5. *3-D contours*: draws contours for the response with respect to 3 experimental factors simultaneously.

6. *3-D mesh*: creates a mesh plot showing the value of the response variable throughout a 3-dimensional experimental region.

The *Factors* button is used to select the factors that define the axes of the plots and the values at which other factors will be held:

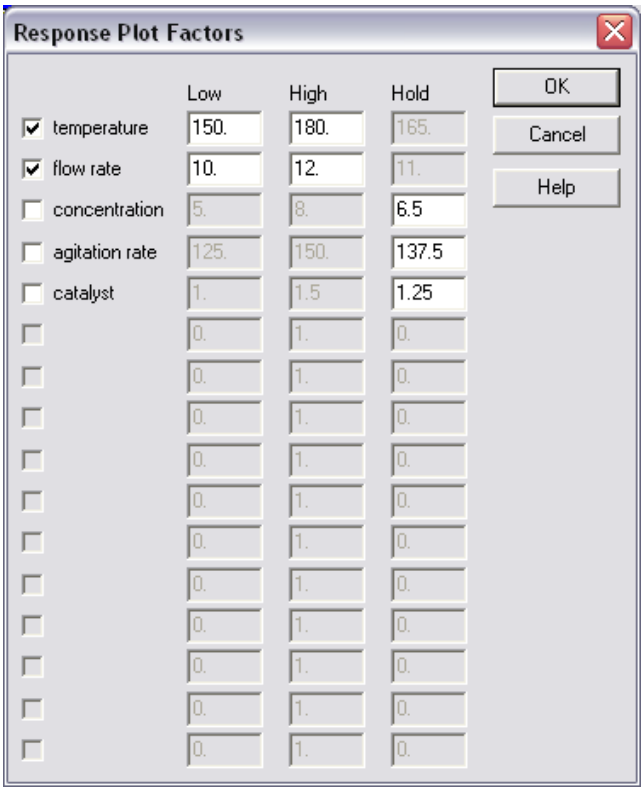


Figure 16-26. Response Plot Factor Options Dialog Box

To create the plot below, the *Contours* field has been set to *Painted*, the *Surface* to *Solid* with *Contours Below*, and the contours have been scaled to range from 81 to 86 by 1:

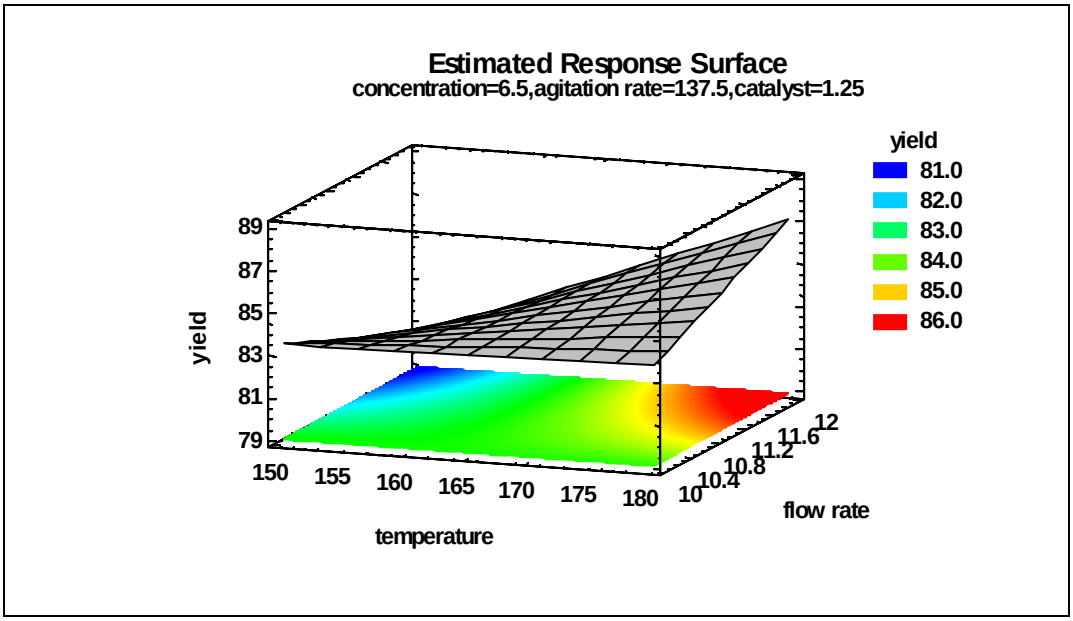


Figure 16-27. Response Surface Plot with Contours Below

The same plot can be displayed as a contour plot rather than a surface plot:

