Statistix 8

User's Manual

Analytical Software
PO Box 12185
Tallahassee FL 32317-2185
(850) 893-9371
www.statistix.com

Publisher: Analytical Software ISBN 1-881789-06-3

P R E F A C E

Statistix is a very fast, easy-to-use data analysis program designed to encourage you to "play" with your data. Manipulating data becomes simple and straightforward, allowing you to focus on your research and not your software.

Capabilities of *Statistix*

- Descriptive statistics
- Nonparametric tests
- Linear regression
- Stepwise regression
- Logistic regression
- Analysis of variance/covariance
- T tests
- Survival analysis
- Residual analysis
- Association tests
- Probability functions
- Time series analysis
- Statistical process control
- Powerful transformations
- Graphs

Features of *Statistix*

- Compact—less than 4M disk space
- *Excel*, *1-2-3*, and *Access* support
- Interactive design
- On-line help
- Spreadsheet data editing
- Free technical support
- Fast
- Accurate

System Requirements

Statistix runs on *Windows* based personal computers. It requires *Windows 95* or a later version of *Windows.* A minimum of 16 MB of RAM is recommended.

What's Included

Statistix for Windows includes this manual, program CD, and a registration card. **Please mail us your registration card to receive:** (1) free technical support, (2) special upgrade prices, and (3) product announcements.

Technical Support

Should you need help using *Statistix,* call our office for free technical support at 850-893-9371 Monday-Friday, 9:00 a.m.-5:00 p.m. Eastern Time. You can fax us at 850-894-1134 or send email to support@statistix.com anytime.

Acknowledgements

We'd like to thank the people who supported the development of *Statistix* over the years with help and advice:

Mark Berenson
Baruch College

David Levine
Baruch College

Douglas Hawkins
University of Minnesota

Donald Richter
New York University

David Hosmer
University of Massachusetts

Harry Roberts
University of Chicago

Ben King
Florida Atlantic University

Herbert Spirer
University of Connecticut

Stanley Lemeshow
University of Massachusetts

Janet Wagner
University of Massachusetts

C O N T E N T S

C H A P T E R

1

Introduction

Welcome to *Statistix 8*, our latest data analysis software designed for the *Windows* operating system. This version introduces several new enhancements making *Statistix* more useful than ever. These include: a new GLM algorithm for improved analysis of unbalanced AOV designs, Latin Squares, Balanced Lattice, Fractional Factorials, Two-Stage Least Squares Regression, and Stepwise Logistic Regression.

The emphasis in the design of *Statistix* has always been to make it quick and easy to obtain concise reports that answer your data analysis questions. Whether you are a professional statistician or a researcher with your own data to analyze, whether you do data analysis every day or only occasionally, you'll find that *Statistix* will quickly and easily help you find the answers you're looking for.

Statistix is fast, compact, and accurate. The user's manual is clear and concise, complete with examples and references.

Fans of *Statistix* love it because it's intuitive and easy to use. *Statistix* encourages the kind of creativity that marks the difference between good and routine data analysis. By its very design, *Statistix* invites you to be adventurous—to explore, play around, and get to know your data.

So jump in and get started.

Using This Manual

If you're like most people, you'll run the software before reading the manual. That's OK with us because *Statistix* is far easier to use than it is to read about. However, if you haven't tried *Statistix* yet, install the software now by following the directions in the Installing *Statistix* section on the next page. Explore the *Statistix* menus and play around with the *Statistix* dialog boxes used to run statistical procedures. You can open one the sample data files supplied with *Statistix* (cholesterol.sx is good one to start with) or enter a small data set of your own. Get a feel for the program; it'll make it easier to understand the manual.

Once you've experimented with the software, read the rest of Chapter 1. The section titled The *Statistix* Menus describes how to use the menus. Two sections—Data in *Statistix* and Getting Data In and Out of *Statistix*—give an overview of how data are handled. *Statistix* Dialog Boxes is an important section that describes how to make efficient use of the dialog boxes used for model specification.

The Preferences procedures are discussed at the end of this chapter. Use them to specify your preferences for variable list order, date formats, graph colors, and other options.

As you develop your *Statistix* skills, make it a high priority to read Chapter 2 on the Data Menu and Chapter 3 on the File Menu. Chapters 4 through 13 describe the statistical analysis procedures available in *Statistix*—Summary and Descriptive Statistics; One, Two, & Multi-Sample Tests; Linear Models; Analysis of Variance; Association Tests; Randomness/Normality Tests; Time Series; Quality Control; Survival Analysis; and Probability Functions.

It's a good idea to at least skim these chapters so that you're aware of the range of *Statistix*' capabilities. If you come across statistical procedures with which you're unfamiliar, study the examples and references until you have a general understanding of when the analyses would be useful. The details of how the analyses are performed are unimportant; you can always look them up when needed. However, to fully utilize *Statistix*, you need to know which tools to apply to which tasks.

The *Statistix* manual provides useful background. Used in conjunction with appropriate references, it's a valuable learning tool.

Installing Statistix

The *Statistix* software comes on one CD. You can't run *Statistix* directly from the distribution CD. You must install *Statistix* on a fixed disk (hard disk). You must be running *Windows 95* or later to run the *Statistix* installation program.

Insert the *Statistix* CD into your computer. On most computers, the installation program will start automatically. If it doesn't, use the Run command to start the installation program. Type **e:setup** in the Open box of the Run dialog. (If your CD is not e:, substitute the correct letter.) Follow the on-screen instructions to complete the installation. A *Statistix* folder is created to store the program files (usually Statistix) and a *Statistix* group is created on the Program folder of your *Windows* Start menu.

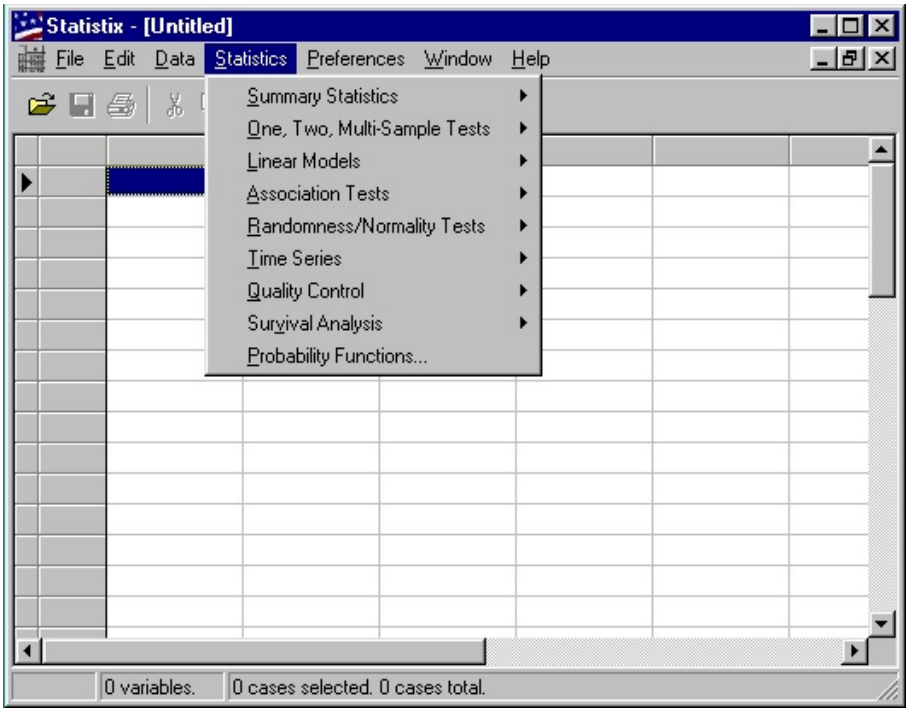
To run *Statistix*, click on the *Statistix* icon. You can view the ReadMe file for additional information by clicking on the ReadMe icon.

The Statistix Menus

When you first run *Statistix*, an empty spreadsheet is displayed and the main menu appears above it. The main menu, or the spreadsheet menu, is visible whenever the spreadsheet window is the active window. All of the items on the main menu are themselves menus. These menus offer a variety of data management and statistical procedures. An example empty spreadsheet and menu appears on the next page.

To select a pull-down menu, use your mouse to point and click on the name of the pull-down menu. You can also select a pull-down menu by holding down the Alt key and pressing the character underlined in the menu name (e.g., Alt-F for File). Point and click to select a menu item from the pull-down menu or press the letter underlined in the name.

The **File** menu includes procedures to open and save *Statistix* data files, and to import data from other programs. The **Data** menu offers a number of procedures to manipulate the spreadsheet data including a powerful



Transformations procedure. The **Statistics** menu lists several topic menus as shown above offering both basic and advanced statistical analyses. The **Window** menu lists the windows within *Statistix*, which include the spreadsheet window and results windows. Use the Windows menu to switch between the windows in *Statistix*.

Exit *Statistix* by selecting the **Exit** procedure from the File menu, or by clicking on the close button in the upper-right corner of the window.

Before attempting to use a statistical procedure, you must either create a data set using the Insert procedure on the Data menu, or retrieve a data set from disk using either the **Open** or **Import** procedure on the File menu.

In addition to the statistics menus listed above, the regression, analysis of variance, and several of the time series procedures display results menus after the initial analysis is specified and computed offering additional analyses.

Data In Statistix

In this section, we give an overview of how data are handled once entered into *Statistix*. The following section describes ways to get your data into *Statistix*. More details about data handling are given in Chapter 2.

Variables and Variable Names

Data in *Statistix* can be viewed as a rectangular table of values. The columns of the data table are called variables, and the rows are called cases. All data in *Statistix* are referenced by variable names that you assign when data are entered. A variable name is one to nine characters in length, must begin with a letter, and can only consist of letters, digits, and the underscore character. You should assign meaningful variable names to help you remember what they represent. There are a few words reserved for other tasks, such as CASE, M, PI, and RANDOM, that you cannot use as variable names.

Variable names are used to manipulate the data. For example, a new variable VOLUME can be created from the variables HEIGHT and BASE using the **Transformations** procedure as follows:

```
VOLUME = PI * HEIGHT * SQR (BASE)
```

PI and SQR are examples of built-in functions available in Transformations, which we'll discuss in detail in Chapter 2.

Variable names are used to specify the source of data for statistical analyses. For example, to specify the regression of HEAT on CHEM1 and CHEM2 using the Linear Regression procedure, select the name HEAT for the dependent variable. Then select the names CHEM1 and CHEM2 for the independent variables.

Data Types

Statistix can handle four types of data: real, integer, date, and string. A variable can only contain values of one data type. The data type of a variable is established when you create the variable and can be changed to a different data type using the Transformations procedure.

The "real" data type is used to represent floating point numbers (e.g., 1.245). This format is the most flexible offered by *Statistix* and is used as the default data type when creating new variables.

Integer data in *Statistix* are whole numbers in the range -32767 to 32767. This data type uses only 25% as much space as the real data type. You can use the integer data type instead of the real data type, when appropriate, to increase the data set capacity of *Statistix*. This will also save disk space by reducing the size of *Statistix* data files.

The "date" data type is used to represent dates (e.g., 12/31/1992). See **General Preferences** on page 15 for an option to select the order of month, day, and year.

The "string" data type is used to enter alphanumeric data, such as a subject's name. String variables can be used as grouping variables for statistical procedures that compute results by group.

Cases and Data Subsets

The rows in the rectangular data table are called cases. The cases are numbered sequentially. The case numbers are listed on the left side of the spreadsheet window. The Case function in **Transformations** and **Omit/Select/Restore Cases** provides a method to refer to the case numbers.

Sometimes you'll want to temporarily work with a subset of all data cases. The **Omit/Select/Restore Cases** procedure can be used to "hide" specified cases from the system. The subset selection is based on a condition that you specify.

```
OMIT IF (HEIGHT > 5) AND (WEIGHT < 100)
```

Until specified otherwise, *Statistix* only "sees" cases not omitted using the omit statement. The cases are not deleted, but hidden. You can easily restore the hidden cases anytime. Further details on Omit/Select/Restore Cases are given in Chapter 2.

Data Set Size

A *Statistix* data set is limited to 500 variables and 200,000 cases. The active *Statistix* data set must completely fit in the available memory. There is a maximum file size of 32 MB, which translates to a spreadsheet with about 4 million cells. Variables and cases compete for space in the data table. The more cases you have in a data set, the fewer variables you can add.

Other programs running compete with *Statistix* for your computer's memory. You can free up memory for use by *Statistix* by closing other applications.

Missing Values

When data are entered into *Statistix*, a value of "M" is used to show a missing value. When data are displayed, a missing value is displayed as an "M". The M function available in Transformations and Omit/Select/Restore Cases is used to assign missing values and to make comparisons to missing values.

All *Statistix* procedures examine the data for missing values and treat them appropriately. If arithmetic is performed on a variable containing missing values using Transformations (e.g., A = B + C), the result of the equation will be missing for the cases that contain the missing values. When a statistic, such as the mean of a variable, is calculated, only the non-missing values for the column are used to compute the result. The Linear Regression procedure will drop cases that contain a missing value for either the dependent variable or any of the independent variables.

Getting Data In and Out of Statistix

You can use three methods to enter data into *Statistix*:

1) Keyboard data entry
2) Text, *Excel*, *Lotus 1-2-3*, *Quattro Pro*, *Access*, *dBase*, or *Paradox* files
3) *Statistix* data files.

Keyboard data entry is performed directly on the spreadsheet window. You create new variables using the **Insert Variables** procedure found on the Data menu (Chapter 2). Keyboard data entry is often preferred when the amount of data being entered is small.

Statistix can read text files (also called ASCII files) created using a word processor, a spreadsheet program, or some other PC program. *Statistix* can also read *Excel*, *Lotus 1-2-3*, and *Quattro Pro* spreadsheet files, and *Access*, *dBase* and *Paradox* files. Text files provide a standard data exchange format between *Statistix* and other programs. Use the *Statistix* **Import** procedure to create or augment data sets using data stored in text, spreadsheet, or database files. Likewise, *Statistix* data can be exported to a variety of programs using the **Export** procedure.

While running *Statistix*, your data set is temporarily stored in random access memory (RAM). Before you exit *Statistix*, you should save your data set using either the **Save** or **Save As** procedure. A *Statistix* data file is a "snapshot" of *Statistix*' data memory, and it's an ideal method of storing data sets for future *Statistix* analyses. The advantages of *Statistix* data files are that they can be read and written very rapidly; are compact in terms of the disk space occupied; and preserve such *Statistix* information as variable names, labels, and the case omit status. *Statistix* data files are described in more detail in Chapter 3. *Statistix* data sets are retrieved using the **Open** procedure.

A *Statistix* data set is dynamic. You can increase or decrease its size in a variety of ways. For example, you can delete cases and variables when you no longer need them to conserve space. New cases and variables can be added from the keyboard, imported from text and spreadsheet files, or merged from *Statistix* data files. You'll often use **Transformations** to create new variables. Some of the statistical procedures can produce new variables, too. For example, **Linear Regression** can save residuals and predicted values as new variables. A *Statistix* data file can be saved at any time.

Statistix 8 can open data files created by earlier versions of *Statistix*. However, earlier versions of *Statistix* can't open *Statistix 8* files. The **Save As** procedure includes an option to save data using an older file format for backward compatibility (see Chapter 3).

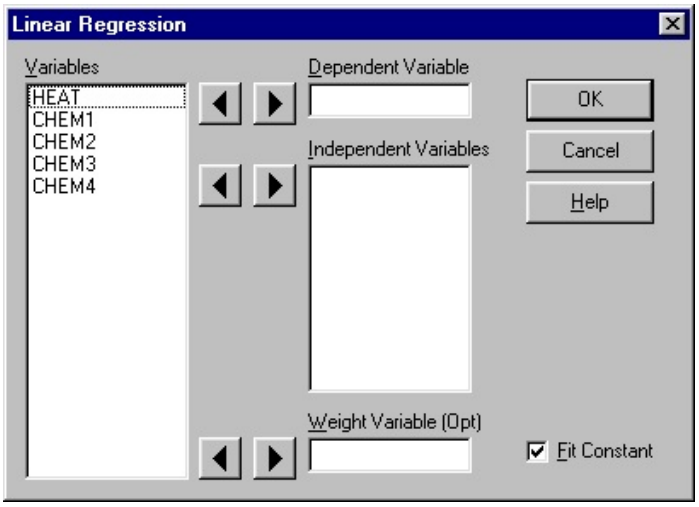
Statistix Dialog Boxes

Once you select a procedure using the menus, a dialog box is displayed on your screen. A dialog box is a window you use to instruct *Statistix* on the details of a data management operation or statistical procedure.

Statistix dialog boxes look like the dialog boxes you've seen in other *Windows* applications. They contain the familiar buttons, check boxes, list boxes, and edit controls.

To illustrate, we'll examine the dialog box for the Linear Regression

procedure displayed below. Like most *Statistix* dialog boxes, this one lists the data set variables in a list box with the heading ***Variables***. You specify the regression model by moving variable names from the Variables list box to the ***Dependent Variable***, ***Independent Variables***, and ***Weight Variable*** list boxes.



The example dialog box also has a check box with the title ***Fit Constant***. You check and uncheck the box by pointing to the box using your mouse and clicking the mouse button. The model will fit the constant when the box is checked.

Some *Statistix* dialog boxes include radio buttons and edit controls. Radio buttons are a group of buttons where you must select one and only one of the choices available. An edit control is a box that you use to enter text.

Once you've specified the model, press the ***OK*** button to compute and display the results. You can exit a dialog box without performing the analysis by pressing the ***Cancel*** button. Press the ***Help*** button to display context sensitive help.

Variable Name Selection

Most *Statistix* procedures require that you specify which variables are to be used to complete the analysis. You do this by moving variable names from the Variables list to one or more "target" list boxes. In the Linear Regression dialog box above, there are three target boxes: Dependent Variable, Independent Variables, and Weight Variable.

First, highlight a variable name in the Variables list box: Point to the variable name using the mouse and click the mouse button once. The background color changes to show that the variable is highlighted. You then press the right-arrow button next to the target box to which you want to move the variable. The name is deleted from the Variables list and is added to the target box.

You can move more than one variable at a time. To highlight several variables, press and hold down the Ctrl key, and click each variable you want to select. Once you've highlighted all the variables you want, press the right-arrow button to move them.

You can highlight a range of sequential variables at once. Click the first variable you want to select, and then drag the cursor to the last item you want to select. The entire range of variables is highlighted and can be moved by pressing an arrow button. The order that the variables are listed in the Variables list box affects the usefulness of this feature. You can have variables listed in alphabetical order or in spreadsheet order (see page 15).

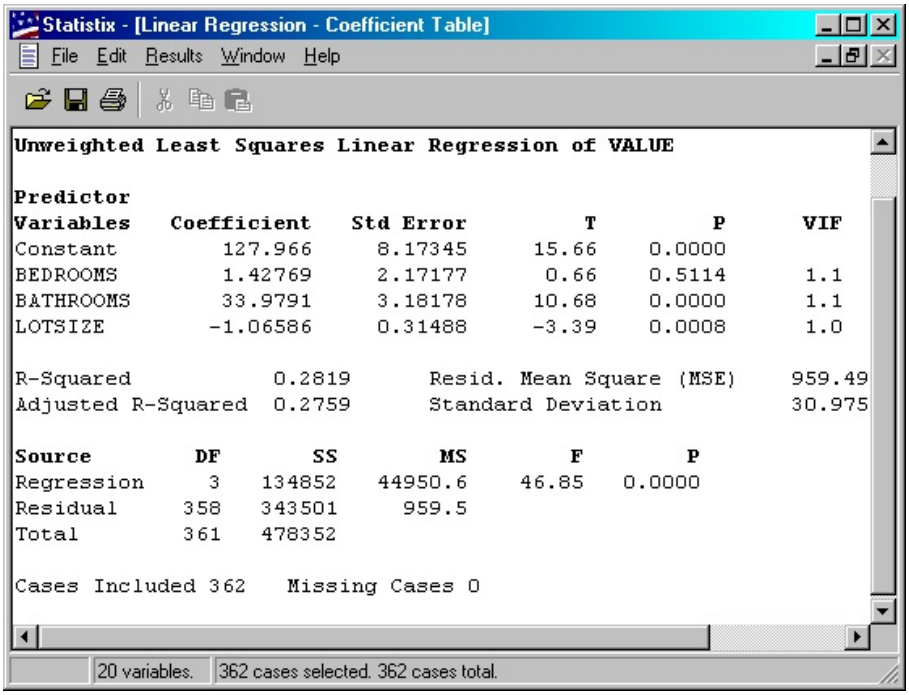
In some situations you can select and move a variable from one list box to another by double-clicking on the name. This is a quick way to move a variable, but it can only be used when there's only one possible destination for the variable selected. For example, double-clicking a variable in the Variables list box in the Linear Regression dialog box on the preceding page doesn't move the variable because there are three possible destination list boxes. However, you can move a variable from the Independent Variables list box to the Variables list box by double-clicking on its name, since the Variables list box is the only possible destination.

Saving Dialog Boxes

The variable lists, file names, and other details that you enter on a dialog box are automatically saved when you press the OK button. When you select a procedure for the second time, the data you entered previously automatically reappear. You can press OK and rerun the same analysis, or you can make changes to the dialog box. When you save a data set using the **Save** procedure, the dialog box details are saved with the spreadsheet data. When you open the data file later, the dialog box options specified earlier are available.

Results Window

When you specify an analysis using a dialog box and then click the OK button, the analysis is computed. The resulting table or graph is displayed in a "results window". The results window for a linear regression analysis is displayed below.



Like other windows, the results window can be resized by dragging the sides or corners of the window. It has minimize, maximize/restore, and close buttons located in the upper-right corner. When the results can't fit in the window, scroll bars appear that you can use to scroll through the report. You can also scroll the window's contents using the arrow and page keys.

When you've finished with a results window, you'll want to close the window (although you can leave it open for later reference). You close the results window by pressing the close button located in the upper-right corner, by double-clicking the control button located in the upper-left corner, or by selecting Close from the File menu.

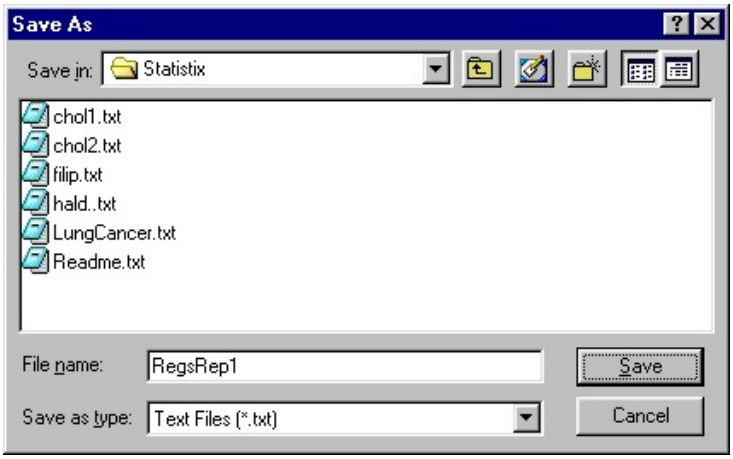
Results windows have their own "results menus" as can be seen in the example above. The results menu replaces the main spreadsheet menu

when a results window becomes the active window. The results menu has five submenus: File, Edit, Results, Window, and Help. The **File** menu has Print and Save As commands you can use to print and save reports and graphs. The **Edit** menu lets you copy reports and graphs to the *Windows* clipboard. The **Results** menu gives you direct access to the dialog box used to generate the report; options for changing a displayed graph; and for some statistical procedures, options for further analysis. The **Window** menu lists the windows currently open in *Statistix*, and can be used to switch control to a different *Statistix* window. The **Help** menu accesses the *Statistix* on-line help.

Printing and Saving Reports

There are two ways to print a report or graph displayed in the active window. You can click on the printer icon on the toolbar, in which case the report is printed immediately using your printer's current settings. You can also select **Print** from the File menu, in which case the print dialog box is displayed. This way you get a chance to select a different printer or change printer settings such as the page orientation. You can also change printer settings using the Printer Setup procedure found on the main File menu.

A results window can contain either a report or a graph. You can save both reports and graphs by clicking on the diskette icon on the toolbar, or by selecting the **Save As** command on the File menu. A Save As dialog box appears, as shown below.



Reports can be saved as text files or rich text files (RTF) . Text files are plain ASCII files. RTF files contain text formatting information including

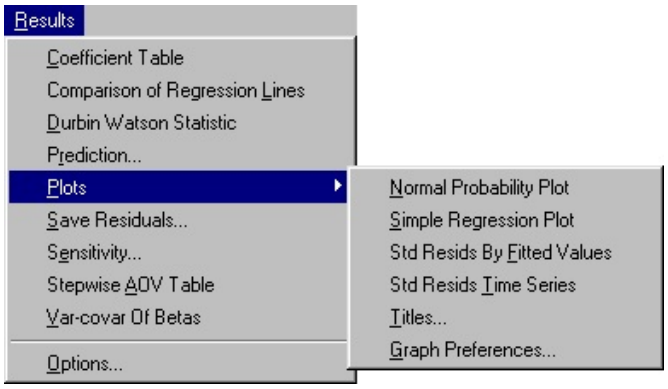
font names and sizes. Both formats are standard file formats used by many *Windows* programs including *Wordpad*, *WordPerfect*, and *Word*. Reports saved as text or rich text files can be viewed and printed using the **View Text File** procedure found on the *Statistix* File menu (see Chapter 3).

Graphs can be saved using a choice of graphics file formats: *Windows* Metafile (WMF), Enhanced Metafile (EMF), or *Windows* bitmap (BMP). The WMF and EMF formats produce compact files that can be imported by many *Windows* programs including word processing and spreadsheet programs. Bitmap files are generally larger, but are supported by many programs including the *Windows Paint* program.

There are two ways to select a graphics format when saving a graph. One way is to select the format from the pull-down list titled ***Save as file type***. You can also specify the format you want by including the corresponding file name extension (.EMF, .WMF, or .BMP) when you enter the file name.

The Results Menu

When you click on the Results menu for the Linear Regression results window pictured on page 11, the resulting pull-down menu offers several opportunities for further analysis.



The menu choices offer additional tables of results, plots, and the option to save residuals. Most of the statistical procedures don't offer further analysis, but all of the Results menus contain the last menu item shown: Options. Selecting Options from a Results menu brings back the dialog box used to specify the analysis. This gives you direct access to the dialog box and a convenient way to make modifications to the model specified.

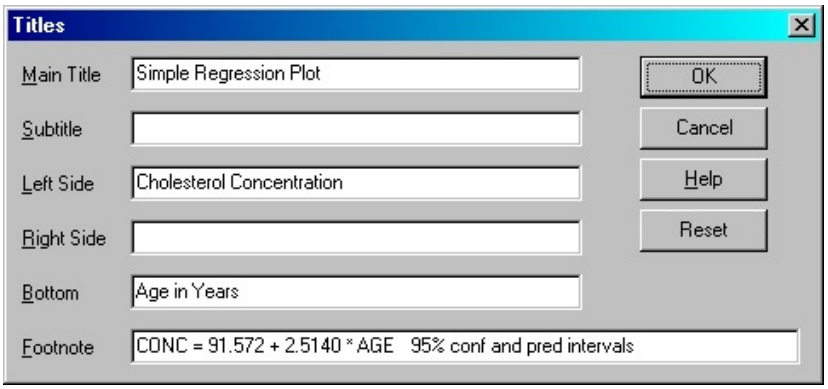
Be sure to look at the Results menus or you may be overlooking

opportunities for further analysis. The procedures that offer further analysis after the initial results are displayed are listed below:

All analysis of variance procedures	Logistic Regression
Exponential Smoothing	Moving Averages
Kaplan-Meier	Poisson Regression
Kruskal-Wallis One-Way AOV	Proportional Hazards Regression
Linear Regression	SARIMA
Log-Linear Models	Two-stage Least Squares Regs

Graph Titles

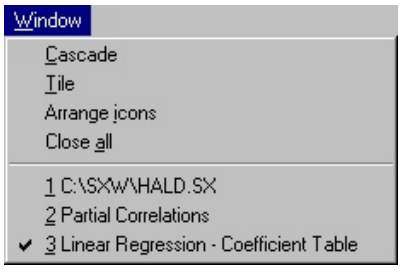
Procedures that produce a graph have an additional item on the Results menu that you can use to modify the titles that appear on graphs. Selecting **Titles** from the results menu displays the Titles dialog box as shown below.



You can edit, add, and delete titles from the graph. Pressing the ***OK*** button redisplays the graph with the new information. Pressing the ***Reset*** button resets all the titles back to those generated automatically by *Statistix*.

Switching Between Windows

The Window menu that appears on the results menu, as well as the main spreadsheet menu, lists all of the *Statistix* windows.



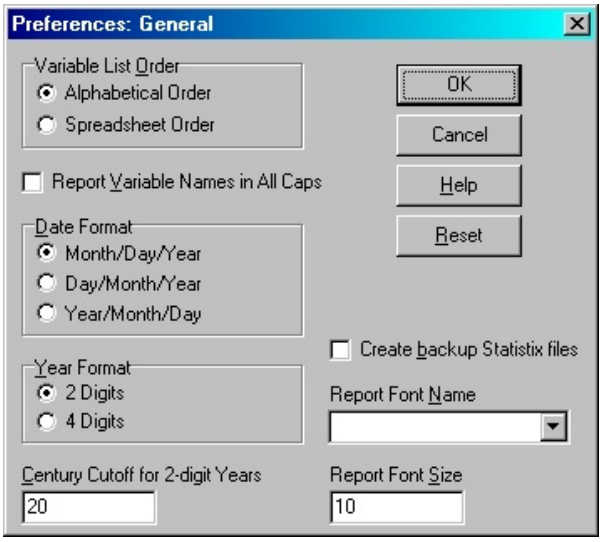
One of the *Statistix* windows listed is the spreadsheet window (HALD.SX in the example on the preceding page). Select the spreadsheet window from the menu to gain access to the spreadsheet menu containing the Data and Statistics menus.

Preferences

The **Preferences** procedures are used to configure *Statistix* to your taste. You can select variable order, date format, and graph colors. Once you've made your selections, they remain in effect until you change them again.

General Preferences

Use the General Preferences procedure to select variable order, date format, and report font attributes.



Statistix displays the variable names of the open data file in a list box for most data management and statistical procedures dialog boxes. The variable name list can be kept in alphabetical order or in spreadsheet order. If you select spreadsheet order, you can reorder the variables using the Reorder Variables command on the Data menu (see Chapter 2).

Variable names are stored using both upper and lower case letters. Check the ***Report Variable Names in All Caps*** box if you'd prefer to have variable names displayed using all capital letters in reports.

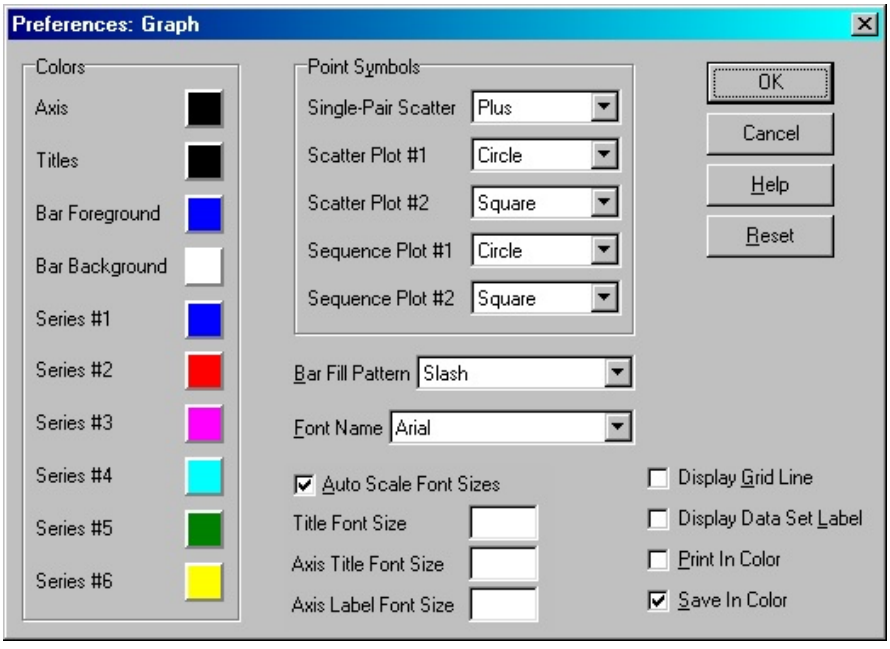
Select the date format you want to use for both data entry and reports—month/day/year, day/month/year, or year/month/day. You can also have dates displayed using either two or four digits for the year. Dates entered with two-digit years less than the ***Century Cutoff for 2-digit Years*** value will be entered as 21st century dates.

Check ***Create backup Statistix files*** if you'd like *Statistix* to create a backup file each time you save a *Statistix* file using the name of a file that already exists. Backup files have the extension .~sx.

You can select the font used to display reports from the ***Report Font Name*** drop-down list. You can also change the ***Report Font Size***.

Graph Preferences

There are many options regarding *Statistix* graphs. These options can be changed by selecting Graph from the Preferences menu. The dialog box is displayed below.



You can select the colors used to display the various components of a graph. Press a color button to change the color of the corresponding component. A color dialog box will appear from which you can select a basic color, or create a custom color.

You can use a number of symbols to mark points on X-Y plots (plus, start, circle, square, etc.). The ***Single-Pair Scatter*** is used to plot points on scatter plots when there is only one X-Y pair of variables. If there is more than one X-Y pair, *Statistix* uses different symbols to mark the different X-Y pairs. You can select the symbols you want for the first two X-Y pairs in multi-pair scatter plots. *Statistix* will select the remaining symbols for plots with more than two pairs.

Select the symbol you want *Statistix* to use for sequence plots (time series plots and control charts). You can plot up to seven series on a single plot using the **Time Series Plot** procedure. You can select the symbols to mark the points for the first two time series variables plotted. *Statistix* will select the symbols for the others when more than two variables are used.

Select your preference for the ***Bar Fill Pattern***. It will be used to shade the inside of the bars in histograms, error bar charts, and Pareto charts.

You can select the font name for text that appears on graphs. Check the ***Auto Scale Font Sizes*** box to have *Statistix* scale text to fit the graph. If you want smaller or larger fonts, uncheck the box and enter specific values for titles, axis titles, and axis labels.

There are also options to display grid lines, display the data set label at the top of graphs, print in color, and save graphs in color.

C H A P T E R

2

Data Menu

The **Data Menu** procedures are used to enter data and manipulate data already entered into *Statistix*. The **Data** menu is available whenever the spreadsheet window is active.

The **Insert** procedures are used to add variables and cases to the active spreadsheet. Use it to create new data files.

The **Delete** procedures are used to remove variables and cases from the active spreadsheet. Cases can be deleted by specifying a range of case numbers, or all omitted cases can be deleted.

The **Fill** procedure is used to fill consecutive cells with a single value.

The **Transformations** procedure is used to modify the values of existing variables and to create new variables using algebraic expressions and built-in transformation functions.

The **Recode** procedure is used to change a list of values of an existing variable to a single new value.

The **Indicator Variables** procedure is used to create variables that use the values 0 and 1 to indicate the absence and presence of a factor.

The **Stack** operation stacks several variables end-to-end to create a single long variable. The **Unstack** operation unstacks one variable into several shorter ones.

The **Transpose** operation transposes a table of data reversing the roles of variables and cases.

The **Omit/Select/Restore Cases** procedure lets you "hide" or "omit" specified cases from the program. These rows are ignored by *Statistix*, although they can be restored at will. If a case is not omitted, we say that the case is selected. The omit status of a case refers to whether it is omitted or selected.

The **Sort Cases** procedure is used to sort the cases of a data set based on the values of one or more variables.

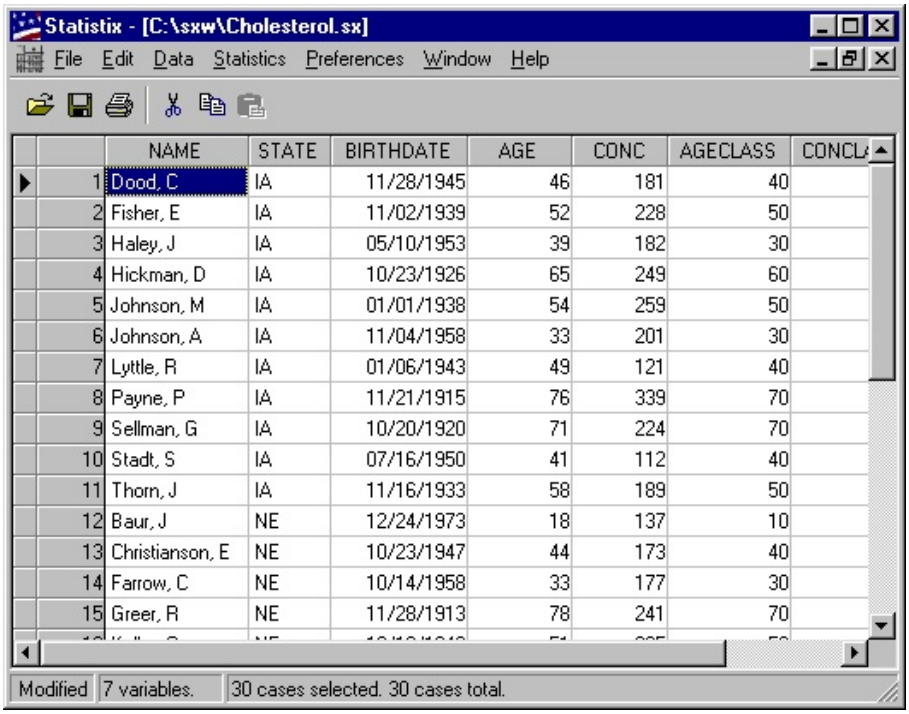
The **Reorder Variables** procedure is used to reorder the variables as they appear on the spreadsheet left to right.

The **Column Formats** procedure is used to control the column widths and numerical formats used to display each variable.

The **Labels** procedures are used to enter a data set label, variable labels, and value labels. These labels appear on reports.

Spreadsheet Window

The current *Statistix* data set is displayed in the spreadsheet window. The file name appears in the window's title bar. The word "Untitled" appears in the title bar for a new data set that hasn't been saved yet. The variable names appear along the top of the spreadsheet window, and the case numbers appear along the left margin. Both selected and omitted cases appear on the spreadsheet. The case numbers of omitted cases are dimmed to remind you that the cases are omitted. The values of omitted cases can be changed.



	NAME	STATE	BIRTHDATE	AGE	CONC	AGECLASS	CONCL
1	Dood, C	IA	11/28/1945	46	181	40	
2	Fisher, E	IA	11/02/1939	52	228	50	
3	Haley, J	IA	05/10/1953	39	182	30	
4	Hickman, D	IA	10/23/1926	65	249	60	
5	Johnson, M	IA	01/01/1938	54	259	50	
6	Johnson, A	IA	11/04/1958	33	201	30	
7	Lyttle, R	IA	01/06/1943	49	121	40	
8	Payne, P	IA	11/21/1915	76	339	70	
9	Sellman, G	IA	10/20/1920	71	224	70	
10	Stadt, S	IA	07/16/1950	41	112	40	
11	Thorn, J	IA	11/16/1933	58	189	50	
12	Baur, J	NE	12/24/1973	18	137	10	
13	Christianson, E	NE	10/23/1947	44	173	40	
14	Farrow, C	NE	10/14/1958	33	177	30	
15	Greer, R	NE	11/28/1913	78	241	70	

When the spreadsheet window is the active *Statistix* window, you can enter data in the cells, scroll through the data using the scroll bars, and manipulate the data using the **Data** procedures described in this chapter. You can make the spreadsheet window the active window by clicking anywhere on the window using your mouse if part of the window is visible, or by selecting it from the **Window** menu on the main menu bar.

The cell at the current position in the spreadsheet is highlighted. The value at the current position can be changed simply by entering a new value and pressing Enter. Enter the letter M for missing values for integer, real, and date variables. To partially change the value of a cell without retyping the

entire entry, first highlight the cell, then press F2 or click on the string in the highlighted cell. Press Esc to undo the changes to the current cell.

You can move the current position around the spreadsheet using the arrow keys, page up, and page down keys. You can move the current position to a different cell by clicking on that cell with your mouse.

The spreadsheet window can be manipulated in the same manner as other windows. It has the minimize, maximize/restore, and close buttons in the upper right corner of the window. You can resize the spreadsheet window by dragging a corner of the window with your mouse. You can move the spreadsheet window by dragging the title bar with your mouse.

Variable Data Types

Statistix can handle four types of data: real, integer, date, and string. A variable can only contain values of one data type.

The "real" data type is used to represent floating point numbers in *Statistix*. This format is the most flexible offered by *Statistix* and is used as the default data type when creating new variables.

Integer data in *Statistix* are whole numbers in the range -32767 to 32767. This data type uses only 25% as much space as the real data type. You use the integer data type instead of the real data type, when appropriate, to increase the data set capacity of *Statistix*. This also saves disk space by reducing the size of *Statistix* data files.

The "date" data type is used to represent dates. The "string" data type is used to enter alphanumeric data, such as a subject's name. String variables can be used as grouping variables for statistical procedures that compute results by group.

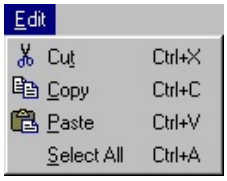
When typing numbers using *Statistix*, you can enter a number in either decimal format (e.g., 2.45) or exponential format (e.g., 1.23E+05). Enter the letter M to indicate a missing value for integer, real, and date variables, but not for string variables. A blank string is the missing value equivalent for string variables.

A variable's data type is established when you create it. Variables can be created using several of the **Data** procedures discussed in this chapter including **Insert Variables** and **Transformations**. Variables are also created using the **Import** procedure discussed in Chapter 3. You indicate

the data type of a new variable by typing the letter I, R, D, or S in parentheses after the variable name for integer, real, date, or string. String types require a number after the letter S to indicate the maximum length. For example, FirstName(s12) creates a string variable named FirstName with maximum length 12.

Edit Menu

In addition to the data management procedures listed on the Data menu, you can manipulate the spreadsheet data using the Cut, Copy, and Paste commands on the **Edit** menu. The Cut and Copy commands copy selected spreadsheet cells to the *Windows* clipboard, and the Paste command retrieves information from the clipboard. You can use these commands to move a block of cells from one place on the spreadsheet to another, or to export the block to another *Windows* application. You can also import data from other programs via the clipboard using the Paste command.



Before you can use the Cut and Copy commands, you have to select part of the spreadsheet to cut or copy. Use your mouse to select cases, variables, or a rectangular block of cells. To select a single case, click on the narrow *row-bar* located just left of the case numbers. The whole case is highlighted to show that it's selected. You can select a range of cases by clicking on the row-bar for the first case in the range, and dragging the mouse to the last case in the range. You can select noncontiguous cases, hold the Ctrl key down and click on the row bar for the cases you want to select.

To select a variable, click on the variable name at the top of the spreadsheet. To select a range of variables, click on the first variable name, then drag the mouse to the last variable name in the range. To can select noncontiguous variables, hold the Ctrl key down and click on the variable names you want to select.

To select a single cell, click on that cell. To select a range of cells, click on one corner of the rectangle, then drag the mouse to the opposite corner of the rectangle. NOTE: To begin selecting a range of cells, you must point your mouse to the extreme left side of the starting cell. The mouse cursor

will change direction, then press and hold the mouse button and begin dragging. The cells are highlighted as you drag the mouse so it's easy to see what you've selected.

Once you've selected cases, variables, or cells, use the Cut or Copy command to copy the selected data to the clipboard. The ***Cut*** command copies the selected cells to the clipboard, and then deletes them from the spreadsheet. The ***Copy*** command copies the selected cells to the clipboard without deleting them. The information remains on the clipboard until you next use the Cut or Copy command in *Statistix* or another *Windows* program. Press Del deletes the selected cases, variable, or cells without first copying the information to the clipboard.

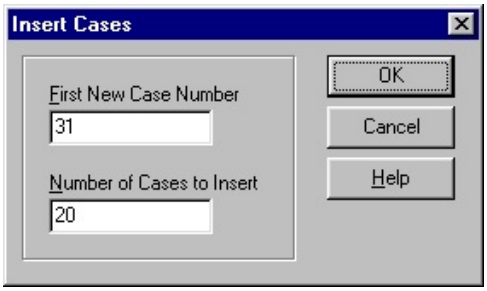
Information on the clipboard, whether it was copied there from *Statistix* or a different program, can be pasted anywhere in the *Statistix* spreadsheet. To paste information, first click on a cell to select the insertion point. Then, select the ***Paste*** command from the Edit menu. The pasted data overwrites the data of existing variables starting at the current position. When pasting data on the spreadsheet, keep in mind that the data must be consistent with the data types of the target variables.