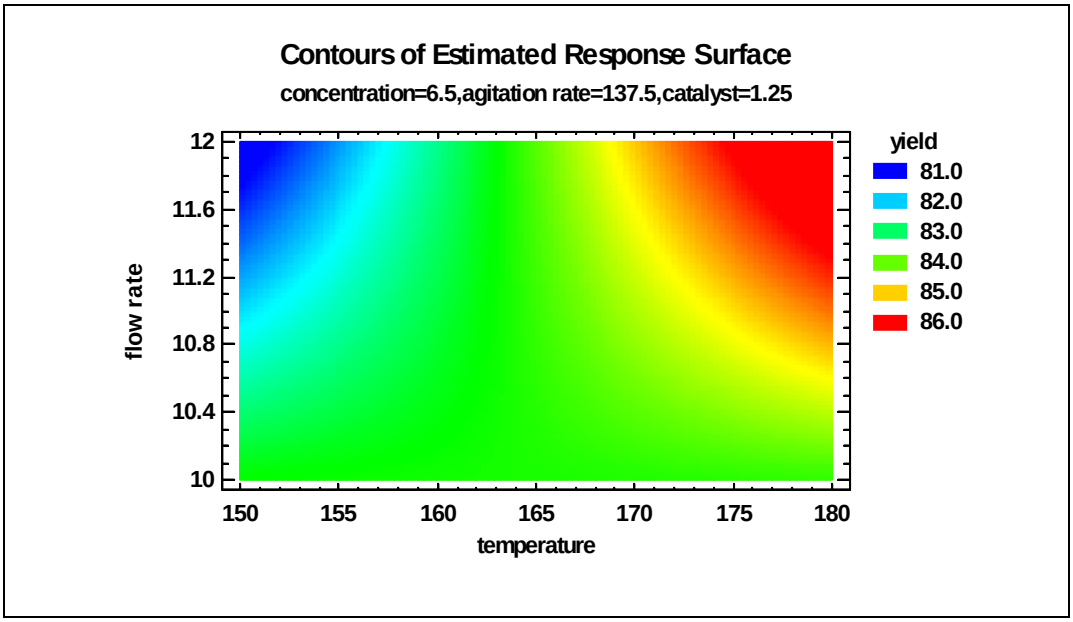


Figure 16-28. Contour Plot of Response Surface for Yield

High values of *yield* are obtained in the upper right corner.
The second response variable measured during the experiment was *strength*. The analysis window for *strength* displays the following Pareto chart:

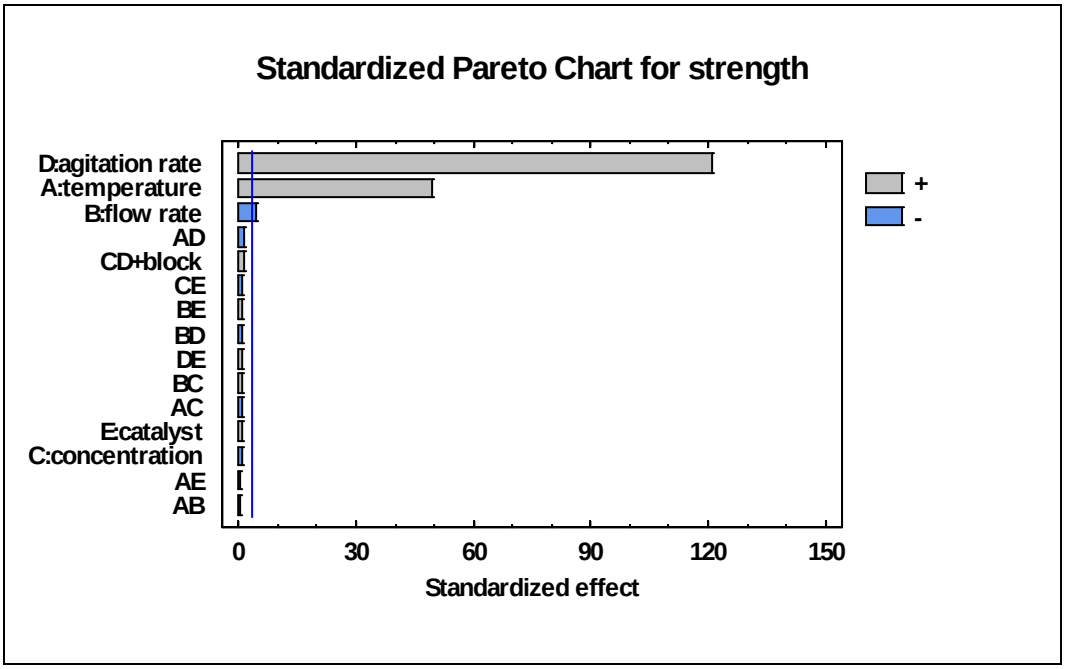


Figure 16-29. Standardized Pareto Chart for Strength

After taking out the insignificant effects, the fitted model is:

strength = -317.288 + 1.02083*temperature - 1.3125*flow rate + 3.005*agitation rate

Note that *agitation rate* impacts *strength*, although it did not have a significant effect on *yield*. The contour plot for the two strongest factors is shown below:

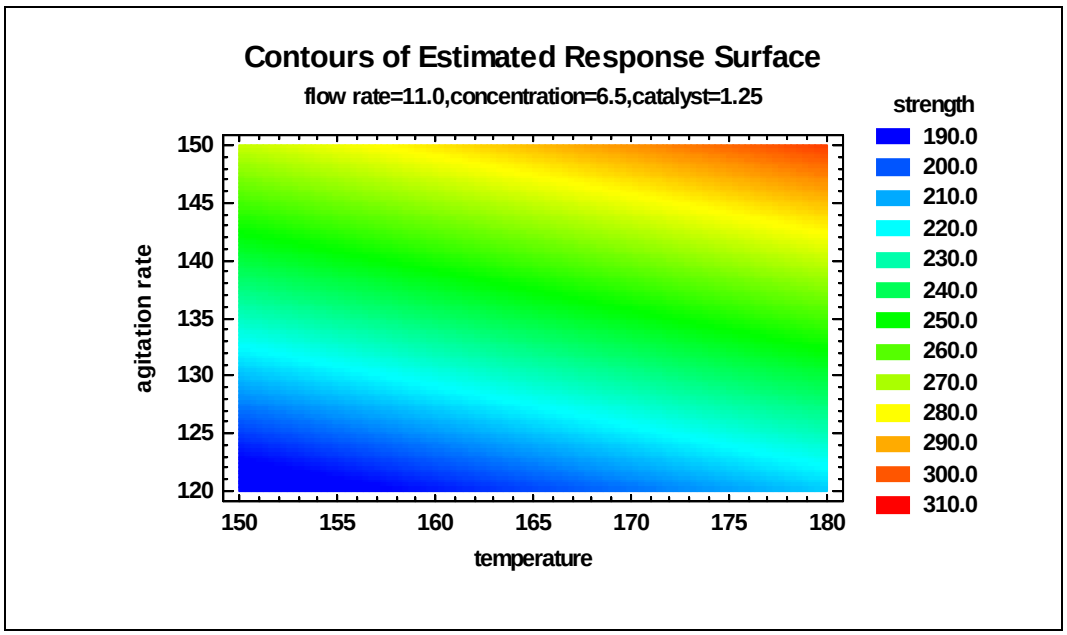


Figure 16-30. Contour Plot of Response Surface for Strength

Step 9: Optimize responses

Having built statistical models for both responses, optimal settings of the factors can now be determined. Recall that that the goal of the experiment was to maximize *yield* while keeping *strength* as close to 250 p.s.i. as possible. If you press the button labeled *Step 9*, the following dialog box will be displayed:

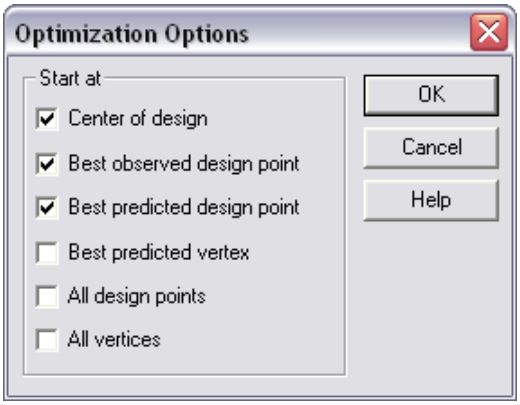


Figure 16-31. Optimization Options Dialog Box

Since the program will use a numerical search to find the best location within the experimental region, it is a good idea to start the search from several points to avoid finding a local optimum.

Press *OK* to start the search. After a few moments, the following message will be displayed:

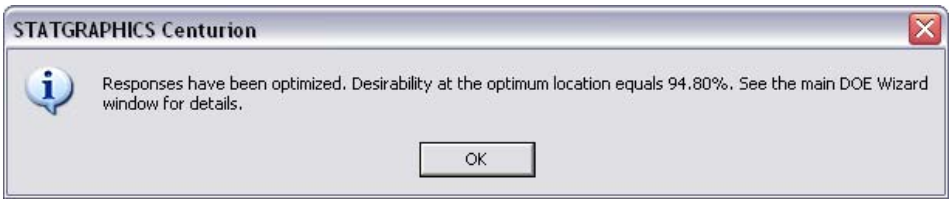


Figure 16-32. Message Displayed After Optimization Has Been Completed

At the same time, the following output will be added to the DOE Wizard's main window:

Step 9: Optimize the responses

Response Values at Optimum

Response	*Prediction*	*Lower 95.0% Limit*	*Upper 95.0% Limit*	*Desirability*
yield	88.6734	78.5663	98.7804	0.867338
strength	250.0	212.56	287.44	1.0

Overall desirability = 0.948027

Factor Settings at Optimum

Factor	*Setting*
temperature	179.999
flow rate	12.0
concentration	8.0
agitation rate	132.875
catalyst	1.5

Figure 16-33. Optimization Summary Added to Main DOE Wizard Window

At the indicated settings of the factors, it is estimated that *yield* will equal 88.67 grams while *strength* will equal 250 p.s.i. The resulting *yield* has a "desirability" quotient equal to 0.867, since it is 86.7% of the way between the specified range of 80 to 90 grams. *Strength* has a desirability quotient equal to 1, since it is exactly on target. The overall desirability equals 0.948, which is calculated by taking the desirability of each response, raising it to the power specified by its *impact*, multiplying the results together, and then raising the product to a power equal to 1 divided by the sum of the impacts. The result is a number between 0 and 1, with more weight given to the response with the higher impact.

If you press the *Tables and Graphs* button on the analysis toolbar, you may create two additional plots. The *Overlaid Contour Plots* shows contours of the two response variables overlaid on each other:

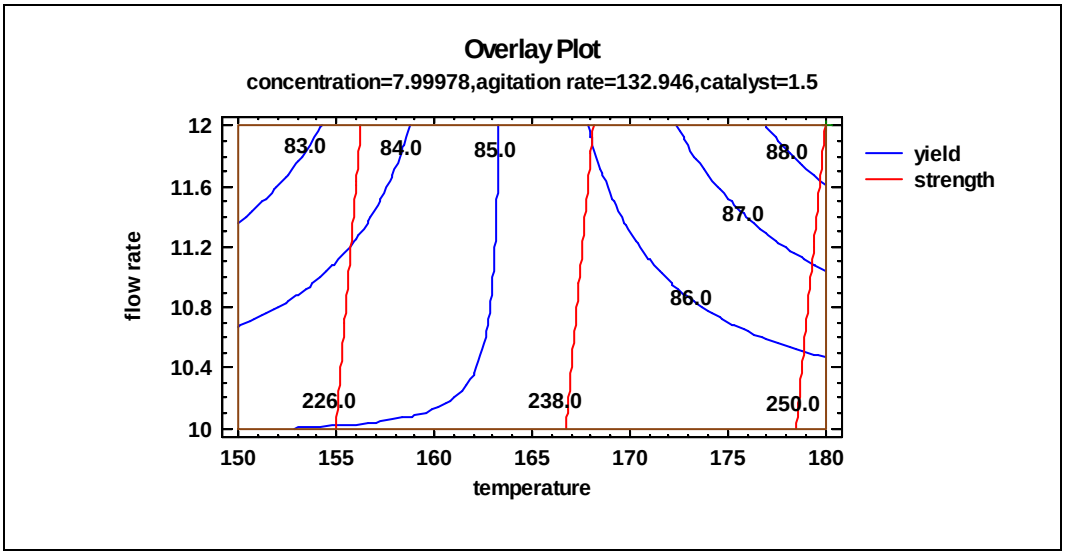


Figure 16-34. Overlaid Contour Plots for the Two Responses

The optimal point is at the upper right corner, where *yield* is maximized along the line for *strength* = 250. The *Desirability Plot* can be used to display the overall desirability versus two or three factors at a time. Selecting a 3-D mesh plot creates the following display:

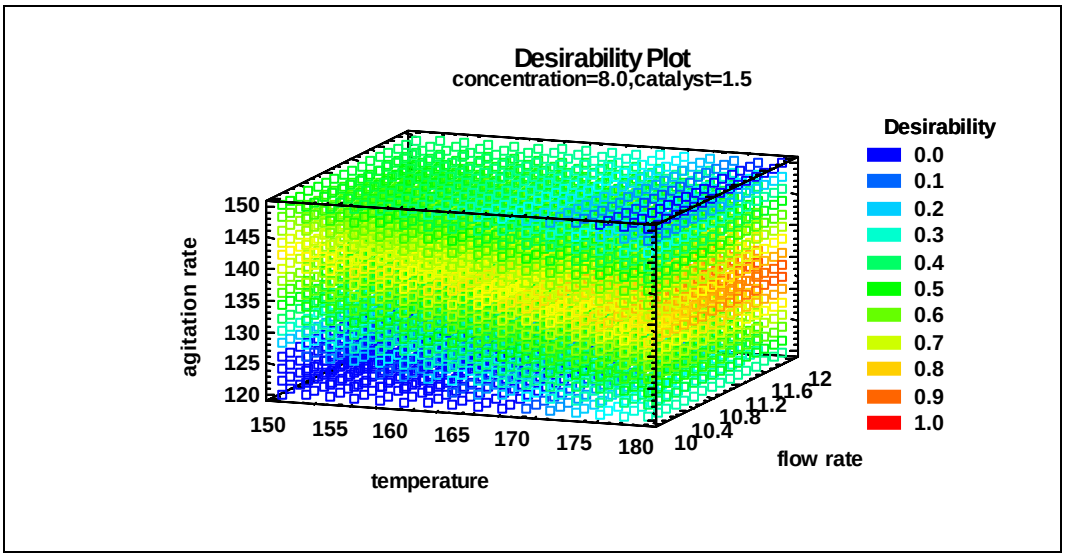


Figure 16-35. 3-D Mesh Plot of Overall Desirability

The best location is shown in red, where both *temperature* and *flow rate* are high, while *agitation rate* is held at a medium value.

Step 10: Save results

To save the results of the analysis and optimization, press the button labeled *Step 10* to save the results in a StatFolio:

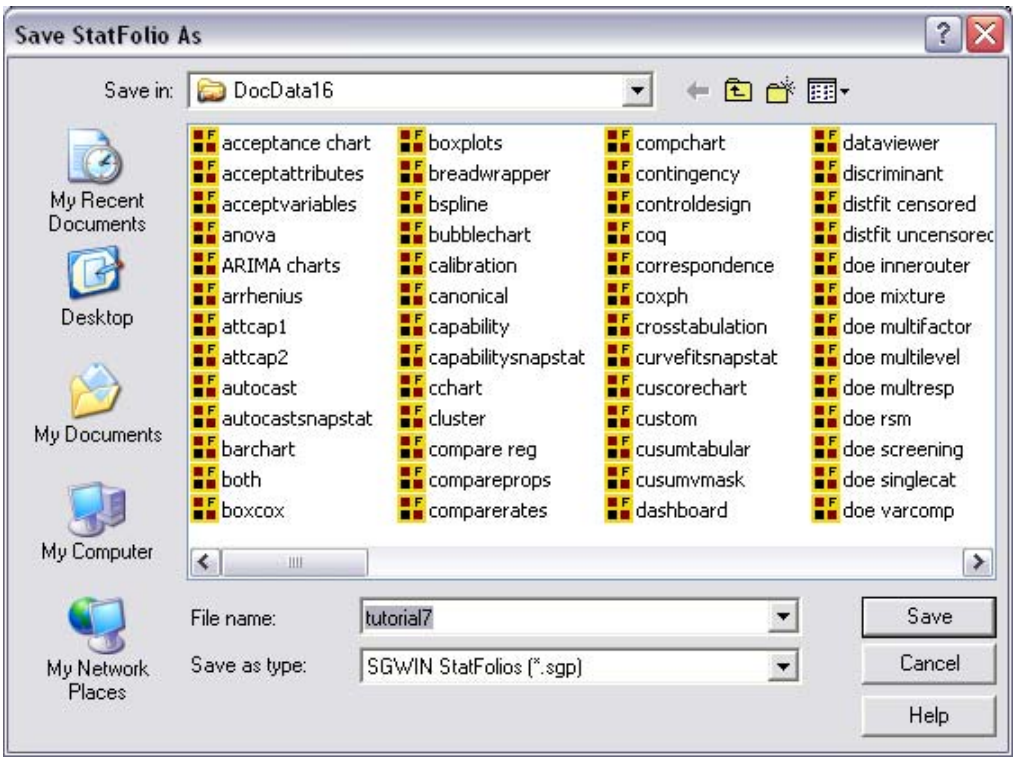


Figure 16-36. Save Results Dialog Box

16.3 Further Experimentation

If further experimentation is desirable, STATGRAPHICS Centurion XVI can help by either augmenting the existing design or generating points along the path of steepest ascent.

Step 11: Augment design

If you press the button labeled *Step 11*, you can add additional runs to the current experiment. It begins by displaying the dialog box shown below:

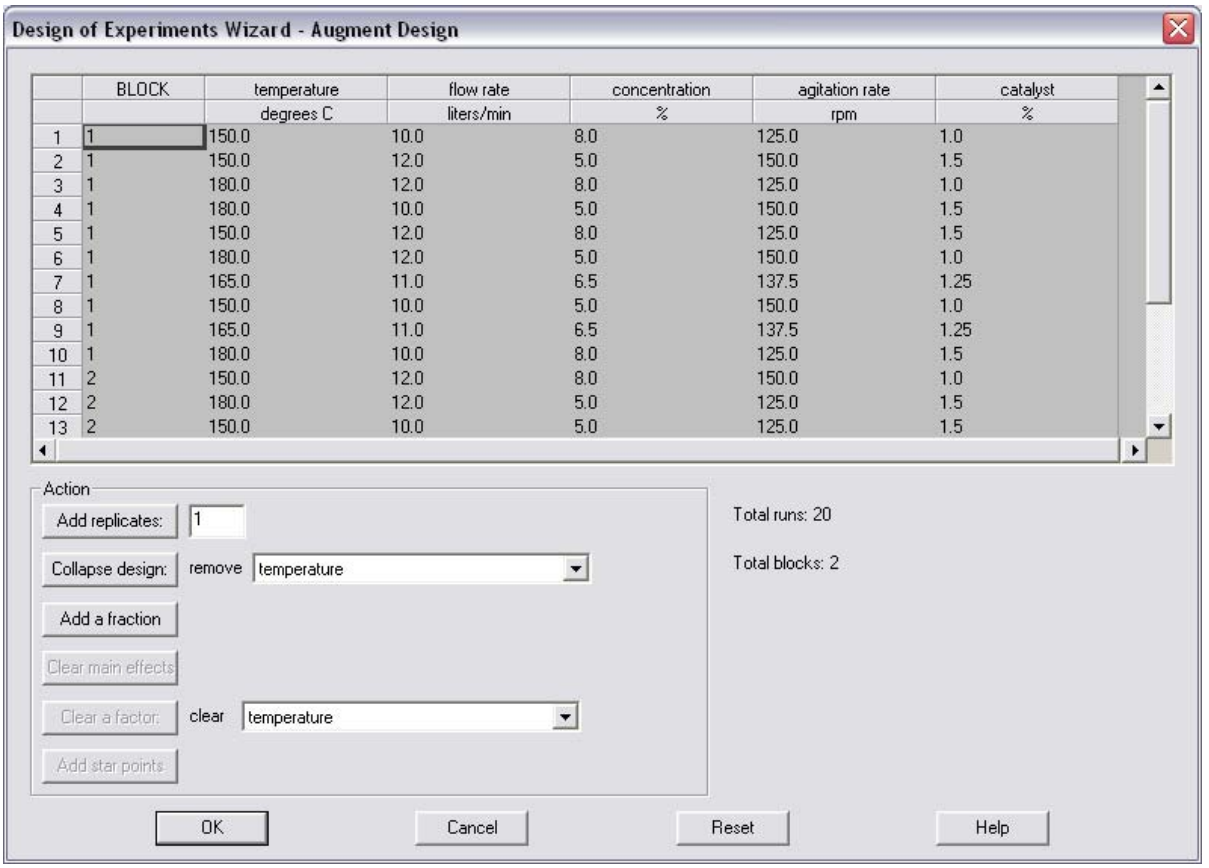


Figure 16-37. Augment Design Dialog Box

Three options are available:

1. *Add replicates:* adds another 20 runs to the design that are identical to the first 20. This would create more degrees of freedom for estimating the experimental error.

2. *Collapse design:* removes a specified experimental factor from the design and the resulting analyses.

3. *Add a fraction:* adds 20 more runs to make the design a full factorial.

Step 12: Extrapolate

You can also generate points along the *path of steepest ascent* in an attempt to move quickly to regions of higher yield. At a specific point in the experimental region and moves in the direction that exhibits the greatest change in the estimated response for the smallest changes in the experimental factors. Following that path can be very effective in obtaining dramatic improvements very quickly.

When you press the button labeled *Step 12*, the following dialog box is displayed:

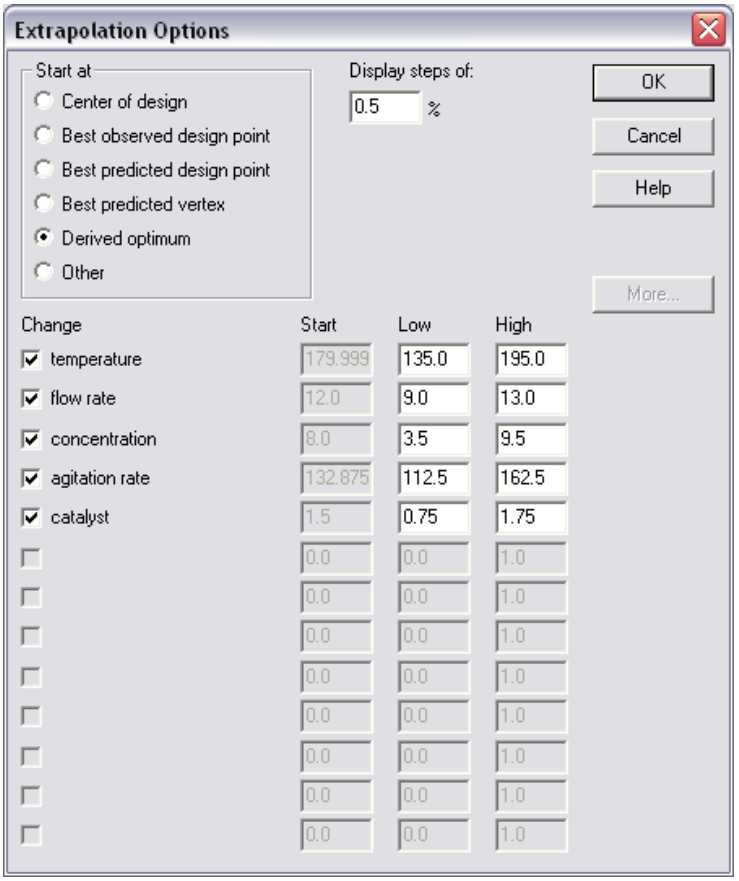


Figure 16-38. Extrapolate Dialog Box

The dialog box settings shown above instruct the program to begin at the derived optimum and let the 5 factors vary between *low* and *high* values that double the width of the experimental region in each dimension. It is instructed to display combinations of the factors whenever the

estimated desirability changes by at least 0.5%. After pressing *OK*, the following table will be added to the DOE Wizard's window:

Step 12: Extrapolate model

Extrapolated Response Values

Step	*Desirability*	*yield*	*strength*
0	0.948018	88.6734	250.001
1	0.953289	88.81	249.974
2	0.959294	88.9548	249.987
3	0.964593	89.0926	249.97
4	0.970278	89.2319	249.984
5	0.975933	89.3718	249.997
6	0.981395	89.5155	250.012
7	0.986799	89.6578	250.018
8	0.991955	89.7919	249.985
9	0.997518	89.9366	250.008
10	0.999936	90.0113	249.995

Factor Settings for Extrapolation

Step	*temperature*	*flow rate*	*concentration*	*agitation rate*	*catalyst*
0	179.999	12.0	8.0	132.875	1.5
1	180.037	12.05	8.04539	132.875	1.50594
2	180.114	12.1	8.08944	132.875	1.51171
3	180.162	12.15	8.13255	132.875	1.51735
4	180.239	12.2	8.16866	132.875	1.52208
5	180.317	12.25	8.20457	132.875	1.52678
6	180.396	12.3	8.24335	132.875	1.53186
7	180.466	12.35	8.28218	132.875	1.53694
8	180.497	12.4	8.32249	132.875	1.54222
9	180.585	12.45	8.3575	132.875	1.5468
10	180.585	12.46	8.41739	132.875	1.55572

Figure 16-39. Extrapolation Summary added to Main DOE Wizard Window

It is estimated that the yield can be raised to its target value of 90 grams while maintaining a strength equal to 250 by increasing temperature to 180.6 degrees, increasing the flow rate to 12.46 liters per minutes, increasing concentration to 8.42%, and increasing catalyst to 1.56%. Since this is an extrapolation of the fitted statistical model outside of the original experimental region, confirmatory runs would need to be done to verify this result.

Suggested Reading

The following books are excellent, readable sources of information about the statistical techniques described in this guide:

Basic statistics: Applied Statistics and Probability for Engineers, 4th edition, by Douglas C. Montgomery and George C. Runger (2006). John Wiley and Sons, New York.

Analysis of variance: Applied Linear Statistical Models, 5th edition, by Michael H. Kutner, Christopher J. Nachtsheim, and John Neter (2004). McGraw Hill.

Regression analysis: Applied Linear Regression, 3rd edition, by Sanford Weisberg (2005). John Wiley and Sons, New York.

Statistical process control: Introduction to Statistical Quality Control, 6th edition, by Douglas C. Montgomery (2008). John Wiley and Sons, New York.

Design of experiments: Statistics for Experimenters: Design, Innovation and Discovery, 2nd edition by George E. P. Box, William G. Hunter, and J. Stuart Hunter (2005). John Wiley and Sons, New York.

Data Sets

93cars.sgd

This data was downloaded from the Journal of Statistical Education (JSE) Data Archive. It was compiled by Robin Lock of the Mathematics Department at St. Lawrence University and is used with his permission. An article associated with the dataset appears in the *Journal of Statistics Education*, Volume 1, Number 1 (July 1993).

bodytemp.sgd

This data was also downloaded from the Journal of Statistical Education (JSE) Data Archive. It was compiled by Allen Shoemaker of the Psychology Department at Calvin College and is used with his permission. The data were derived from an article in the *Journal of the American Medical Association* (1992, vol. 268, pp. 1578-1580) entitled "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich" by P. A. Mackowiak, S. S. Wasserman, and M. M. Levine. An article associated with the dataset appears in the *Journal of Statistics Education*, Volume 4, Number 2 (July 1996).

Journal of Statistical Education (JSE) Data Archive web site:

http://www.amstat.org/publications/jse/jse_data_archive.htm

Index