When you're pasting data into an empty Statistix spreadsheet, Statistix will create new variables to hold the data. The first row of the clipboard data can be used for variable names, but the names must conform to the rules of *Statistix* variables names. Names must start with a letter, consist of only letters, digits, and the underscore character. Variable names can include data type information (e.g., lastname(s15)). If the first row doesn't have variable names, Statistix will generate names starting with V001.

Insert

When you select **Insert** from the **Data** menu, a pop-up menu with two alternatives appears: Cases and Variables. Select **Variables** to add variables to a new or existing data set. Select **Cases** to insert cases into an existing data set.

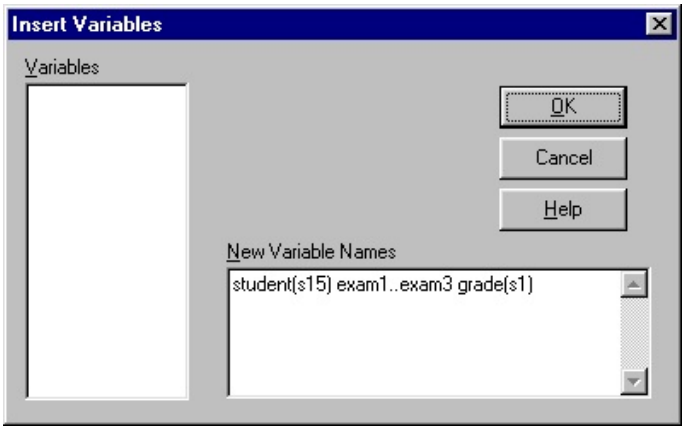
Insert Cases



First enter the case number where you want to insert the new cases. Then enter the number of new cases to insert. Press ***OK***.

New cases appear in the spreadsheet window as rows of M's representing missing values. Enter your data by typing in the actual values over the M's.

Insert Variables



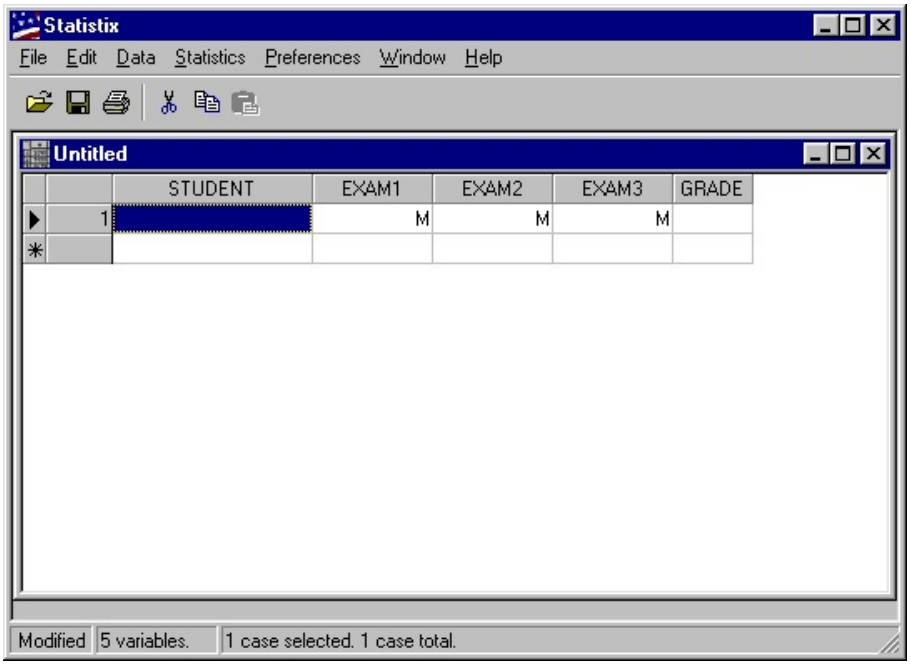
The existing variables, if any, are listed in the ***Variables*** list box for your reference. You list a variable name for each column of data you intend to enter in the ***New Variable Names*** edit control. Variable names must start with a letter and can contain letters, digits, and the underscore character. Variable names can be up to nine characters long.

In *Statistix*, data can be integer, real, date, or string. However, a particular column can only be used to store one type of data. You can specify the data type of variables when you list the variable names. Use the letters I, R, D, and S in parentheses after the variable names to identify integer, real, date, and string types. String types require a number after the letter S to indicate the maximum length. If you don't specify a data type, real is assumed.

In the dialog box on the preceding page, the variable STUDENT is a string variable with a maximum length of 15 characters and the variable GRADE is a string variable with length 1. No data type is specified for the variables EXAM1, EXAM2, and EXAM3, so they are assigned the real data type.

The variable list EXAM1, EXAM2, and EXAM3 was abbreviated using what we call VAR01 .. VAR99 syntax. A data type entered at the end of the list is applied to all of the variables in the list (e.g., Q1 .. Q15(I)).

After listing the variable names, press ***OK*** to begin data entry. For a new data set, an empty spreadsheet will appear, as in the following example.



To enter data, simply enter values for each cell and press Enter. Pressing Enter or Tab advances the cursor toward the right. Press Shift-Tab to move to the left. Press the up arrow to move up and the down arrow to move down. Enter the letter M for missing values in integer, real, and date variables, but not for string variables.

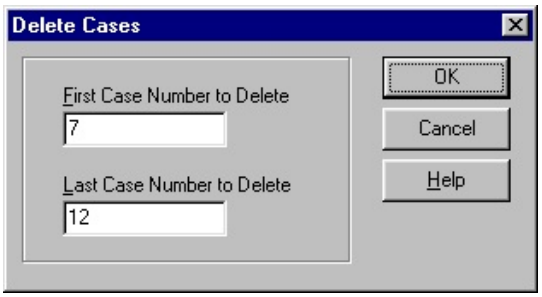
You can use the arrow keys to go back and correct errors anytime. Simply type in a new value over the existing value in a cell. Press F2 to edit the string at the current cell. You can also move to a different cell by clicking the mouse on that cell.

Delete

Select **Delete** from the **Data** menu to delete cases or variables. You will be presented with a pop-up menu giving you four choices: Cases, Omitted Cases, Selected Cell, and Variables.

Delete Cases

Use the delete cases command to delete one case or to delete a contiguous block of cases.



Enter the first and last case numbers of the range of cases you want to delete, then press the ***OK*** button.

An easier way to delete cases is to highlight one or more cases using your mouse, and then pressing the Delete key. To select a single case, click on the narrow *row-bar* located just left of the case numbers. The whole case is highlighted to show that it's selected. You can select a range of cases by clicking on the row-bar for the first case in the range, and dragging the mouse to the last case in the range. You can select noncontiguous cases, hold the Ctrl key down and click on the row-bar for the cases you want to select.

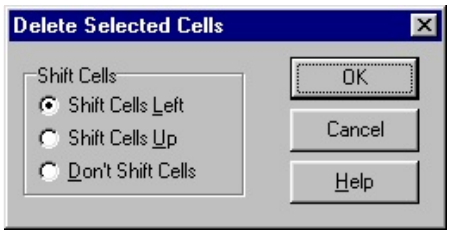
Delete Omitted Cases

You can also delete all omitted cases. This method of deleting cases gives you greater flexibility in selecting which cases to delete. The **Omit/Select/Restore Cases** procedure discussed later in this chapter can be used to omit cases based on data values rather than case numbers.

Delete Selected Cells

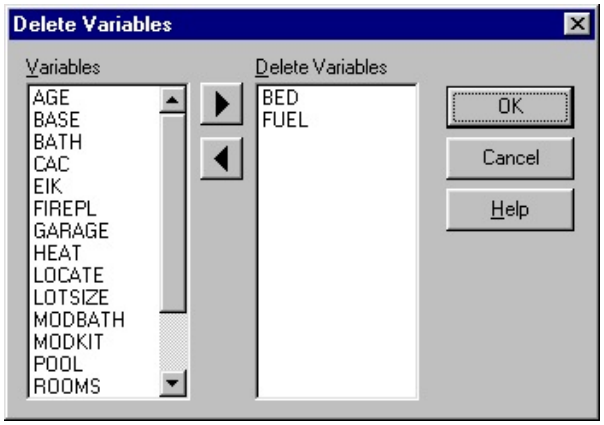
This procedure lets to delete an arbitrary rectangle of cells, and optionally shift the remaining cells up or to the right to fill the void. First highlight the cells you want to delete. To select a single cell, click on that cell. To select a range of cells, click on one corner of the rectangle, then drag the mouse to the opposite corner of the rectangle. NOTE: To begin selecting a range of cells, you must point your mouse to the extreme left side of the starting cell. The mouse cursor will change direction, then press and hold the mouse button and begin dragging. The cells are highlighted as you drag the mouse so it's easy to see what you've selected.

Once you've highlighted a range or cells, select Delete Selected Cells from the Data menu (or press the Delete key), and the dialog box shown below appears.



Select one of the Shift Cells radio buttons. Select ***Shift Cells Left*** to have data to the right of the deleted rectangle shifted left. Select ***Shift Cell Up*** to have data below the deleted rectangle shifted up. Select ***Don't Shift Cells*** to fill the deleted cells with missing values.

Delete Variables



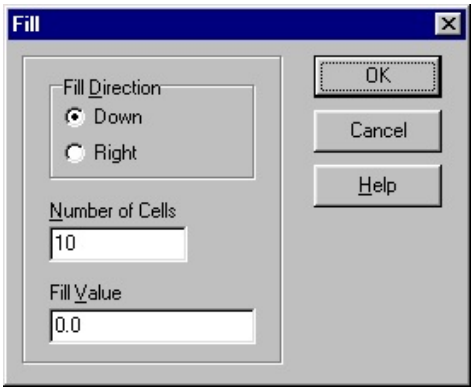
Move the variables you want to delete from the ***Variables*** list box on the left to the ***Delete Variables*** list box on the right. First highlight one or more

variables in the Variables box, then press the right-arrow button to move them. You can move a single variable by double-clicking on the name with your mouse. Press the ***OK*** button to delete the variables in the Delete Variables box.

An alternative method of deleting variables is to highlight one or more variables using your mouse, and then pressing the Delete key. To select a variable, click on the variable name at the top of the spreadsheet. To select a range of variables, click on the first variable name, then drag the mouse to the last variable name in the range. To can select noncontiguous variables, hold the Ctrl key down and click on the variable names you want to select.

Fill

The **Fill** command is used to fill a number of contiguous cells with a particular value.



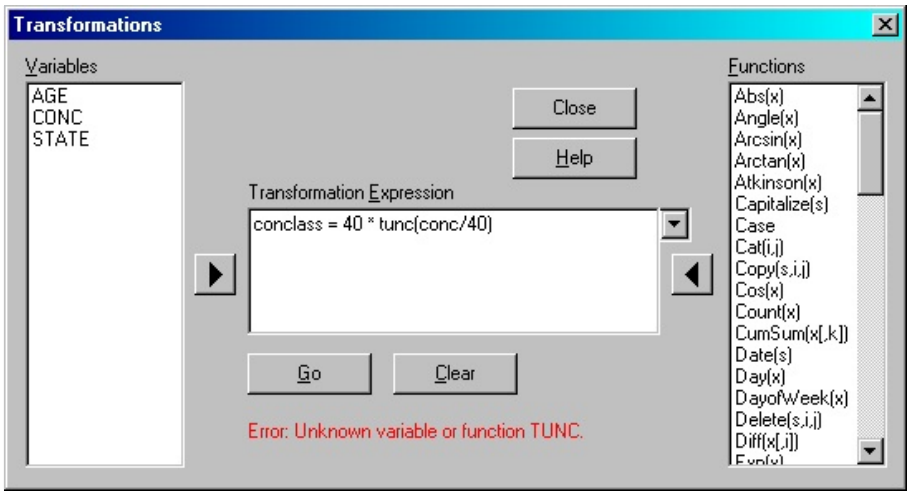
Cells will be filled starting at the current position. First select the direction in which you want the filling to proceed from the ***Fill Direction*** radio buttons, either down or right. Enter the number of cells you want to fill in the ***Number of Cells*** edit control. Enter the value you want to us to fill the cells in the Fill Value edit control. If you select ***Down*** for the fill direction, new cases will be added to the bottom of the spreadsheet if necessary. If you select ***Right***, then filling will automatically wrap around when the end of a case is reached and will continue on the next row.

Transformations

This powerful procedure is used to create new variables or to alter the values of the existing ones. It has two general forms: simple assignment and conditional assignment using the If-Then-Else construct.

If you're not familiar with arithmetic and logical expressions, you should read the **Arithmetic and Logical Expressions** section at the end of this chapter.

Specification



The ***Transformations Expression*** edit control in the center of the dialog box is where you build your transformation. The data set variables are listed in the ***Variables*** list box on the left. The *Statistix* built-in functions are listed in the ***Function*** list box on the right. Below the Transformation Expression box is an area where error messages are displayed. The error message "Unknown variable or function TUNC" in the example is reporting that the built-in function Trunc has been misspelled.

The transformation expression must be typed in manually. You can, however, select variables from the Variables box and built-in functions from the Functions box and insert them into the Transformation Expression edit control at the cursor's current position. You can move the cursor in the edit control using the arrow keys or the mouse. To copy a variable name or function, highlight the variable or function name with your mouse, then press the corresponding arrow button.

You're only allowed to enter a single transformation in the Transformation Expression edit control. The control has several lines so that you can enter a long expression.

Once you've entered an expression, press the ***Go*** button. If *Statistix* finds an error in your expression, an error message is displayed below the Transformation Expression box. Edit your expression to correct the error and press Go again. Press the ***Clear*** button to erase the contents of the Transformation edit control.

Clicking on the down-arrow to the right of the Transformations Expression edit control displays a drop-down list of previously entered expressions. Selecting an expression from the list copies it to the main expression box.

Simple Assignment

Suppose the variable NEWVAR is a new one you want to create or an existing one that you want to alter. Such a variable is called a **target variable**. In a simple assignment, a target variable is simply equated with an arithmetic expression:

```
{target variable} = {arithmetical expression}
```

To give a specific example, suppose you want NEWVAR to be the sum of the variables A, B, and C.

```
NEWVAR = A + B + C
```

A new variable called NEWVAR has now been created with the sum of the variables A, B, and C. If a variable called NEWVAR already exists, the variable's values will be replaced by the sum A + B + C.

A target variable's name can appear on both sides of an assignment statement provided the variable already exists.

```
TARGET = 2.75 + SQRT (TARGET)
```

Sqrt is an example of a built-in function that computes the square root of the argument. The built-in functions are described on page 54.

Statistix can handle four types of data: real, integer, date, and string. Use the letters I, R, D, and S in parentheses after the name of the new variable to identify integer, real, date, and string types. String types require a number after the letter S to indicate the maximum length. If you don't specify a data type, real is assumed. The examples on the next page illustrate how integer, date, and string variables can be created.

```
COUNT (I) = A + B + C
DUEDATE (D) = SALEDATE + 30
FULLNAME (S30) = FIRSTNAME + " " + LASTNAME
```

Date and string constants must be enclosed in quotation marks. For example:

```
SALUTE (S20) = "Ms. " + LASTNAME
LENGTH = LASTDAY - "1/1/95"
```

Conditional Assignment

On occasion you may want to apply one assignment to the target variable when some condition is met and another assignment when the condition is not met. Conditional assignments employing the If-Then-Else construct are designed for this purpose. Its general form is:

```
IF {logical expression}
THEN {target variable} = {some arithmetic expression}
ELSE {target variable} = {some other arithmetic expression}
```

If a specified logical expression for a particular case is true, the THEN clause is performed; otherwise, the ELSE clause is performed. This construct is quite flexible, as the following examples illustrate.

Suppose you want to create a new variable—AGEGROUP—based on the values of a variable AGE. You want AGEGROUP to be the tens digit of age, but you want to group all ages of 60 or greater into one group.

```
IF AGE < 60
THEN AGEGROUP = TRUNC (AGE/10)
ELSE AGEGROUP = 6
```

The three key words—IF, THEN, and ELSE—need not be on separate lines. A short statement can be written on one line. The ELSE expression can be omitted from a conditional transformation, in which case the target variable is left unchanged when the logical expression is not true.

```
IF AGEGROUP > 6 THEN AGEGROUP = 6
```

The logical expression can include any valid arithmetic expressions, and the arithmetic expressions can include the target variable if it already exists.

```
IF (A + B) <= (1.25 * SIN (C))
THEN A = 0.0
ELSE A = A + D + E
```

Converting Variable Types

All of the values of a particular variable must be of the same data type, either integer, real, date, or string. The data type of an existing variable can

be changed using a transformation. For example, suppose that the variable CONC was originally created as a real variable and contains values for cholesterol concentration. You can convert the variable CONC to an integer variable using the transformation:

```
CONC (I) = CONC
```

If a value for CONC exceeded the limits for an integer variable (-32767 to 32767), the result would be missing. An integer variable can just as easily be converted to a real variable:

```
HEIGHT (R) = HEIGHT
```

Sometimes you end up with string variables that contain numbers or dates. Since *Statistix* can't do arithmetic using string variables, it's best to convert string variables of these types using the special string conversion functions Date and Number.

```
BIRTHDATE (D) = DATE (BIRTHDATE)
LEVEL (R) = NUMBER (LEVEL)
```

The maximum length of a string variable can be changed using a transformation. Suppose you had created a variable NAME (S15) but later decided that 15 characters were insufficient. You could increase the maximum length using the transformation:

```
NAME (S20) = NAME
```

Missing Values

If a number used in an arithmetic expression is missing, the result of the expression is also missing. After all, you can't perform arithmetic on numbers that don't exist. Consider the transformation:

```
A = B + C
```

If the value for the variable C is missing for a particular case, the expression B + C is missing and the target variable A is assigned the missing value.

In the logical expression of a conditional transformation, it makes sense to make tests of equality using missing values. The expressions IF X = Y and X <> Y are evaluated normally if either X or Y is missing. However, a number can't be less than or greater than a missing value. So when X or Y is missing in an expression like IF X < Y, the logical expression can't be evaluated and the target variable is assigned a missing value.

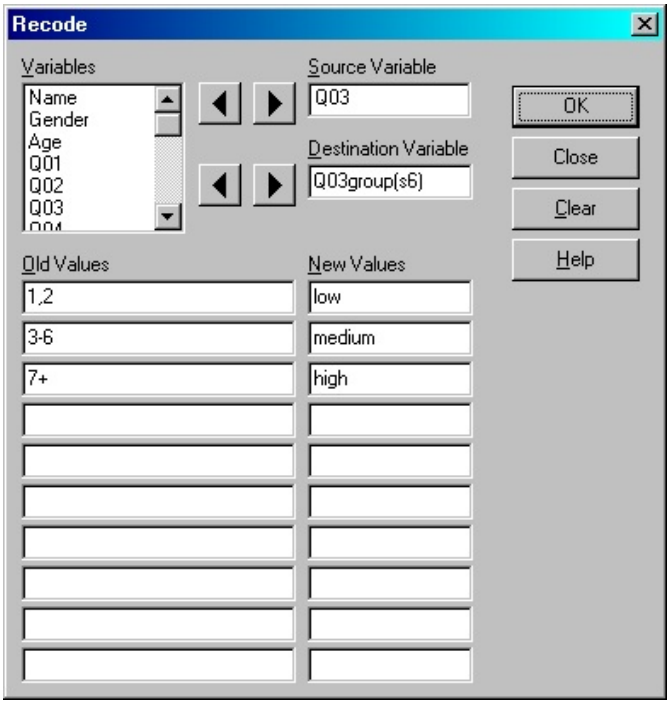
The treatment of missing values is discussed at greater length in the **Arithmetic and Logical Expressions** section on page 49.

Omitted Cases

Cases omitted using **Omit/Select/Restore Cases** are ignored when a transformation is performed. If the target variable is a variable that already exists, omitted cases retain their old values, i.e., they are not transformed. If the target variable doesn't already exist, omitted cases are assigned the missing value when the transformation is performed.

Recode

The **Recode** procedure is used to replace (or "recode") a list of values of a variable with a new value. It's most useful when you want reduce the number of values a variable has by replacing them with new values that identify groups of the original values.



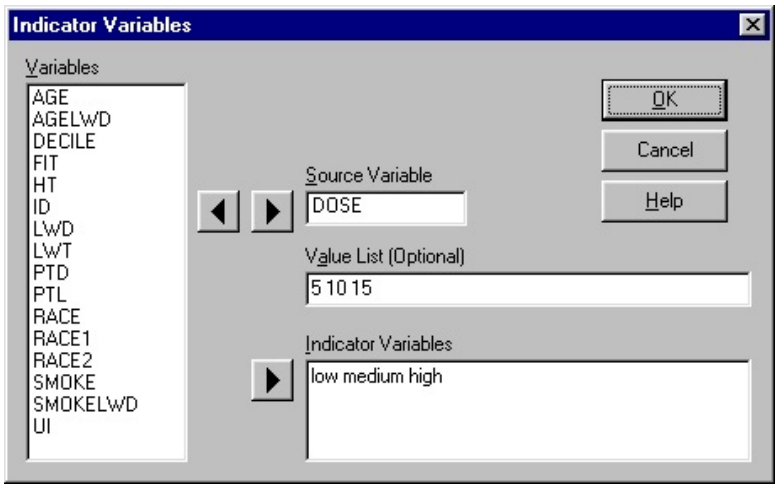
First select the variable you want to recode from the ***Variables*** list and copy it to the ***Source Variable*** box. Next, identify the ***Destination Variable***. You can select an existing variable, or enter the name of a new variable. The

source and destination variables can have any data type (integer, real, date or string) and they needn't have the same type.

Finally, enter lists of values of the source variable in the column of boxes titled ***Old Values***, and a corresponding new value for the destination variable in the column of boxes titled ***New Values***. When entering a list into an Old Values box, separate values with commas or spaces. You can specify a range of values using a dash (-) or colon (:). Use the plus sign to indicate an open ended range of values. When listing the values of a string variable, long strings, or strings that contain spaces or commas must be enclosed in quotation marks (e.g., "George Washington", "Abe Lincoln").

Indicator Variables

An indicator variable (also called a dummy variable) uses the values 0 and 1 to indicate the absence or presence of a factor or condition. This procedure creates indicator variables based on the values of an existing variable.



For example, suppose you have a categorical variable named DOSE that contains the values 5, 10, and 15, and you want to create three indicator variables named LOW, MED, and HIGH. In the dialog box above, the variable DOSE was moved from the ***Variables*** list box to the ***Source***

Variable box. The three values 5, 10, and 15 were then entered into the ***Value List*** edit control. The names of the three new variables LOW, MED, and HIGH were then entered into the ***Indicator Variables*** edit control.

For cases where DOSE has the value 5, the new variable LOW will be given the value 1; otherwise, the value would be 0. For cases where DOSE has the value 10, the new variable MED will be given the value 1; otherwise, the value would be 0. For cases where DOSE has the value 15, the new variable HIGH will be given the value 1; otherwise, the value would be 0.

The source variable containing the existing levels can be an integer, real, date, or string variable. The levels listed for an integer or a real variable must be nonnegative whole numbers. String values in the list of levels must be enclosed in quotation marks.

The list of levels can be omitted if the levels are consecutive numbers starting with 1. For example, if DOSE had contained the levels 1, 2, and 3, the Value List could have been left empty.

Stack

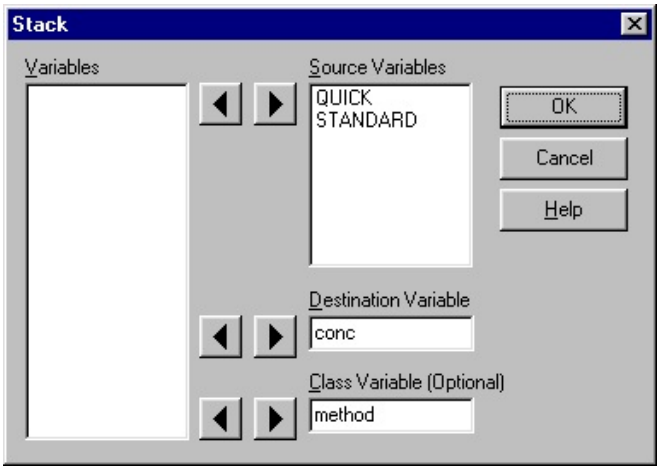
The **Stack** procedure is used to stack several variables end-to-end to create a single long variable.

Data to be analyzed can often be broken down into groups of interest. For example, you may be interested in cholesterol concentration by age group, or fat absorption of doughnuts by the type of fat used in cooking. There are two ways data of this type can be presented. One method is to store the data for each group in its own variable. A second method is to put all the data into one variable, and use a second categorical variable to identify the groups. You can use the Stack procedure to transform data stored in the first format to data stored in the second format.

An example of the Stack dialog box appears on the next page. First, move the names of the variables you want to stack together from the Variables list box to the ***Source Variables*** list box. Then, enter the name of a new or

existing variable to capture the stacked data in the ***Destination Variable*** box.

You can also specify a ***Class Variable***. The class variable is assigned a number for each source variable, starting with number 1 for the first source variable. You can use it to tell from which source variable each item of data in the destination variable originated.



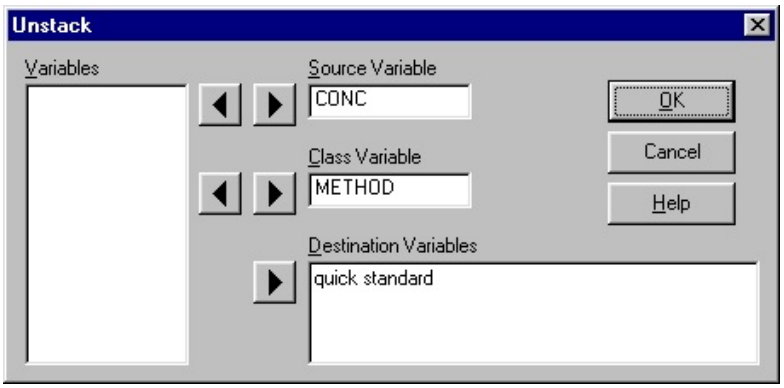
In the example dialog box above, the variables QUICK and STANDARD are selected as the source variables. A new variable CONC has been entered as the destination variable. The new variable METHOD has been entered as the class variable and will capture the group numbers 1 and 2 for the source variables QUICK and STANDARD, respectively. The resulting data set, which contains both the original variables and the new variables, is presented below.

QUICK	STANDARD	CONC	METHOD
23	25	23	1
18	24	18	1
22	25	22	1
28	26	28	1
17	M	17	1
25	M	25	1
19	M	19	1
16	M	16	1
M	M	25	2
M	M	24	2
M	M	25	2
M	M	26	2

Unstack

The **Unstack** procedure is used to break one long variable into two or more shorter variables. The data in the original variable is divided up between the new variables based on the values of a class variable.

The Unstack procedure is the reverse of the Stack procedure discussed on the preceding page. We'll illustrate the Unstack procedure using the same data presented above.



Suppose you start with the variables CONC and METHOD, and you want to make the two variables QUICK and STANDARD. You'd use the Unstack procedure to do this, as illustrated in the dialog box above. Both the ***Source Variable*** and the ***Class Variable*** are required. The class variable can be any data type (integer, real, date, or string). There can be no more than 500 groups. List the ***Destination Variables***. You can use VAR1 .. VAR99 syntax to abbreviate the list. If there are n values for the class variable, you must enter exactly n variables in the list.

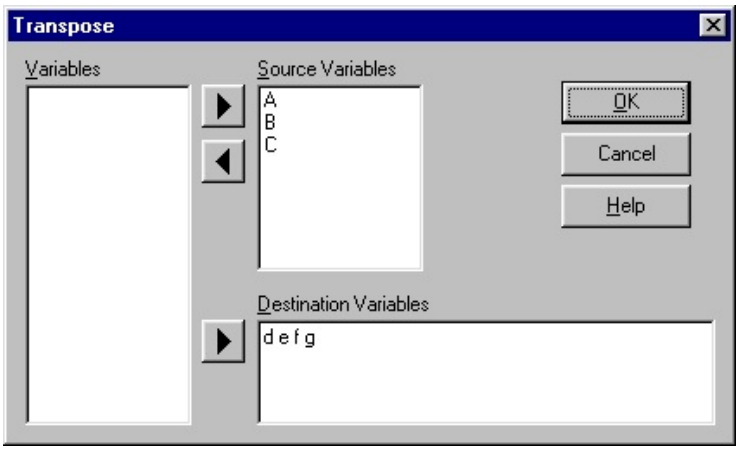
Transpose

Most *Statistix* procedures operate on columns of data. You may on occasion want to obtain statistics for rows of data. The transpose operation is used to copy data from rows of selected variables to a new set of variables, thus reversing the role of cases and variables.

This is best explained using an example. Suppose we have three variables A, B, and C with four cases of data:

CASE	A	B	C
1	1	2	3
2	4	5	6
3	7	8	9
4	10	11	12

To transpose this table of data, we specify the variables A, B, and C as the ***Source Variables***. Since there are four cases, we need four ***Destination Variables***. We list the new variables D, E, F, and G, as illustrated in the dialog box below.



The results are shown below.

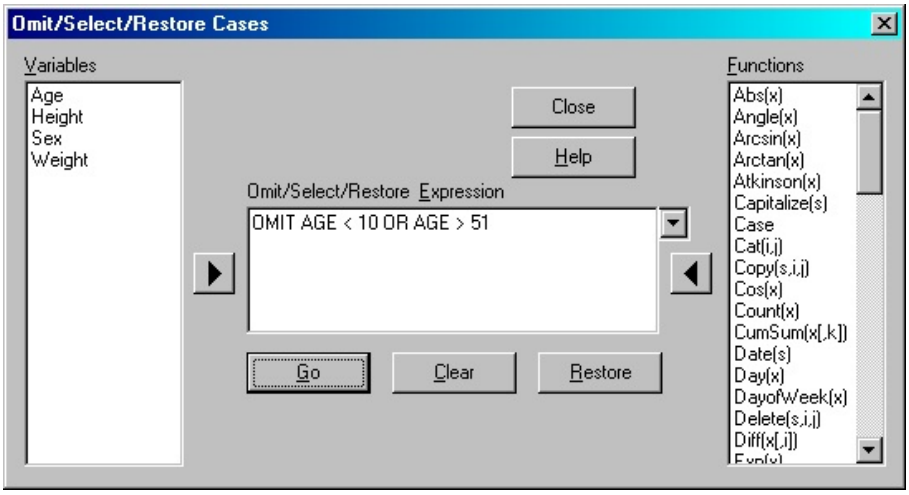
CASE	A	B	C	D	E	F	G
1	1	2	3	1	4	7	10
2	4	5	6	2	5	8	11
3	7	8	9	3	6	9	12
4	10	11	12	M	M	M	M

Note how the values 1, 2, and 3, reading left to right for variables A, B, and C now appear reading from top to bottom for variable D.

Omit/Select/Restore Cases

If you want to analyze a select subset of the cases in *Statistix*, the **Omit/Select/Restore Cases** procedure lets you temporarily "hide" some of your data from the program. Once cases are omitted, they are ignored by the statistical procedures until they are "restored". Omitted cases can also be selectively restored using a "select" statement. The omit status of cases are saved when you use the **Save** procedure.

You specify the cases you want to omit using logical expressions. If you're not familiar with logical expressions, please refer to the **Arithmetic and Logical Expressions** section on page 49.



The ***Omit/Select/Restore Expression*** edit control in the center of the dialog box is where you build your omit statement. The data set variables are listed in the ***Variables*** list box on the left. The *Statistix* built-in functions are listed in the ***Function*** list box on the right.

The omit expression must be typed in manually. You can, however, select variables from the Variables box and built-in functions from the Functions box and insert them into the Omit/Select/Restore Expression edit control at the cursor's current position. You can move the cursor in the edit control using the arrow keys or the mouse. To copy a variable name or function, highlight the variable or function name with your mouse, then press the corresponding arrow button.

You're only allowed to enter a single omit statement in the Omit Expression edit control. The control has several lines so you can enter a long expression.

Once you've entered an expression, press the ***Go*** button. If *Statistix* finds an error in your expression, an error message is displayed below the Omit Expression box. Edit your expression to correct the error and press ***Go*** again. Press the ***Clear*** button to erase the contents of the Omit Expression edit control.

Clicking on the down-arrow to the right of the Omit Expression edit control displays a drop-down list of previously entered expressions. Selecting an expression from the list copies it to the main expression box.

You can enter three types of statements: "omit", "select", and "restore". The omit statement has the general form OMIT {logical expression}. Likewise, the select statement has the form SELECT {logical expression}. The restore statement is simply the word RESTORE, which instructs *Statistix* to restore the omit status of all cases to be selected.

The omit status of a case can be either selected or omitted. When a data set is created, all cases are selected. You can use the omit statement to selectively omit cases that currently have the status selected; the omit statement doesn't change the status of cases already omitted. Use the select statement to change the status of omitted cases back to selected; the select statement doesn't change the status of already selected cases.

Once you've entered and successfully executed an omit expression, a message box reports the number of cases omitted or selected. The number of cases currently selected is displayed on the status bar that appears at the bottom of the *Statistix* window.

The successful omit or select expression remains in the Omit Expression edit control and can be edited to create a new omit expression. To delete a previous expression from the edit control, press the ***Clear*** button.

The effects of sequential omit expressions are cumulative; a second omit expression will only act upon the cases not omitted by the first expression. Thus, the two expressions "OMIT AGE < 10" and "SEX <> 'F'" entered one after the other have the same effect as the single expression "OMIT AGE < 10 OR SEX <> 'F'".

The omit expression can be as complex as you like and can span all of the lines in the edit control. You can press Enter while typing an omit expression to advance to the next line.

A useful tip for using the omit command takes advantage of the fact that the arithmetic involving missing values always results in a missing value. Thus, the omit expression

```
OMIT AGE + HEIGHT + SEX + WEIGHT = M
```

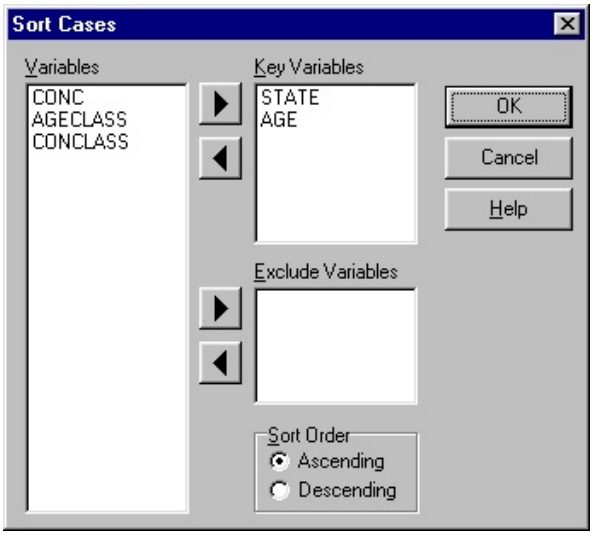
will omit a case if any of the variables have a missing value for the case and is easier to type than the alternative:

```
OMIT AGE = M OR HEIGHT = M OR SEX = M OR WEIGHT = M
```

Please see the **Arithmetic and Logical Expressions** section on page 49 for more information about logical expressions. The built-in functions are defined on page 54.

Sort Cases

Statistix offers a **Sort Cases** procedure to sort the cases of a data set into ascending or descending order based on the values of selected key variables.



Highlight the variables you want to use as the key variables in the ***Variables*** list box and press the right-arrow button to move them to the ***Key Variables*** box. The order in which the variables appear in the Key Variables list box is important. The cases will be sorted by the first variable first, then by the second variable within the first, and so on.

Any type of variable (integer, real, date, or string) can be used as a key variable. Sorting using string keys is not sensitive to upper and lowercase letters (e.g., the keys "Iowa" and "IOWA" will be treated the same).

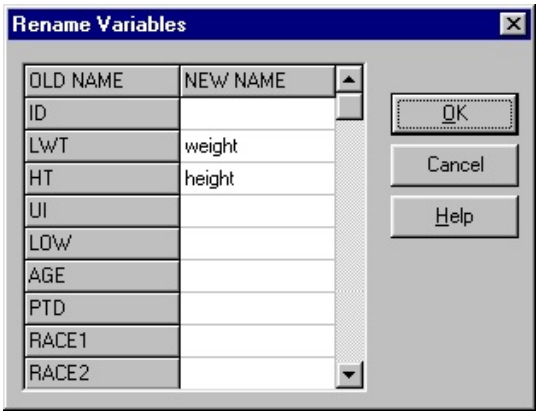
In most applications, you'll want the cases kept intact so that the values for the non-key variables are moved to their new positions along with the key variable values. Occasionally, you'll want to sort the values of a variable without disturbing the order of some or all of the remaining variables. In such cases, move the variables you want excluded to the ***Exclude Variables*** list box.

You can have your data sorted into ascending or descending order. Indicate your preference by selecting one of the ***Sort Order*** radio buttons.

Omitted cases are sorted along with selected cases. The omit status of a case moves with the key variables.

Rename Variables

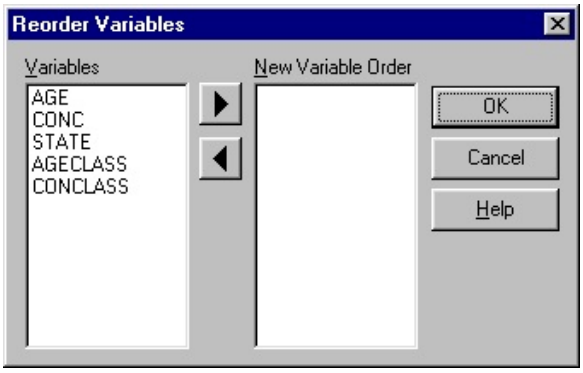
Use this procedure to rename variables in your open data file.



Simply type in new variable names in the ***New Names*** box next to the old names of the variables you want to rename. Leave the space blank for variables you don't want to rename.

Reorder Variables

The **Reorder Variables** procedure is used to change the order in which the variables are displayed in the spreadsheet window. If you've selected

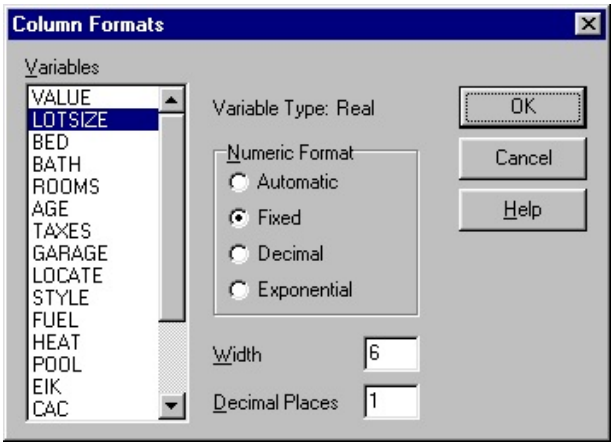


spreadsheet order rather than alphabetical order using the **Preferences** procedure (see Chapter 1), it will also change the order that variables are listed in the Variables list boxes that appear in dialog boxes.

The variables in the ***Variables*** list box are listed in the order they currently appear in the spreadsheet. You move the variables to the ***New Variable Order*** list box in the order you want them to appear in the spreadsheet. First highlight one or more variables in the Variables box, then press the right-arrow button to move them. You can move a single variable by double-clicking on the name with your mouse. Press the ***OK*** button to reorder the variables.

Column Formats

This procedure allows you to control the column width of data stored in variables, and the numerical format of real variables. The widths and formats you choose affect both the appearance of the columns in the spreadsheet window and columns displayed in the **Print** report (see Print in Chapter 3).



First, highlight one or more variable names in the ***Variables*** list box. The data type of the first variable you highlight appears on the dialog box. Only variables with the selected data type will be changed. In the example, the highlighted variable LOTSIZE is a real variable.

For string, date, and integer variables, you can only change the column width (number of characters). For real variables, choose a format by selecting one of the four ***Numeric Format*** radio buttons, then enter values for ***Width*** and ***Decimal Places*** (***Significant Digits*** for decimal and exponential formats). Be sure to enter a width large enough to account for the decimal point and the minus sign for negative numbers. Press the ***OK*** button to save your changes.

You can apply a single column format to more than one variable at a time. To do so, highlight one of the variables you want to format; the current column format values for that variable appear in the dialog. Next, highlight additional variables you want to format. You can do this by clicking on the variable names while holding down the control key. You can also select a range of contiguous variables by clicking on the first variable in the range and dragging the pointer to the last variable in the range while holding the mouse button down.

The ***Automatic*** format displays numbers using an integer format for whole numbers, or a decimal format when the number contains a fraction. For numbers with a fraction, as many digits as possible will be displayed, but trailing zeros are trimmed. The automatic format works well for data you've entered manually because the numbers generally are displayed just as you've entered them. For computed variables, such as variables created using the Transformations procedure, the automatic format often displays nonsignificant digits.

The ***Fixed*** format displays numbers in a decimal format with a fixed number of decimal places.

The ***Decimal*** format displays numbers using a decimal format where the decimal pont is free to move to maximize the number of digits displayed. You specify the number of ***Significant Digits***.

The ***Exponential*** format displays numbers using scientific notation. A number displayed in exponential format has two parts, the mantissa and the exponent. The mantissa is a number displayed as a decimal number that's always greater than or equal to 1.0 and less than 10.0. The exponent is displayed using the letter E followed by a signed integer. The number represented in this fashion is the mantissa multiplied by 10 raised to the exponent. For example, the number 4.23E-02 is equal to 4.23×10^{-2}, or 0.0423. You specify the number of ***Significant Digits***, which is the number of digits in the mantissa.

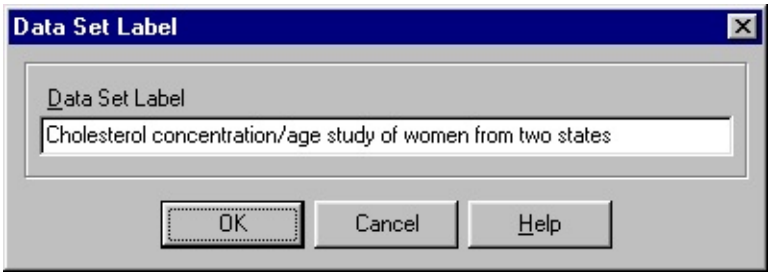
Labels

The **Labels** procedure is used to enter or change the current values of three kinds of labels: the data set label, variable labels, and value labels. These labels are used to annotate *Statistix* reports and graphs. After selecting Labels from the Data menu, you're presented with a pop-up menu with the three selections: Data Set, Variable, and Value.

The **File Info** report on the **File** menu (see Chapter 3) displays all of the labels for the open file.

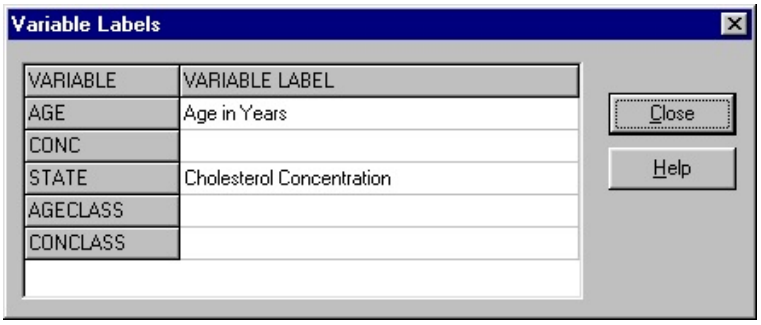
Data Set Label

The **Data Set Label** is a one line heading used to describe the data set. It's printed at the beginning of each *Statistix* report.



Variable Labels

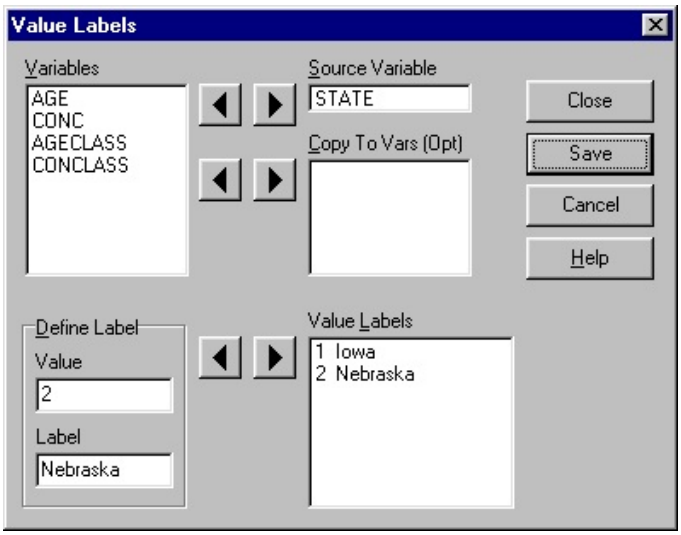
Variable labels are descriptive comments for variables. Variable labels will remind you what data the variables contain and how they were created. Variable labels are incorporated into the heading of some reports, such as the stem and leaf plot. They are also used for axis labels for graphs (e.g., scatter plots and histograms).



The variables and variable labels are presented in a table as shown in the example on the preceding page. Type in your variable labels beside the corresponding variable names. Variable labels can be up to 40 characters long.

Value Labels

Value labels are descriptive strings attached to individual values for a variable. For example, coding states using numbers may be convenient for data entry purposes, (1 for Iowa and 2 for Nebraska in the example dialog below.) Value labels serve as comments to remind you what the codes represent. These labels also appear in *Statistix* reports (e.g., cross tabulations) to improve readability.



First, highlight the variable of interest and move it to the ***Source Variable*** box. The current value labels of the selected variable, if any, are displayed in the ***Value Labels*** list box. If you want to use the same labels for other variables, move the variables to the ***Copy To Vars*** list box.

To define a label, enter a number in the ***Value*** edit control, and a string in the ***Label*** edit control, then press the right-arrow button to add the value-label pair to the Value Labels list. Value labels can be up to ten characters long.

To delete a defined label, highlight the value-label pair in the Value Labels box, then press the left-arrow button. The value and label appear in the Define Label edit controls so you have the opportunity to edit the label and

add it back to the list.

When you've finished defining value labels for the source variable, press the ***Save*** button to save the labels. Then select another source variable, or press the ***Close*** button to exit.

You can copy the labels already defined for one variable to one or more other variables. First, move the variable that has labels defined to the Source Variable box. The variable's labels are displayed in the Value Labels box. Then, move the unlabeled variables to the Copy To Vars list box. Then press the Save button.

Arithmetic and Logical Expressions

The data management procedures **Transformations** and **Omit/Select/ Restore** are powerful data manipulation tools. To appreciate the full potential of *Statistix*, you must understand the principles of arithmetic and logical expressions discussed in this section. This material will be familiar to people who are experienced in either database management software or programming languages.

Arithmetic Expressions

The following arithmetic operators are available:

^ Exponentiation. For example, A ^ B is A raised to the B-th power.

* Multiplication. A * B is the product of A and B.

/ Division. A / B is A divided by B.

+ Addition. A + B is the sum of A and B.

- Subtraction or reversal of sign. A - B is B subtracted from A. The expression -A, unary negation, reverses the sign of A.

These operators are used to form arithmetic expressions with constants, variable names, and built-in functions. A constant is simply a number, such

as 1.96 or 3.1416. String and date constants must be enclosed in quotation marks (e.g., “Great Scott”, “10/12/55”). Built-in functions are described in more detail on page 54.

An example of an arithmetic expression is A + 2.75 ∗ B, where A and B are variable names and 2.75 is a constant. *Statistix* evaluates this expression for each case by first taking the product of the constant 2.75 and the variable B. The value of variable A is then added to get the final result.

The order in which operators in an expression are evaluated is determined by some simple rules of precedence. The rules of precedence *Statistix* uses to evaluate arithmetic expressions are the same as those used in algebra. If all of the arithmetic operators in an expression are of equal precedence, they are evaluated in order from left to right. However, not all operators share equal precedence. The following table ranks the arithmetic operators according to precedence.

Highest Precedence:	- (unary negation)
	^
	∗, /
Lowest Precedence:	+, -

Unary negation has the highest precedence and is always performed first if it occurs anywhere in an expression. Exponentiation (^) is performed next, followed by multiplication (∗) and division (/). Multiplication and division are of equal precedence, so the order in which they appear in the expression determines which is done first. Addition (+) and subtraction (-) share the lowest precedence among arithmetic operators.

Any expression within parentheses is evaluated before expressions outside the parentheses. For example, in the expression (A + 2.75) ∗ B, the sum A + 2.75 is evaluated first and the result is then multiplied by B. Parenthetical expressions can be nested, with the innermost ones being evaluated first. An example is (A + B ∗ (C + D)) ∗ E.

The simplest arithmetic expression is just a constant or variable by itself, in which case no arithmetic operators are involved.

Date and String Arithmetic

Some arithmetic can be performed on dates and strings. A date can be subtracted from another date to compute the number of days between two events. For example, AGE = (“05/23/92” - BIRTH) / 365. You can also

add a constant to a date. Multiplication and division of dates aren't allowed.

Addition can be performed using string variables and constants. For example, the expression FIRST + " " + LAST concatenates the strings in the variables FIRST and LAST with a space in between.

Logical (Boolean) Expressions

When arithmetic expressions are evaluated, they return numerical values by case. On the other hand, logical expressions return the boolean values TRUE or FALSE by case. You, as a *Statistix* user, will never actually see the values TRUE and FALSE. What you will see are the consequences of some action that was based on whether the expression was TRUE or FALSE, for example, whether a case becomes omitted or not (see Omit/Select/Restore Cases for details).

We now introduce two new classes of operators—relational operators and logical operators. These operators are used to construct logical expressions, expressions that take the boolean values TRUE or FALSE when evaluated. The relational operators are:

=	equal to	<>	not equal to
<	less than	<=	less than or equal to
>	greater than	>=	greater than or equal to

The logical operators are NOT, AND, and OR.

Relational operators require arithmetic expressions for arguments—one to the left and one to the right of the operator (remember that the simplest arithmetic expression is just a constant or a variable name). Relational operators return the boolean values TRUE or FALSE when evaluated. Some typical examples of simple logical expressions using relational operators are:

```
A + B > C
A = 999
A ^ 3.45 >= B / C
```

The embedded arithmetic expressions are evaluated before the relational operators. All relational operators have the same precedence.

The logical operators NOT, AND, and OR are used to construct more complex logical expressions. Logical operators require boolean arguments,

or to put it another way, the arguments for logical operators must be logical expressions. (Note: This is why they are called logical operators; they operate on logical expressions.) The NOT operator requires only one argument to the right; both AND and OR require two arguments, one on either side. The truth table below summarizes the action of these operators (T stands for TRUE, and F for FALSE).

ARGUMENT VALUES		VALUE RETURNED		
X	Y	X AND Y	X OR Y	NOT X
T	T	T	T	F
T	F	F	T	F
F	T	F	T	T
F	F	F	F	T

In their most general form, logical expressions are built with relational and, when needed, logical operators. Some further examples of logical expressions are:

```
(A > B) AND (A = 1)
NOT ((A + B) > C)
((A = B) AND (B = C)) OR ((A <> D) AND (A < 1.96))
```

There are often many ways to express the same condition. Use the one that is clearest to you, not necessarily the most “elegant”. In such expressions, embedded arithmetic operators are evaluated first, followed by relational operators. Logical operators have the lowest precedence of any operator. NOT takes precedence over AND and OR; AND is evaluated before OR. The order of evaluation is easy to control with the use of parentheses. Be careful to use enough parentheses to insure that things are evaluated in the intended order. The following table summarizes the precedence ordering of *Statistix* operators and built-in functions.

Highest Precedence:	parenthesized expressions
	built-in functions
	- (unary negation)
	^
	*, /
	+, -
	=, >, <, >=, <=, <>
	NOT
	AND
Lowest Precedence:	OR

Machine Precision and Tests for Equality

Computers do not perform decimal arithmetic exactly. While the rounding error that occurs during arithmetical operations is usually negligibly small, there is one situation where it is extremely important: tests for exact equality. For example, suppose you want to perform the transformation Y=9∗X/9. You might expect Y to equal X, but it may not because of very small rounding errors. Therefore, it's not certain that the logical comparison IF Y = X will return the value TRUE. It's safer to perform the comparison using the expression IF Abs (X - Y) < d, where d is some small number and Abs is the absolute value function.

Handling of Missing Values

Missing values require special consideration when arithmetical or logical expressions are evaluated. In arithmetical expressions, if any of the arguments have the value missing, the expression is automatically evaluated as missing. This is only reasonable—you can't perform arithmetic on numbers that don't exist.

Logical expressions are somewhat trickier. Different actions are taken depending on the context in which the relational expression is used. The following truth table shows the rules used with missing values for Omit/Select/Restore Cases.

ARGUMENT VALUES		VALUE RETURNED					
X	Y	X=Y	X<>Y	X>Y	X<Y	X>=Y	X<=Y
M	NOT M	F	T	F	F	F	F
M	M	T	F	F	F	F	F

The If-Then-Else construct has the same truth table for X = Y and X <> Y. However, it avoids the issue for the other inequalities by assigning the target variable the value missing.

You may think it would make sense to always assign the target variable the value missing. After all, how do you decide if two unknown values are equal? The reason we use these rules is so you can manipulate the missing values. For example, perhaps you want to replace all of the missing values with 0.0:

```
IF X = M
THEN X = 0
ELSE X = X
```

You should be aware, however, that sometimes you may not be satisfied with these rules. Consider, for example:

```
IF X = Y  THEN NEWVAR = 1 ELSE NEWVAR = 2
```

If both X and Y are missing, *Statistix* will evaluate the ELSE expression and assign 2 to NEWVAR. You really want NEWVAR to be missing as well. This is easily done with a second transformation, as shown below.

```
IF (X = M) OR (Y = M)
THEN NEWVAR = M
ELSE NEWVAR = NEWVAR
```

Remember that if the value for any variable in an arithmetic expression is missing, the expression is evaluated as missing. This can be exploited in a variety of ways. For example, suppose you have the variables A, B, C, D, and Y, and you want Y to be missing whenever A, B, C, or D is missing. One way of doing this is:

```
IF (A = M) OR (B = M) OR (C = M) OR (D = M)
THEN Y = M
ELSE Y = Y
```

But a more compact method is:

```
IF A + B + C + D = M  THEN Y = M  ELSE Y = Y
```

Built-in Functions

There are 66 built-in functions that can be included in arithmetic expressions. Most of them require arguments. The arguments of most functions can be any valid arithmetic expression. This will be represented as "x" in the following descriptions. Note that x can include built-in functions, including the function it is an argument for, such as Sqrt (Sqrt (x)). If x has the value missing when it is evaluated, the function is also assigned as missing. In a few cases, an integer constant is required for an argument, such as CAT (5,1). Integer constants are represented as "i" or "j" in the following descriptions. String constants or variables are represented as "s". The row functions (e.g., Rowmean) require a variable list for an argument and are represented as "v1..vn" below. Functions Case, M, Pi, Random, and Selcase do not require input arguments. The available built-in functions are listed in the table on the next page.

The function names appear in a list box when you are using the Transformations or Omit/Select/Restore Cases procedures. Functions can be selected directly from the ***Functions*** list box and inserted into the expression box. You can also type in the function name into an expression, in which case you can abbreviate the function name using enough of the first characters to distinguish it from the other function names.

Abs(x)	Exp(x)	NRandom(m,sd)	SD(x)
Angle(x)	Factorial(i)	Number(s)	SelCase
Arcsin(x)	GeoMean(x)	Percentile(x)	Sin(x)
Arctan(x)	Insert(s1,s2,i)	Pi	Sqr(x)
Atkinson(x)	Lag(x[,i])	Pos(s1,s2)	Sqrt(x)
Capitalize(s)	Length(s)	Power(x,y)	String(x[,i])
Case	Ln(x)	Random	Studentize(x)
Cat(i,j)	Log(x)	Rank(x)	Tan(x)
Copy(s,i,j)	LowCase(s)	Round(x)	Total(x)
Cos(x)	M	RowCount(v1..vn)	Trunc(x)
Count(x)	Max(x)	RowMax(v1..vn)	Unitize(x)
CumSum(x[,k])	Mean(x)	RowMean(v1..vn)	UpCase(s)
Date(s)	Median(x)	RowMedian(v1..vn)	Variance(x)
Day(x)	Min(x)	RowMin(v1..vn)	Year(x)
DayofWeek(x)	Modulo(i,j)	RowSD(v1..vn)	ZInverse(x)
Delete(s,i,j)	Month(x)	RowTotal(v1..vn)	ZProb(x)
Diff(x[,i])	Normalize(x)		

Most of the functions operate using one case at a time. For example, Sqrt(x) computes the square root of the expression x, case by case. Some functions compute a single value for an entire column x. The Mean (x) is an example of this type; it computes the mean of the column x. Column functions are normally used as part of a larger expression (e.g., x - Mean(x)). The nine column functions are Count, Geomean, Max, Mean, Median, Min, SD, Total, and Variance.

Most of the functions expect numerical arguments and return numerical results. There are ten functions that require string arguments (Upcase, Lowcase, Length, Pos, Copy, Delete, Insert, Capitalize, Date, and Number) and four functions that require date arguments (Day, Month, Year, and Dayofweek).

A description of each of the functions follows.

Abs (x) Absolute value of x.

Angle (x) Computes the angular transformation (also called the arcsin-square root transformation) for proportions. Proportions near 0 or 1 are spread out so as to increase their variance (Snedecor and Cochran, 1980). The argument x must be a proportion between 0 and 1. You can apply the function to a percentage by first dividing the percentage by 100.

Arcsin (x) Arcsine of x in radians.

Arctan (x) Arctangent of x in radians.

Atkinson (x) Computes the transformation for the Atkinson score method used to determine what power transformation, if any, is needed in a linear regression analysis (Weisberg, 1985). The argument of the transformation is the dependent variable in the regression analysis. The transformed variable is added as an independent variable in the regression analysis to test for a power transformation (see Weisberg for an example).

Capitalize (s) Capitalizes the first letter of each word in the string s and converts all other letters to lowercase.

Cat (i, j) Categorical index generator. This function generates index values. The integer argument i gives the number of categories in the index, and the integer argument j is the repeat factor. This function generates the numbers 1 through i, repeating each value j times. First j 1's are generated, then j 2's, etc., up to j i's. After i x j values are generated, the process repeats.

Both i and j must be specified as positive integers.

Some examples are shown below. The example assumes that there are 12 cases in the data set, and all have been selected.

Y = CAT (3,2) Y : 1,1,2,2,3,3,1,1,2,2,3,3

Y = CAT (3,3) Y : 1,1,1,2,2,2,3,3,3,1,1,1

Y = CAT (4,3) Y : 1,1,1,2,2,2,3,3,3,4,4,4

Y = CAT (5,1) Y : 1,2,3,4,5,1,2,3,4,5,1,2

Y = CAT (2,4) Y : 1,1,1,1,2,2,2,2,1,1,1,1

Although this may initially appear to be a rather simple-minded function, it is extremely useful for creating categorical variables needed for numerous types of analyses. For example, suppose you have data from a randomized block design with 4 blocks and 3 treatments applied within each block. Then suppose your data are ordered such that the 3 observations for block 1 came first, followed by the 3 observations for block 2, etc. To create your block index, you would specify: BLOCK = CAT (4,3). The treatment index is created as: TREAT=CAT (3,1).

Note that if the data had been ordered such that the 4 values for treatment 1

came first, followed by the 4 values for treatment 2, etc., the indices would be generated as BLOCK = CAT (4, 1) and TREAT = CAT (3, 4).

Copy (s,i,j) — Copies j characters of text from the string s starting at the i-th character.

Case — Case index. Case indicates the position of a value in the data set. If the data set is thought of as a rectangular table of numbers, the variable names identify the columns and the case indices identify the rows. Case is incremented for cases that are omitted. The Selcase function discussed on page 60 skips omitted cases.

Cos (x) — Cosine of x. Units of x are assumed to be radians. Angles can be converted from degrees to radians as: RADIANS = DEGREES / 180 * PI.

Count (x) — Number of usable cases of x. This function returns the total number of usable cases (i.e., selected and not missing). It is not a running counter; please see Case above for such a function.

Cumsum (x[,k]) — This function is two functions in one. When used with a single argument x, it computes the running sum of x. The value it returns for the i-th case is the sum of the first i cases of x.

When used with the two arguments x and k, it computes the decision interval cusum. The value returned for the i-th case S_i is defined as $S_i = \max (0, S_{i-1} + x_i - k)$.

Date (s) — Converts the string value s to a date value. This function is used to convert string variables to date variables.

Day (x) — Day of month. The argument must be a date.

Dayofweek (x) — Day of week (1 = Sunday, 2 = Monday, etc.). The argument x must be a date.

Delete (s,i,j) — Deletes j characters of text from the string s starting at character index i.

Diff (x, i) — Difference of x and x lagged by i cases. The value of this function for the j-th case is x at case j minus x at case j-I. The argument i may be omitted, in which case i = 1 is assumed. See also Lag (x, i).

Factorial (x) — Computes x!

Exp (x)	Exponentiation; e raised to the power x.
Geomean (x)	Geometric mean of the column x.
Insert (s1,s2,i)	Inserts the string s1 into the string s2 at character index i.
Lag (x, i)	Lag x by i cases. The value of this function for the j-th case is the (j - i)-th case of x. Cases 1 up to i receive the value missing. Only selected cases are treated. If the (j - i)-th case of x is missing, Lag will return the value missing for the j-th case. The integer lag factor i is optional; it defaults to 1 if omitted.
Length (s)	Computes the length (number of characters) in the string s.
Ln (x)	Natural (base e) log of x.
Log (x)	Base 10 log of x. Use the Power function discussed on the next page to compute antilogs (e.g., Power (10, x)).
Lowcase (s)	Converts the string s to all lowercase.
M	Missing value indicator. This can be used to assign a variable a missing value (e.g., VAR = M). It can also be used to test a variable for missing data (e.g., IF VAR = M).
Max (x)	Maximum value of x over all selected cases.
Mean (x)	This function returns the mean of x over the selected cases.
Median (x)	This function returns the value of the median for x over all selected cases.
Min (x)	This function returns the minimum value of x over all selected cases.
Modulo (x, y)	Computes the modulus of x by y (the remainder of x divided by y). x and y may be expressions but must have integer values less than 99,999. Example: Modulo (12, 5) = 2.
Month (x)	The month of the year. The argument x must be a date.
Normalize (x)	Normalize scales x such that the sum of all values of x equals 1.

NRandom (m, sd)	Normal random number generator. Generates a normal random number with mean m and standard deviation sd. See also Random.
Number (s)	Converts the string value s to a number. This function is used to convert string variables containing numeric strings to real or integer variables.
Percentile (x)	Percentile of x. Suppose you have a variable SCORE that contains the test scores for all students in a class. The transformation PSCORE = Percentile (SCORE) creates a new variable PSCORE that contains the percentiles of each test score.
Pi	The constant pi, 3.1415926.
Pos (s1, s2)	Returns the starting position of the string s1 where it appears in the string s2. If the string s1 does not appear in the string s2, zero is returned. Example: Pos (“def”, “abcdefg”) = 4.
Power (x, y)	Raises the value x to the power y. Produces the same result as x ^ y. This function is defined for nonnegative values for x and for negative values for x when y is a positive whole number.
Rank (x)	Ranks of x. The value of this function for the I-th case of x is the rank of that case. If some cases are tied, these cases receive the appropriate average rank.
Random	Generates a uniformly distributed random number on the interval 0.0 to 1.0. You change the scale of the random numbers produced by multiplying the function result by a constant. See also NRandom.
Round (x)	Rounds x to the nearest whole integer number.
Rowcount (v1..vn)	Counts the number of nonmissing values in the variable list for each case. As with the other row functions described below, the variable list can include VAR1 .. VAR99 syntax and the keyword ALL.
Rowmax (v1..vn)	The maximum value among the variables listed for each case.
Rowmean (v1..vn)	The mean of the variables listed for each case. The mean is computed ignoring missing values.

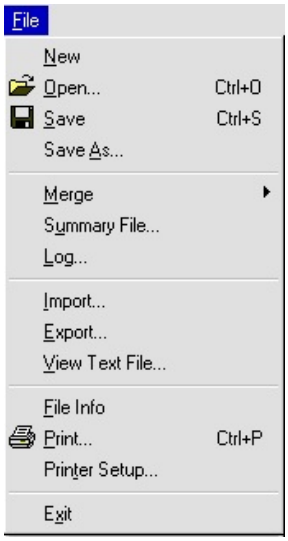
Rowmedian (v1..vn)	The median of the variables listed for each case. The median is computed ignoring missing values.
Rowmin (v1..vn)	The minimum value among the variables listed for each case.
RowSD (v1..vn)	The standard deviation of the variables listed for each case.
Rowtotal (v1..vn)	The sum of the variables listed for each case. The total is computed ignoring missing values.
SD (x)	Standard deviation of x. This function returns the sample standard deviation for the column x. This is the so-called unbiased estimate of the standard deviation; the divisor is the square root of n - 1, where n is the number of usable cases.
Selcase	Selected case index. Selcase indicates the position of a value in the data set with respect to the selected cases. If the data set is thought of as a rectangular table of numbers, the variable names identify the columns and the case indices identify the rows. Selcase is not incremented for cases that are omitted. Selcase ranges from one to the maximum number of selected cases. Case counts all cases regardless of their omit status.
Sin (x)	Sine of x. Units of x are assumed to be radians.
Sqr (x)	Squares the value of x.
Sqrt (x)	Square root of x.
String (x[,i])	Converts a number or date to its string equivalent. The optional constant i specifies the number of decimal places when converting numbers.
Studentize (x)	Studentizes x. A variable is studentized by subtracting the sample mean from the original values and dividing the deviations from the sample mean by the sample standard deviation. Once x has been studentized, it'll have a mean of zero and a standard deviation of one. If x was originally normally distributed, x will be nearly standard normally distributed after studentizing.
Tan (x)	Tangent of x. Units of x are assumed to be radians.
Total (x)	This function returns the sum of all selected cases for x.

Trunc (x)	Truncates the decimal portion of a number. For example, 1.98 when truncated becomes 1.0.
Unitize (x)	Scales x so its vector norm equals one. The norm, or length, of a vector is the square root of the sum of the squares of the elements of the vector.
Upcase (s)	Converts the string s to all uppercase.
Variance (x)	This function returns the sample variance for x. This is the so-called unbiased estimate of the variance; the divisor is n - 1, where n is the number of usable cases.
Year (x)	Year of a date. The argument x must be a date.
Zinverse (x)	Inverse of the standard normal distribution. If the value of the argument x is between 0 and 1, Zinverse returns the inverse of the standard normal distribution. That is, the value returned is the z value (standard normal value) for which the probability of a smaller value is the value of x.
Zprob (x)	The standard normal probability of x. This function returns the probability of a value smaller than x from a standard normal distribution. In other words, this function returns the lower tail probability.

C H A P T E R

3

File Menu



Statistix offers flexible and easy-to-use file management procedures. These procedures are used to manipulate data files stored on fixed disks, diskettes, and CDs. You'll use the file management procedures to open and save data files, import data created by other programs, and view text files without leaving *Statistix*.

Data that you create using *Statistix* are temporary. Data that you enter, import, or create using transformations are not stored on disk until you

explicitly create a disk file to permanently store your *Statistix* data.

The **New** procedure is used when you want to create a new *Statistix* data file. It closes the current data set and displays an empty spreadsheet.

The **Save** and **Save As** procedures are the usual methods of saving *Statistix* data. These procedures create a high speed compact binary file representation of a *Statistix* data set. *Statistix* data files have the file name extension ".SX".

The **Open** procedure is used to retrieve *Statistix* data files created previously using the Save and Save As procedures.

The **Merge** procedures combines data from your active data set and a second data set stored in a *Statistix* data file. There are three separate procedures to merge (1) cases, (2) variables, and (3) labels, transformation expressions, and dialog box settings.

The **Summary File** procedure is used to create a new *Statistix* data file containing summary statistics of the active data set.

The **Log File** procedure is used to start recording a log of the *Statistix* procedures you will perform.

The **Import** procedure is used to read data from text, *Excel*, *Lotus 1-2-3*, *Quattro Pro*, *Access*, *dBase*, and *Paradox* files into *Statistix*.

The **Export** procedure is used to create text, *Excel*, *1-2-3*, *Quattro Pro*, *Access*, *dBase*, or *Paradox* file versions of your current *Statistix* data set so that the data can be accessed by other programs.

The **Print** procedure is used to print the contents of the active window. The active window can be the data set window, a report, or a graph.

The **File Info** procedure produces a report that lists details about the open file, such as, variable names, variable types, column formats, variable labels, and value labels.

The **Printer Setup** procedure is used to select the printer you want to use, and to select printer options, such as, page orientation.

The **View Text File** procedure is used to view the contents of text files.

New

The **New** procedure is used when you want to start a new *Statistix* data set. It closes the open data set, if any, and displays an empty spreadsheet. You then add variables to the new data set using either the **Import** or **Insert Variables** command.

If you have not saved the active data set since it was last modified, a warning message appears on the screen. This warning gives you an opportunity to save the active data set before erasing it from memory.

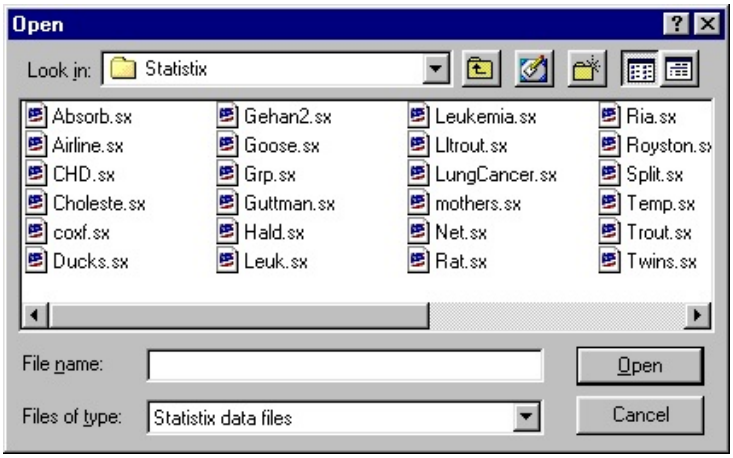
Press the ***Yes*** button to save your data, then close. If your data hasn't been given a file name yet, you'll be prompted to enter one. Press the ***No*** button to close without saving. Press the ***Cancel*** button to cancel the New command altogether.

Open

The **Open** procedure is used to read a *Statistix* data file previously created using the **Save**, **Save As**, or **Summary File** procedure. The file data set becomes the active *Statistix* data set and is displayed on the spreadsheet. This procedure can't be used to open data stored in formats other than the *Statistix* format. Use the Import procedure described later in this chapter to import data from text, *Excel*, *Lotus 1-2-3*, *Quattro Pro*, *Access*, *dBase*, and *Paradox* files.

Statistix can only have one data set open at a time. If you already have a data set open, it will be replaced with the data contained in the file you open. If you haven't saved the active data set since it was last modified, *Statistix* will warn you and give you a chance to save it.

The Open dialog box is similar to the open dialogs of other *Windows* applications. There's a ***File name*** edit control where you can enter the name of the file you want to open. There's a file list that lists the *Statistix* files in the current folder. The name of the current folder is displayed in the drop-down list labeled ***Look in***.



Your cursor starts at the ***File Name*** edit control. You can enter the name of the file you want to open, and then press the ***OK*** button to open the file.

You can also select a file from the file list. Double click on the name of the file you want to open. Use the scroll bar to scroll down the list if the file you want isn't visible. You can change the file list to a different folder using the ***Look in*** drop-down list. Click on the down arrow on the Look In list to display the list of drives and folders on your computer.

Save

Statistix data files are ideal for saving your *Statistix* data for future *Statistix* analyses. All information about your *Statistix* data is preserved, including variable names, case omit status, missing values, value labels, and *Statistix* dialog box settings. The **Save** takes a "snapshot" of the data set's present state for future use. *Statistix* data files store the data in a compact binary format—these files can be read and written rapidly.

The current file name of your active data set is displayed in the spreadsheet's title bar. If you've just created a new data set, then the work "Untitled" appears on the title bar.

If your data set is untitled when you select the Save procedure from the menu, the Save As dialog box described below is displayed and you'll be asked to enter a file name. If your data set already has a file name, then that file is updated with the current state of your data.

Use the Save As procedure described below if you'd like to save a named data set but with a different file name.

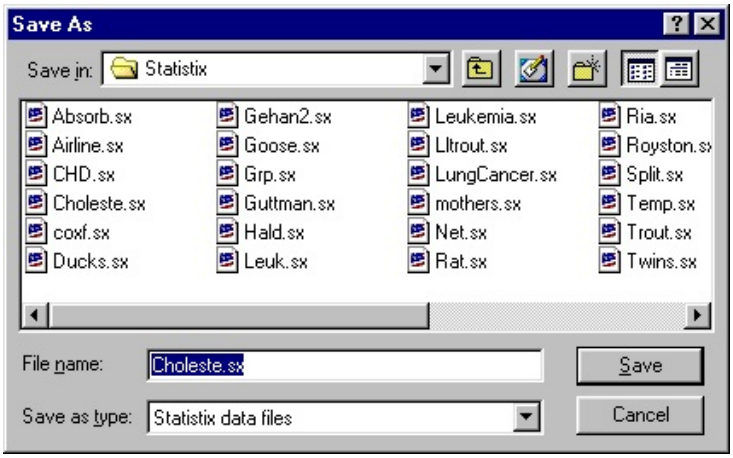
Save As

Use the **Save As** procedure to save your *Statistix* data for future *Statistix* analyses. All information about your *Statistix* data is preserved, including variable names, case omit status, missing values, value labels, and *Statistix* dialog box settings.

The current file name of your active data set is displayed in the spreadsheet's title bar. If you've just created a new data set, then the word "Untitled" appears on the title bar.

Use the **Save As** procedure to save an untitled data set for the first time, or to save a titled data set using a different file name. Use the **Save** procedure discussed on the preceding page if you want to update the file using the name that appears on the spreadsheet title bar.

The Save As dialog box has a ***File name*** edit control where you can enter the name of the new file. There's a file list that lists the *Statistix* files in the current folder. The name of the current folder is displayed in the drop-down list labeled ***Save in***.



Your cursor starts at the ***File name*** edit control. You can enter the name of the file you want to create, and then press the ***OK*** button to save the file. If you don't specify a drive or path, the current drive and path as displayed in the ***Save in*** box are used. The ".SX" file name extension will be added for you.

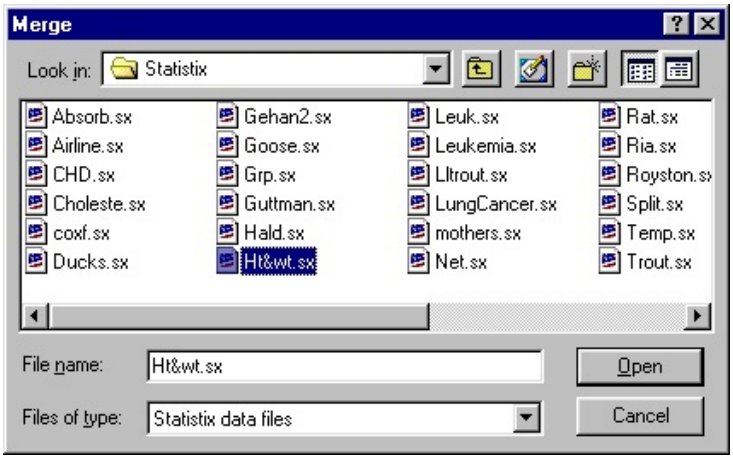
More often than not you'll want to use a new file name when using the Save As procedure. But you can select a file from the file list. Double click on the name of the file you want to use to save your data. Use the scroll bar to scroll down the list if the name you want isn't visible. You can change the file list to a different folder using the ***Save in*** drop-down list. Click on the down arrow on the ***Save in*** list to display the list of drives and folders on your computer.

Statistix 8 files can't be opened using earlier versions of *Statistix*. Click on the ***Save as type*** drop down box to select an older file format. *Statistix 4* data files can be opened using *Statistix 4.0* and later versions (including *Statistix for Windows 1* and *Statistix for Windows 2*). *Statistix 7* data files can be opened using *Statistix 7.0* and later versions. Saving data using an older format will result in the loss of some information. *Statistix 7* data files don't store the transformation expressions history list. *Statistix 4* data files don't store dialog box settings.

Merge Cases

The **Merge Cases** procedure is used to combine the data of your active data set with data stored in a second *Statistix* data file. Cases of data stored in a file are appended to the existing variables in the active data set.

There are two steps to using this procedure. The first step is to select the merge file using the dialog box displayed below.



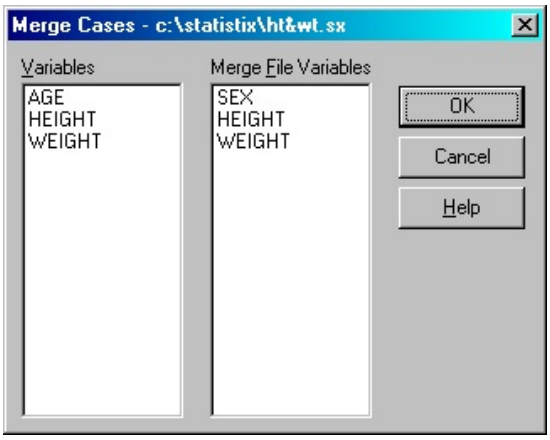
Your cursor starts at the ***File name*** edit control. You can enter the name of the file you want to merge with your current data or double-click on a file , and then press the ***Open*** button.

You can also select a file from the file list. Double click on the name of your merge file. Use the scroll bar to scroll down the list if the file you want isn't visible. You can change the file list to a different folder using the ***Look in*** drop-down list. Click on the down arrow on the Look in list to display the list of drives and folders on your computer.

Once you've specified the name of the merge file (HT&WT in our example), a second dialog box is displayed, as shown on the next page. The variables in your current data set are listed in the ***Variables*** list. The variables found in the merge file are listed in the ***Merge File Variables*** box.

The active data set has three variables named AGE, HEIGHT, and WEIGHT. The file HT&WT has three variables named SEX, HEIGHT, and WEIGHT. Only cases of variables that match variable names in your

current data set can be appended (HEIGHT and WEIGHT in this example).



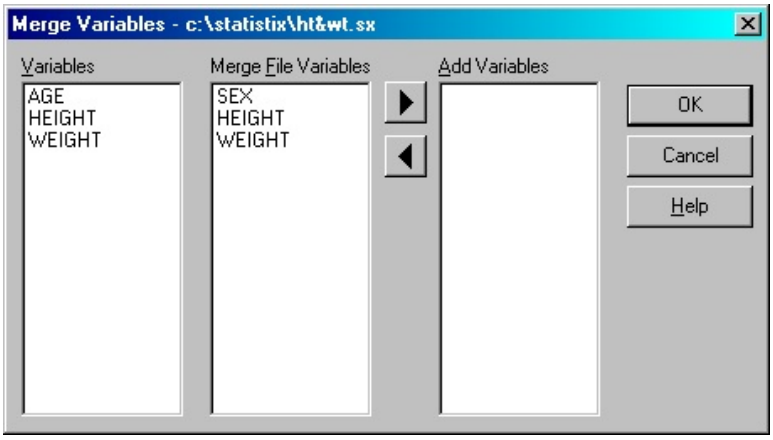
Press the ***OK*** button to start the merge. Once the merge is completed, the data for the HT&WT file variables HEIGHT and WEIGHT will be added at the bottom of the spreadsheet to the active data set variables HEIGHT and WEIGHT. Since the HT&WT variable SEX does not match any variables in the current data set, its data will be ignored.

Merge Variables

The **Merge Variables** procedure is used to combine the data of your active data set with data stored in a second *Statistix* data file. Variables of data stored in a file are added to the existing variables in the active data set.

This procedure is best explained with an example. There are two steps. The first step is to select the merge file. The dialog box and procedure for selecting a file are the same as the Merge Cases procedure discussed on the preceding page.

Once you've specified the name of the merge file (we'll use the same HT&WT file for this example), a second dialog box is displayed, as shown on the next page.



The variables in your current data set are listed in the ***Variables*** list. The variables found in the merge file are listed in the ***Merge File Variables*** box. Select the variables you want to merge from the Merge File Variables box and move them to the ***Add Variables*** box. Only variables with names that don't match variables in your current data set can be selected (SEX is the only such variable in this example).

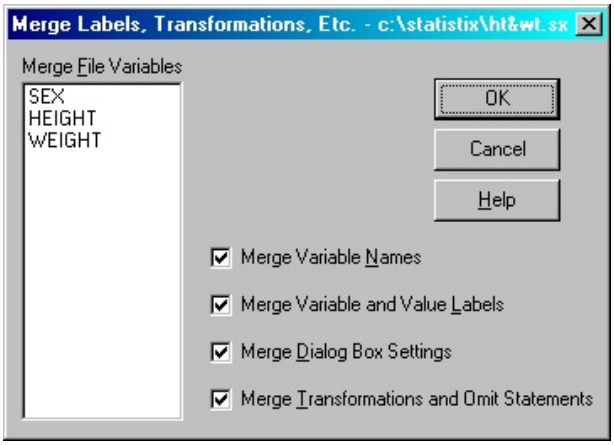
Press the ***OK*** button to start the merge. Once the merge is completed, there'll be four variables in the current data set—AGE, HEIGHT, WEIGHT, and SEX.

Merge Labels, Transformations, Etc.

A *Statistix* data file stores a table of data consisting of numbers, dates, and strings. But it also stores other useful information that you may want to share between separate data files: variable names, variable labels, value labels, dialog box settings, transformation expressions, and omit expressions. The **Merge Labels, Transformations, Etc.** procedure is used to merge these kinds of data into an active data set. Unlike the Merge Cases and Merge Variables procedures discussed on the preceding pages, this procedure can also be used to import data into a new data set. In this way, any *Statistix* data file can be used as a template to create a new data file with the same structure, but with a new table of values.

There are two steps to using this procedure. The first step is to select the merge file. The dialog box and procedure for selecting a file are the same as the Merge Cases procedure discussed on page 69.

Once you've specified the name of the merge file, a second dialog box is displayed, as shown below.



There are four check boxes on the dialog box, one for each kind of data that can be merged into the active data set using this procedure. Simply check the boxes for the types of data you want to merge.

You can merge variable names into an active data set with other variables, or with an empty data set that doesn't have any variables yet. The case data for merged variable names are ignored.

When merging variable labels and value labels, labels for any variables in the merge file that match the names of variables in the active data set replace the original labels, if any, in the active data set.

Data box settings refer to the values of edit controls, list boxes, check, boxes, etc. that appear on data, file, and statistics dialog boxes.

Beginning with *Statistix 8*, lists of transformation and omit cases expressions performed on a data set are saved along with the rest of the data when you use the **Save** procedure to create a *Statistix* data file. These expressions can be reused by the **Transformations** and **Omit/Select Restore Cases** procedures (see Chapter 2). By merging these lists from one data file into the active data set, you can reuse these expressions without having to retype them.

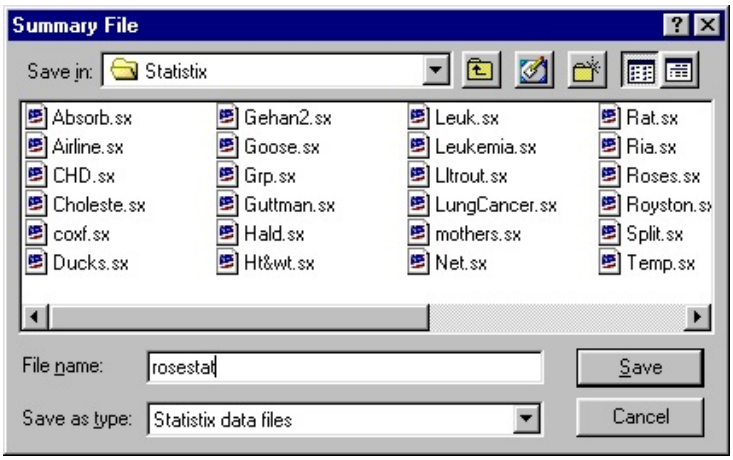
Summary File

The **Summary File** procedure is used to create a new *Statistix* data file containing summary statistics of your active data set. Summary statistics, such as the mean and standard deviation, can be computed for selected variables broken down by one or more categorical variables. The file will contain one case for each unique combination of values for the categorical variables. The current data set isn't modified by the procedure. Use the **Open** procedure to retrieve the summary statistics for further processing.

Be sure not to confuse this procedure with the **Save** procedure, which is used to save an exact copy of your data set.

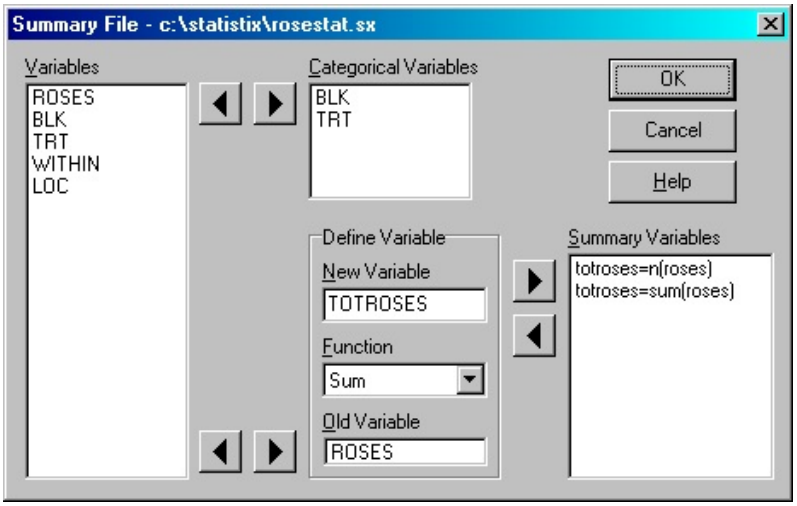
Specification

You must first specify a name for the file you want to create using the dialog box below.



Your cursor starts at the ***File name*** edit control. You can enter the name of the file you want to create, and then press the ***OK*** button. If you don't specify a drive or path, the file will be saved in the folder displayed in the ***Save in*** box. If you don't use the ".SX" file name extension, it'll be added for you.

Once you've specified the file name, you're presented with the Summary File dialog box used to specify the details about the new file, as shown on the next page.



First select the ***Categorical Variables***. The categorical variables contain discrete values that are used to identify groups. The categorical variables can contain numbers, dates, or strings. These variables automatically become variables in the new data file. To select the categorical variables, first highlight the variables you want in the ***Variables*** list, then press the right-arrow button next to the ***Categorical Variables*** list box.

The next step is to define summary variables. The summary variable definition has three parts: the summary variable name, a statistical function name, and the existing variable used to compute the new variable. Enter a new name in the ***New Variable*** edit control. Select a function name from the ***Function*** pull-down list. Select the ***Old Variable***: Highlight a variable in the Variables list box, then press the right-arrow button next to the Old Variable box to copy the variable name. Once you've completed all three parts, press the right-arrow button next to the ***Summary Variables*** list to add the summary variable definition to the list.

The example dialog box above shows the definition for the summary variable TOTROSES. The variable TOTROSES is a new variable that will be included in the summary file. It will be computed using the Sum function based on the current data set variable ROSES.

There are nine statistical functions that can be used with the Summary File procedure:

N	number of observations in the group with non-missing values
Missing	number of observations in the group with missing values
Mean	mean
SD	standard deviation
Min	minimum value
Max	maximum value
Sum	sum
Variance	variance
SE	standard error of the mean

Data Restrictions

There must be at least one and no more than five categorical variables. Numeric values of categorical variables cannot exceed 99,999 and will be truncated to whole numbers. String values of a categorical variable will be truncated to ten characters.

Example

To illustrate the summary file procedure, consider the data from a split block design where the number of saleable roses was counted (Bingham and Fienberg, 1982). Five treatments were applied in two replicates.

CASE	ROSES	BLK	TRT	WITHIN
1	102	1	1	1
2	M	1	1	2
3	84	1	2	1
4	81	1	2	2
5	67	1	3	1
6	83	1	3	2
7	71	1	4	1
8	M	1	4	2
9	53	1	5	1
10	M	1	5	2
11	71	2	1	1
12	79	2	1	2
13	76	2	2	1
14	M	2	2	2
15	74	2	3	1
16	M	2	3	2
17	51	2	4	1
18	63	2	4	2
19	63	2	5	1
20	61	2	5	2

Please refer to the Summary File dialog box on the preceding page. The variables BLK and TRT have been selected from the Variables list and copied to the Categorical Variables list box. These two variables will be included in the summary file.

Two summary variables were defined and are listed in the Summary Variables list box. The first variable N is computed using the N function with the variable ROSES as the argument. It will contain the number of observations per category for ROSES. The second variable TOTROSES is computed using the Sum function and the variable ROSES. It will contain the group sums of ROSES.

The resulting data file ROSESTAT.SX contains the following data:

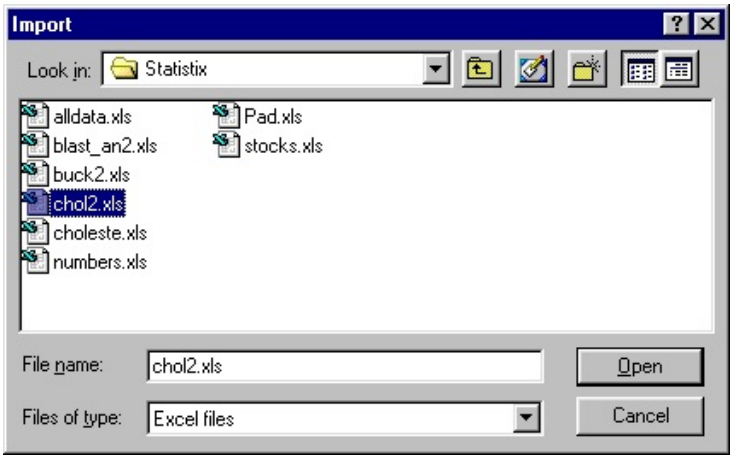
CASE	BLK	TRT	N	TOTROSES
1	1	1	1	102
2	1	2	2	165
3	1	3	2	150
4	1	4	1	71
5	1	5	1	53
6	2	1	2	150
7	2	2	1	76
8	2	3	1	74
9	2	4	2	114
10	2	5	2	124

Import

The **Import** procedure is used to import data from files created by other programs. The file formats supported by *Statistix* are *Excel*, *Lotus 1-2-3*, *Quattro Pro*, *Access*, *dBase*, *Paradox*, and text files. The Import procedure adds variables to a new or existing *Statistix* data set. Each column imported from the file becomes a new variable in your *Statistix* data set.

An alternative method of importing data is to paste data from the *Windows* clipboard directly into the *Statistix* spreadsheet. See Chapter 2 for details.

The first step in importing data is to select an input file.

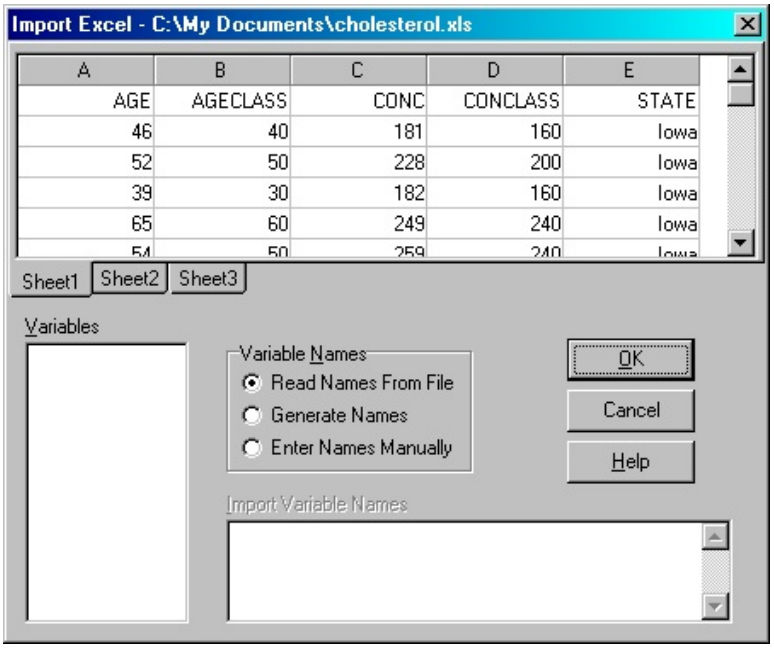


Your cursor starts at the ***File name*** edit control. You can type in the name of the file from which you want to import data, or you can select the file from the list of files. You can change the type of files listed in the file list box by clicking on the ***Files of type*** arrow and selecting a different file type (*1-2-3*, *Access*, etc.). You can select a different drive or folder from the ***Look in*** pull-down list.

Once you've specified the file name, *Statistix* opens one of three types of dialog boxes depending upon the type of the file you selected: one for spreadsheet files, one for data base files, and one for text files. An example dialog box for an *Excel* spreadsheet file is displayed on the next page. An example dialog box for an *Access* data base file is displayed on page 80. The text file dialog box is presented on page 81.

Import *Excel*, *Lotus 1-2-3*, & *Quattro Pro*

After specifying the name of an *Excel*, *Lotus 1-2-3*, or *Quattro Pro* file as discussed on the preceding page, the following Import dialog appears.



The contents of the spreadsheet file are displayed at the top of the dialog box. You can use the scroll bars to view different portions of the file. If the file you select has more than one page, the page names appear in tabs below the rows of data. Click on a page-name tab to import data from that page. You can only import data from one page at a time.

You can import the entire file, select a subset of columns, or select a rectangular array of data. To select one column, click on the column heading for that column. To select a range of columns, drag the mouse over the column headings. To select two or more columns that aren't contiguous, click on the column headings while holding down the control key. You can also select an arbitrary range of cells. Drag your mouse from the top-left cell to the bottom-right cell of the rectangle you want to import. To import the entire file, don't select any columns or cells (to cancel a selection, click on a single cell).

Select one of the ***Variable Names*** radio buttons. You must indicate whether you want to have the variable names imported from the file, have names generated automatically, or enter the names manually. Select a method by clicking on the appropriate radio button.

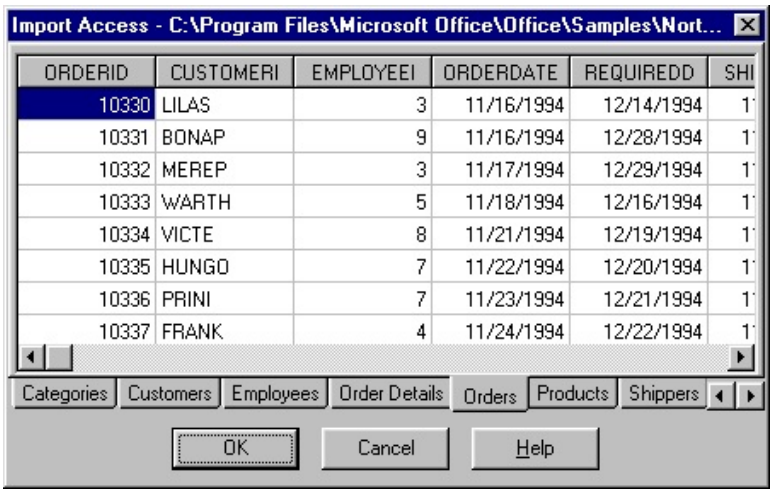
If you select ***Read Names From File***, the first row of the spreadsheet file is scanned for legal variable names. If the text in the cell used for the variable name includes a data type (I for integer, R for real, D for date, and S for string), then the resulting variable will be the data type specified. The letter representing the data type follows the variable name inside parentheses. String types require a number after the letter S to indicate the maximum length (e.g., STUDENT(S15)). If the data type is not specified, the variable type is determined automatically by data found on the first rows of the input file.

If you select ***Generate Names***, then the standard one- or two-letter column headings (A, B, C, . . . AA, AB, AC, etc.) are used for variable names. In this case, the actual data should start in the first row of the file.

If you select the ***Enter Names Manually*** method, you must enter a list of variable names in the ***Import Variable Names*** edit control, one name for each column of data you want to import. You can specify the data types of the variables when you list the variable names. Use the letters I, R, D, and S in parentheses after the variable names to identify integer, real, date, and string types. String types require a number after the letter S to indicate the maximum length. If you don't specify a data type, real is assumed.

One common problem associated with importing data from spreadsheet files is that variables are sometimes assigned an inappropriate data type (real, date, or string). In particular, variables that should be assigned the real type are assigned the string type. Sometimes this happens when there are extra headers in the files, and other times it's caused by blanks appearing in a column. There are several approaches to working around this problem. If you're reading variables names from the first line in the file, try explicitly declaring the data type as part of the variable names (e.g., change AGE to AGE(R)). Another approach is to select ***Enter Names Manually*** and explicitly declare the data types in the variable name list. A third approach is to change the data types of variables one at a time using transformations (e.g., AGE(R) = NUMBER(AGE)).

After specifying the name of an *Access*, *dBase*, or *Paradox* file as discussed on page 77, the following Import dialog appears.

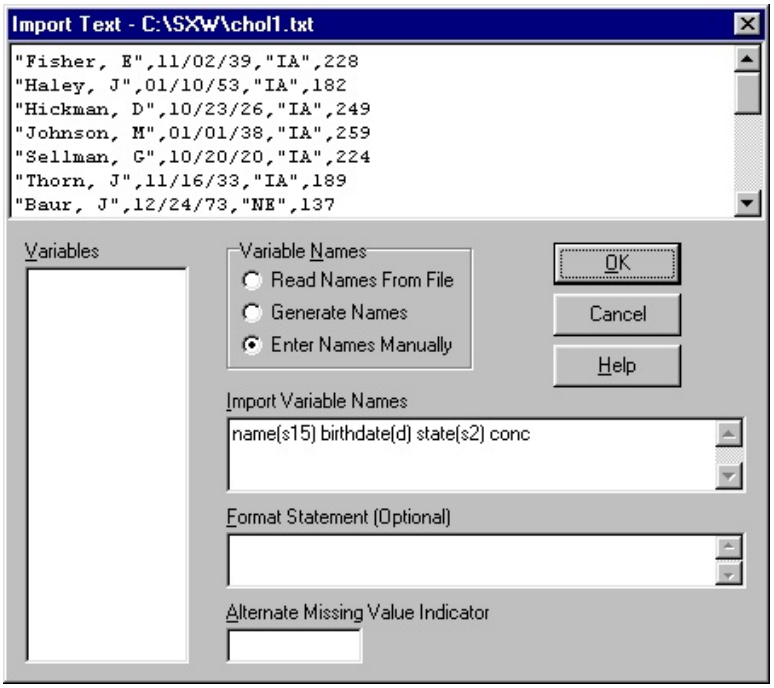


The contents of the spreadsheet file are displayed at the top of the dialog box. You can use the scroll bar to view columns to the right of the display area. If the file you select has more than one table, the table names appear in tabs below the rows of data. Click on a table-name tab to import data from that table. You can only import data from one table at a time.

You can import all the columns from the file, or select one or more columns for importing. To select one column, click on the column heading for that column. To select a range of columns, drag the mouse over the column headings. To select two or more columns that aren't contiguous, click on the column headings while holding down the control key.

Import Text File

After specifying the name of a text file as discussed on page 77, the following **Import Text File** dialog box appears. The contents of the file you selected are displayed at the top of the dialog box. The dialog lists the variables in your active data set, if any, in the ***Variables*** box. It has radio buttons to indicate from where the new variable names will. It has text boxes used to enter new variable names, a format statement, and a string used in the import file to indicate missing values.



The Import procedure is used to add variables to a new or existing data set. You must provide a valid variable name for each column of data you want to import. Use the ***Variable Names*** radio buttons to indicate where you want *Statistix* to find the variable names. Select ***Read Names From File*** if the text file you're importing has variable names on the first line of text. Select ***Generate Names*** to have *Statistix* generate variable names starting with V001. Select ***Enter Names Manually*** if you want to type a list of variable names in the ***Import Variable Names*** text box.

In *Statistix*, you can use integer, real, date, and string data. However, a particular column can only be used to store one type of data. You can specify

the data type of variables when you list the variable names. Use the letters I, R, D, and S in parentheses after the variable names to identify integer, real, date, and string types. String types require a number after the letter S to indicate the maximum length. In the example dialog box on the preceding page, the variable NAME is a string variable with a maximum length of 15. The variable BIRTHDATE is a date variable. The entry for the variable CONC does not declare a data type, so it's assigned the real data type. These data type rules apply to variable names entered manually in the dialog box or read directly from the input file.

You can use the VAR1 .. VAR99 syntax in the variable list to abbreviate a long list of variables. Specify the data type for the entire list of variables by entering the data type at the end of the list (e.g., Q1 .. Q15(I)).

The ***Format Statement*** edit control is required when the fields of the import file are not delimited with comma, spaces, or tabs. The format statement is used to indicate where the data for each variable begins and ends. The format statement is discussed in more detail below.

Statistix interprets the letter M and a period as a missing value. You can enter an additional string (e.g., 999 or N/A) in the ***Alternate Missing Value Indicator*** field to flag missing values.

Comma and Quote Files

A "Comma and Quote" text file is a particular text format made popular by spreadsheet and database programs. In a comma and quote file, columns of data are separated by commas, spaces, or tabs. String data, such as a person's name, are enclosed in quotation marks ("" or ''). One line of text corresponds to one case in *Statistix*. The dialog box on the preceding page shows an example of a comma and quote file. If the file you want to import data fits the description of a Comma and Quote text file, then you needn't use the Format Statement, which greatly simplifies the import process.

Format Statement

The Format Statement in the Import procedure is used to import columns of data from a text file when the data are arranged with fixed field widths. We call files of this type a Formatted file to distinguish them from Comma and Quote files. The format statement is used to specify the exact locations of data for each input variable.

The use of the format statement can be tedious and should be avoided whenever possible. The reasons for using a format statement include:

- Columns of data are not comma, tab, or space delimited.
- String data are not enclosed in quotation marks.
- You want to skip unwanted columns of data.
- The data for one case requires more than one line in the input file.

Consider the formatted file below.

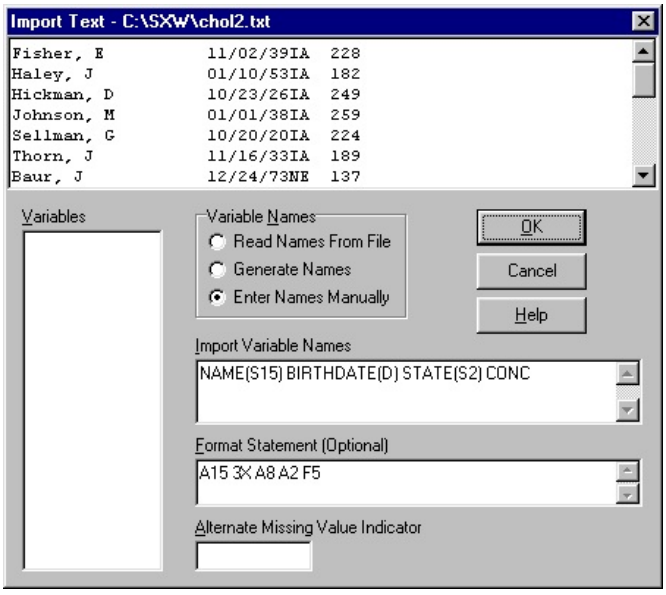
```
Fisher, E          11/02/39IA  228
Haley, J           01/10/53IA  182
Hickman, D         10/23/26IA  249
Johnson, M         01/01/38IA  259
Sellman, G         10/20/20IA  224
Thorn, J           11/16/33IA  189
Baur, J            12/24/73NE  137
Christianson, E    10/23/47NE  173
Farrow, C          10/14/58NE  177
Greer, R           11/28/13NE  241
Keller, G          12/13/40NE  225
Stanley, J         12/15/28NE  337
Steele, A          12/09/72NE  189
Stone, O           11/26/61NE  140
Swanson, D         12/15/44NE  196
Taylor, E          11/20/33NE  262
Thompson, B        11/09/21NE  261
Tucker, T          10/26/24NE  356
Williams, G        10/04/70NE  191
Wright, O          11/21/35NE  197
```

The strings in the first column of data aren't enclosed in quotation marks. Also, there aren't any delimiting characters between the column of dates and the two-letter state abbreviations. A format statement is required to indicate where each column begins and ends.

The Import Text File dialog box to import this file is displayed on the next page. The format statement is:

```
A15 3X A8 A2 F5
```

Each variable requires a format specification. A single format specification consists of a letter followed by a number indicating the field width. There are five format specifications that can be used for variable data. The A format (A for automatic) can be used for string, date, and numerical variables. There are four additional format specifications that can be used for integer and real variables: D - Decimal format, E - Exponential format, F - Fixed format, and I - Integer format. The differences between these types are more important when used with the **Export** and **Print** procedures discussed later in this chapter. Any integer or real variable can be imported using the F format.



In all cases, enter a variable's format specification by typing the letter followed by a number representing the total field width for the variable. The field width can include blank spaces before and after the actual field of data.

In the example above, the format specification A15 is used for the first variable NAME, indicating that the first 15 characters of each input line contains the string value for that variable. The format specification 3X is used in the example to skip three spaces on the input line between the data for the variables for NAME and BIRTHDATE. The last format specification—F5 for the variable CONC—includes two characters for the leading spaces and three characters for the three digit numbers.

The F format can be used to insert a decimal point into a column of numbers at a specific position. Suppose that the column of data for the variable CONC was entered as ten times the actual value, such that the number for the first case 228 represented the value 22.8. By specifying the format F5.1 instead of F5, a decimal point would be inserted one position from the right side.

If field widths repeat, we can use a shorthand notation rFw to reduce the size of the format statement. For example, the format F5.1 F5.1 F5.1 can be abbreviated as 3F5.1.

A repeat factor can be applied to a list of formats inside parentheses too. For example, the format statement A8 I2 F8.2 A8 I2 F8.2 can be abbreviated using the statement 2(A8 I2 F8.2).

The X format is used to skip spaces between variables. The general format is rX where r indicates the number of spaces to skip. For example, we'd use 15X to skip the first 15 characters of the input line and import only the variables BIRTHDATE, STATE, and CONC.

Long records are sometimes split into several lines in an input file. Suppose you wanted to read 17 variables from a file, but the first ten variables were listed on one line and the last seven variables were listed on the following line. You would then need to use the / character in the format statement to show where the line break occurs:

```
10F8 / 7F8
```

Importing a Single Variable

If a file contains the data for a single variable, you can use the "single" option to read all the data items, regardless of how many columns there are in the file. Just enter the word "single" in the space provided for the input format statement. The values will be read left to right, top to bottom.

Comment Lines

Statistix ignores any lines in a text file that begin with the characters "∗", "$", or "#". This lets you add comment lines to your text files.

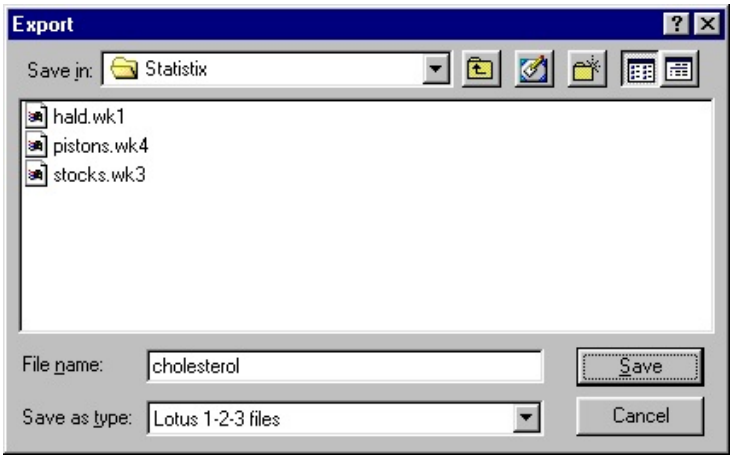
Export

Statistix files can't be accessed directly by other programs. Use the Export procedure to create files containing *Statistix* data that can be used by other programs. The file formats you can export data to are *Excel*, *Lotus 1-2-3*, *Quattro Pro*, *Access*, *dBase*, *Paradox*, and text files. Most programs can accept text files, and many can accept data from one or more of these popular spreadsheet programs.

An alternative method of exporting data to other programs is to use the *Windows* clipboard. You can copy *Statistix* spreadsheet data to the

clipboard and then paste the data into another application.

Specify a name for the file you want to create using the Export dialog shown below.

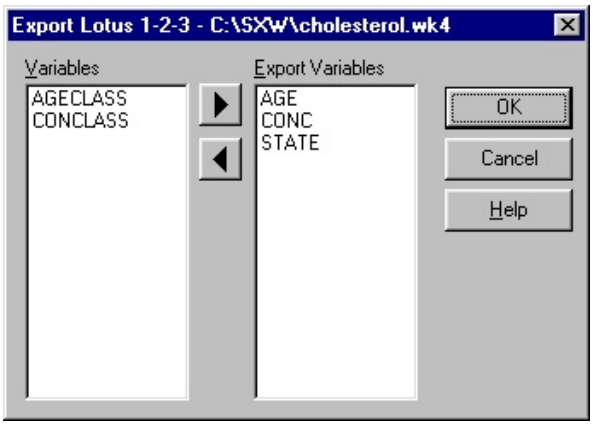


The file you create will be saved in the folder displayed in the ***Save in*** box. You can select a different disk or folder by clicking on the ***Save in*** arrow and making a new selection from the list. Enter a file name in the ***File name*** edit control. If you include a file name extension in the file name, it must be a registered extension (see file type in *Windows* help), or another extension may be added. Select the file type you want to create by clicking on the arrow for the ***Save as type*** box and making your selection from the list.

After you enter a file name, press the OK button. There are three possible Export dialog boxes that will appear depending upon the file type. The dialog box shown on the next page is used for spreadsheet program files (*Excel*, *1-2-3*, and *Quattro Pro*). The dialog box shown on page 88 is used for database program files (*Access*, *dBase*, and *Paradox*). The dialog box of page 89 is used for text files.

Export *Excel*, *Lotus 1-2-3*, & *Quattro Pro*

The dialog box shown below appears after you specify an export file name for an *Excel*, *Lotus 1-2-3*, or *Quattro Pro* file.

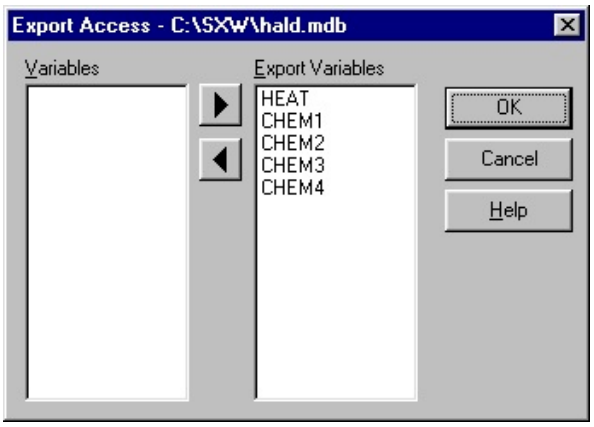


All you need to do now is to select the variables you want to export. Highlight one or more variables in the ***Variables*** list, then press the right-arrow button to move the highlighted variables to the ***Export Variables*** list. You can highlight all the variables in the list by clicking on the first variable in the list, and then dragging the mouse to the last variable. Press the ***OK*** to create the file.

The selected variables will appear in the new file in the order that you've placed them in the Export Variables list. Variable names are listed in the first row of the new file. Omitted cases are not exported.

Export *Access*, *dBase*, & *Paradox*

This procedure is used to export *Statistix* data to a database program file. After you specify the name of an *Access*, *dBase*, or *Paradox* file in the Export dialog box shown on page 86, the dialog box shown below appears.

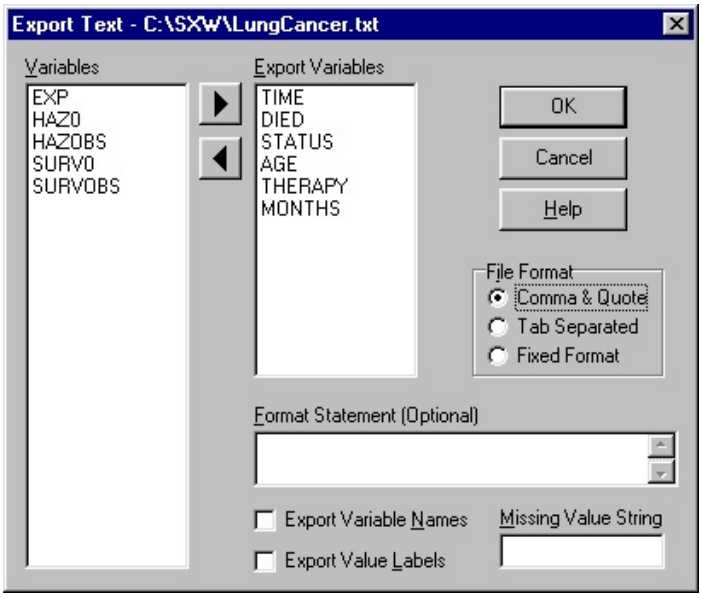


All you need to do now is to select the variables you want to export. Highlight one or more variables in the ***Variables*** list, then press the right-arrow button to move the highlighted variables to the ***Export Variables*** list. You can highlight all the variables in the list by clicking on the first variable in the list, and then dragging the mouse to the last variable. Press the ***OK*** to create the file.

The selected variables will appear in the new file in the order that you've placed them in the Export Variables list. *Statistix* variable names become field names in the new file. Omitted cases are not exported.

Export Text File

An ASCII file is a standard text file format that is commonly used to transfer data between different programs. The dialog box shown below appears after you specify an export file name for a text file as described on page 86.



First select the variables you want to export. Highlight one or more variables in the ***Variables*** list, then press the right-arrow button to move the highlighted variables to the ***Export Variables*** list. You can highlight all the variables by clicking on the first variable, and then dragging the mouse to the last variable in the list. The selected variables will appear in the new file in the order that you've placed them in the Export Variables list.

Next select a ***File Format***. A ***Comma and Quote*** text file separates columns of data using commas, and string data, such as a person's name, are enclosed in quotation marks. An example comma and quote file is displayed on page 81. A ***Tab Separated*** file uses the tab character to separate columns of data.

In a ***Fixed Format*** file, each column has a fixed width such that columns of data line up vertically. String data needn't be enclosed in quotation marks and commas aren't used to separate columns. You have the option of

specifying a ***Format Statement*** that specifies the field widths and numeric formats for each variable. The format statement is discussed in detail below.

Check the ***Export Variable Names*** check box to have the variables included on the first line of the export file. Variable types for variables other than real variables are included in the list of names in a comma and quote file.

Check the ***Export Value Labels*** to have value labels written to the file for numeric variables for which you've defined value labels, rather than the numeric codes themselves.

Format Statement

A format statement is a list of format specifications used to indicate how you want the data for each variable to appear. A variable format specification consists of a letter, a field width, and sometimes a number for decimal places. If you don't enter a format statement, *Statistix* will construct a default format statement using the **Column Format** information discussed in Chapter 2. By omitting the format statement, the data are formatted more or less as it's displayed in the spreadsheet window. The default format guarantees at least one space between variables.

The different format specifications are listed in the table below. In the table, "w" is the field width, "s" is the number of significant digits, "d" is the number of digits to the right of the decimal point, and "r" is the repeat factor (defined below).

Name	General Format	Example Format	Example Appearance	Notes
Automatic	Aw	A10	John Smith 01/05/92 65 12.34567	
Decimal	Dw.s	D11.5	12.345 3.4523E-03	s <= w - 4
Exponential	Ew.s	E10.4	1.234E+01 3.452E-03	s <= w - 4
Fixed	Fw.d	F7.2	12.34 0.003	d < w - 2
Integer	Iw	I2	12 0	
Space	rX	10X		inserts spaces
New line	/	/		inserts line feed

The A format can be used for variables of any data type. The D, E, F, and I formats are used for integer and real variables, each resulting in a different numeric format.

The A format displays numbers using an integer format when the number is a whole number, or a decimal format when the number contains a fraction. For numbers with a fraction, as many digits will be displayed as possible but trimming any trailing zeros. The automatic format works well for data you've entered manually because the numbers generally are displayed just as you've entered them. For computed variables, such as variables created using the Transformations procedure, the A format often displays nonsignificant digits.

The D format displays numbers using a decimal format where the decimal is free to move about to maximize the number of digits displayed.

The F format displays numbers in a decimal format with a fixed number of decimal places.

The E format displays numbers in exponential format, or scientific notation. A number displayed in exponential format has two parts, the mantissa and the exponent. The mantissa is a number displayed as a decimal number that's always greater than or equal to 1.0 and less than 10.0. The exponent is displayed using the letter E followed by a signed integer. The number represented in this fashion is the mantissa multiplied by 10 raised to the exponent. For example, the number 4.23E-02 is equal to 4.23×10^{-2}, or 0.0423. This format is useful for very small and very large numbers.

The I format displays numbers as a whole number. When a number that includes a fraction is displayed using this format, the number is rounded off to the nearest whole number.

The X and / are not used as format specifications for variables but are used to insert spaces and new lines into the output record.

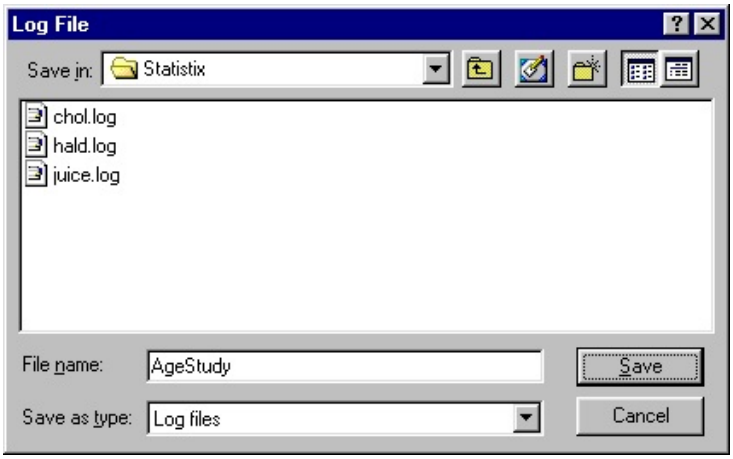
Any of the format specifications can have a number in front called the repeat factor. This is used to abbreviate the format statement when several variables are to be formatted in a similar manner. A repeat factor can also be placed in front of a list of format specifications inside parentheses, as in:

```
3I5 2(I1 1X F4.2 F6.2) E10.4
```

When using format specifications, remember to make “w” large enough to account for the minus sign for negative numbers and for extra space between variables. String data are left justified, so you should always use the X format in front of a format for a string variable to insert a space.

Log File

A log file is a text file that lists the procedures performed during a *Statistix* session. Log files are particularly useful for verifying that a series of transformations or omit cases statements were performed as intended. A log file can be viewed and printed using the *Statistix* **View Text File** procedure during a *Statistix* session to review the work performed. Each procedure is date- and time-stamped so that log file entries can be matched with printed output.



To start a log file, select the **Log File** procedure and enter a file name. If you don’t enter a file name extension, the extension .LOG will be added to the file name.

If you enter the name of a file that already exists, you can choose to have new entries appended to the existing file or you can choose to replace the old file with a new log file.

Once you've started a log file, it continues to record commands until you either start a new log file, or until you exit *Statistix*.

The example log file below lists the procedures used during a short *Statistix* session.

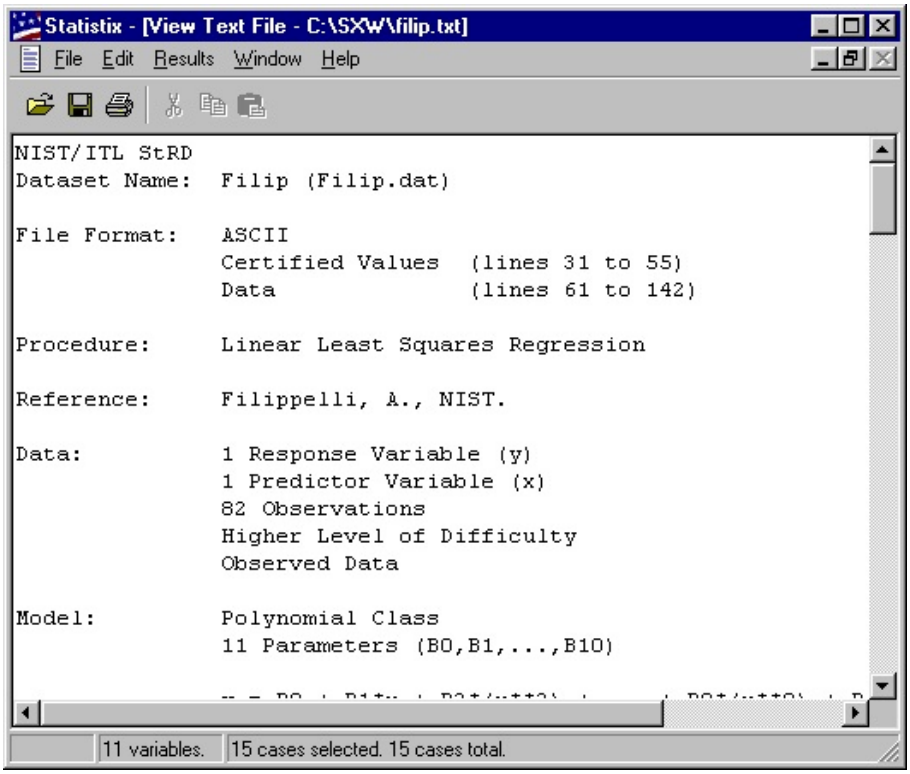
```
Log File, 05/15/00, 15:46
   c:\statistix\agestudy.log
Open, 05/15/00, 15:46
   c:\statistix\choleste.sx
Transformations, 05/15/00, 15:47
   ageclass = 10 * Trunc (age / 10)
Transformations, 05/15/00, 15:47
   conclass = 40 * Trunc (conc / 40)
Print, 05/15/00, 15:48
   AGE,AGECLASS,CONC,CONCLASS,STATE
Histogram, 05/15/00, 15:49
   AGE
   Normal Curve
Cross Tabulation, 05/15/00, 15:50
   AGECLASS,CONCLASS
Save, 05/15/00, 15:52
   c:\statistix\choleste.sx
```

The first entry in the log file shows when the file was started and gives the name of the log file. The remaining entries list the activities that followed: a *Statistix* file named CHOLESTE.SX was opened, two transformations were performed, the data were printed, a histogram was displayed, and a cross tabulation report was obtained. Finally, the modified data set was saved.

View Text File

This procedure is used to view a text file on the screen. There are a number of text files you may be interested in viewing without leaving *Statistix*. You may want to check the contents of a *Statistix* log file to refresh your memory about transformations you've made. Or you may want to review a *Statistix* report file you've saved after running an analysis of your data. You may even want to look at a text file you want to import data from.

Once you've selected a file to view from the standard open dialog box, the file is displayed in a window on the screen, as shown on the next page.



You can scroll forward and backward through the file using the scroll bar and the page up and page down keys. You can print the file by selecting Print from the File menu. You can select another file for viewing by selecting Options from the Results menu.

File Info

File Info is a report that provides basic information about the *Statistix* file you currently have open. The report lists variable names, variable data types, column formats, variable labels, and value labels. The report is first displayed in a window on the screen, but can also be printed or saved in a file. An example report is shown on the next page.

```
File Name   C:\Statistix\Choleserol.sx
File Label  Cholesterol concentration/age study of women from two states

Variables          5
Selected Cases    30
Omitted Cases      0
Total Cases       30

Variable    Data Type   Format   Variable Label/Value Labels
AGE         Real        F 9.0
CONC        Real        F 9.0    Cholesterol concentration
STATE       Integer     A 1
                                 1 Iowa            2 Nebraska
AGECLASS    Integer     A 6
CONCLASS    Integer     A 6
```

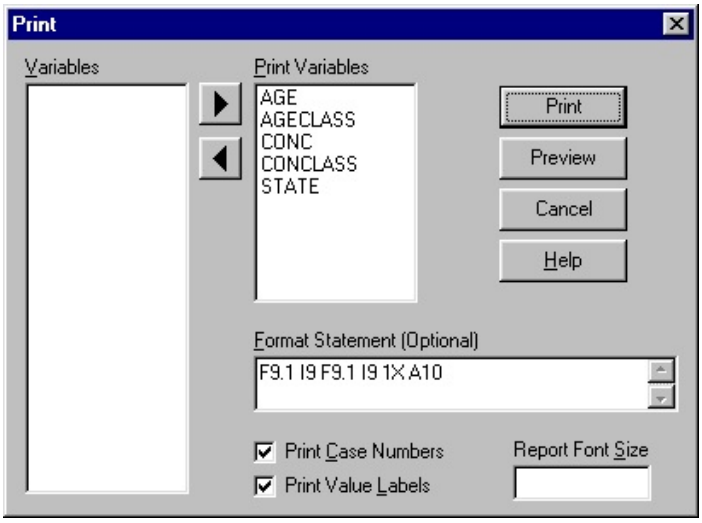
The report is largely self-explanatory. The column formats are coded: A-automatic, D-decimal, E-exponential, F-fixed, and I-integer. The column format letter is followed by the column width and sometimes the number of decimal places. See **Column Formats** in Chapter 2 for more information about column formats.

The File Info report can be printed or saved in a file just like other *Statistix* reports. Select Print from the File menu to print the report. Select Save As from the File menu to save the report in a file.

Print

The **Print** procedure is used to print the contents of the active window. In the case of *Statistix* reports and graphs, the report or graph is simply printed on the default printer. When the spreadsheet is the active window, the Print dialog box appears as shown on the next page.

First select the variables you want to print. Highlight one or more variables in the ***Variables*** list, then press the right-arrow button to move the highlighted variables to the ***Print Variables*** list. You can highlight all the variables by clicking on the first variable in the list, and then dragging your mouse to the last variable. The selected variables will be printed in the order that you've placed them in the Print Variables list.



You have the option of specifying a ***Format Statement*** that specifies the field widths and numeric formats for each variable. The format statement is discussed in detail below.

Check the ***Print Case Numbers*** box if you'd like to have case numbers printed on each line of the report.

If you've defined value labels for any of your variables, check the ***Print Value Labels*** box to have the labels printed rather than the numeric codes.

You can specify the font size by entering a number in the ***Report Font Size*** box. Typical values range from 10 to 12, but you can enter a smaller value to squeeze more data on a page.

When you finish making your selections, you can press the ***Print*** button to send the report directly to the printer. It's often a better idea to press the ***Preview*** button instead so you can look at the report on the screen to verify that the report is formatted correctly. While previewing the report, you have the option of printing the report, or saving it to a file.

Format Statement

A format statement is a list of format specifications used to indicate how you want the data for each variable to appear on the report. A variable format specification consists of a letter, a field width, and sometimes a number for decimal places as well. If you don't enter a format statement,

Statistix will construct a default format statement for you using the **Column Format** information discussed in Chapter 2. By omitting the format statement, the data are formatted more or less as it's displayed in the spreadsheet window. The default format guarantees at least one space between variables.

The different format specifications are listed in the table below. In the table, "w" is the field width, "s" is the number of significant digits, "d" is the number of digits to the right of the decimal point, and "r" is the repeat factor.

Name	General Format	Example Format	Example Appearance	Notes
Automatic	Aw	A10	John Smith 01/05/92 65 12.34567	
Decimal	Dw.s	D11.5	12.345 3.4523E-03	s <= w - 4
Exponential	Ew.s	E10.4	1.234E+01 3.452E-03	s <= w - 4
Fixed	Fw.d	F7.2	12.34 0.003	d < w - 2
Integer	Iw	I2	12 0	
Space	rX	10X		inserts spaces
New line	/	/		inserts line feed

The A format can be used for variables of any data type. The D, E, F, and I formats are used for integer and real variables, each resulting in a different numeric format.

The A format displays numbers using an integer format when the number is a whole number, or a decimal format when the number contains a fraction. For numbers with a fraction, as many digits will be displayed as possible but trimming any trailing zeros. The automatic format works well for data you've entered manually because the numbers generally are displayed just as you've entered them. For computed variables, such as variables created using the Transformations procedure, the A format often displays nonsignificant digits.

The D format displays numbers using a decimal format where the decimal is free to move about to maximize the number of digits displayed.

The F format displays numbers in a decimal format with a fixed number of decimal places.

The E format displays numbers in exponential format, or scientific notation. A number displayed in exponential format has two parts, the mantissa and the exponent. The mantissa is a number displayed as a decimal number that's always greater than or equal to 1.0 and less than 10.0. The exponent is displayed using the letter E followed by a signed integer. The number represented in this fashion is the mantissa multiplied by 10 raised to the exponent. For example, the number 4.23E-02 is equal to 4.23×10^{-2}, or 0.0423. This format is useful for very small and very large numbers.

The I format displays numbers as a whole number. When a number that includes a fraction is displayed using this format, the number is rounded off to the nearest whole number.

The X and / are not used as format specifications for variables but are used to insert spaces and line feeds into the output record.

Any of the format specifications can have a number in front called the repeat factor. This is used to abbreviate the format statement when several variables are to be formatted in a similar manner. A repeat factor can also be placed in front of a list of format specifications inside parentheses, as in:

```
3I5 2(I1 1X F4.2 F6.2) E10.4
```

When using format specifications, remember to make "w" large enough to account for the minus sign for negative numbers and for extra space between variables. String data are left justified, so you should always use the X format in front of a format for a string variable to insert a space.

The sample report on the next page shows the results for the format statement that appears in the example dialog box on page 96.

CASE	AGE	AGECLASS	CONC	CONCLASS	STATE
1	46.0	40	181.0	160	Iowa
2	52.0	50	228.0	200	Iowa
3	39.0	30	182.0	160	Iowa
4	65.0	60	249.0	240	Iowa
5	54.0	50	259.0	240	Iowa
6	33.0	30	201.0	200	Iowa
7	49.0	40	121.0	120	Iowa
8	76.0	70	339.0	320	Iowa
9	71.0	70	224.0	200	Iowa
10	41.0	40	112.0	80	Iowa
11	58.0	50	189.0	160	Iowa
12	18.0	10	137.0	120	Nebraska
13	44.0	40	173.0	160	Nebraska
14	33.0	30	177.0	160	Nebraska
15	78.0	70	241.0	240	Nebraska
16	51.0	50	225.0	200	Nebraska
17	43.0	40	223.0	200	Nebraska
18	44.0	40	190.0	160	Nebraska
19	58.0	50	257.0	240	Nebraska
20	63.0	60	337.0	320	Nebraska
21	19.0	10	189.0	160	Nebraska
22	42.0	40	214.0	200	Nebraska
23	30.0	30	140.0	120	Nebraska
24	47.0	40	196.0	160	Nebraska
25	58.0	50	262.0	240	Nebraska
26	70.0	70	261.0	240	Nebraska
27	67.0	60	356.0	320	Nebraska
28	31.0	30	159.0	120	Nebraska
29	21.0	20	191.0	160	Nebraska
30	56.0	50	197.0	160	Nebraska

Printer Setup

Selecting Printer Setup from the *Statistix* File menu gives you access to the *Windows* printer setup dialog box. You can use the procedure to select a printer or change the printer's properties.

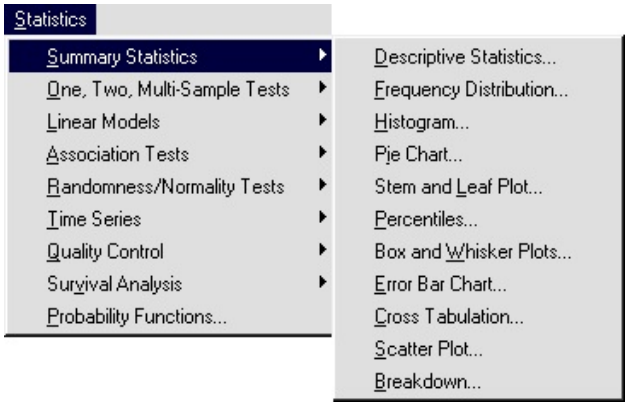
Exit

The Exit command is used to exit *Statistix*. If your open data file has been modified since you last saved it, *Statistix* warns you and gives you a chance to save it before exiting.

C H A P T E R

4

Summary and Descriptive Statistics



These procedures are designed to help you condense, summarize, and display data. You'll use them in the preliminary stages of analysis because they allow you to recognize general patterns and they suggest directions for further analysis. They're particularly useful for detecting "unusual" values.

The utility of these procedures isn't restricted to the preliminary stages of analysis, however. They're important tools for evaluating the results of a variety of analyses. For example, after fitting models to your data, you can use these procedures to inspect the resulting residuals.

The **Descriptive Statistics** procedure computes the mean, standard

deviation, confidence intervals, median, minimum, maximum, and other descriptive statistics for a list of variables. A grouping variable can be used to compute and display statistics broken down by group.

The **Frequency Distribution** procedure tabulates frequency tables, including counts and percentages, for discrete or continuous data.

A **Histogram** is a bar-chart used to graphically represent the frequencies of discrete data and the frequency density of continuous data.

A **Pie Chart** graphically represents frequencies, sums, or means using the slices of a pie.

A **Stem and Leaf Plot** is a frequency graph similar to a histogram where the digits of the data are used to construct the bars of the graph.

The **Percentiles** procedure is used to compute arbitrary percentiles for a list of variables.

A **Box and Whisker Plot** graphically presents the center and the spread of a variable. A grouping variable can be used to produce a box plot for each group.

The **Error Bar Chart** graphically represents the mean and standard deviation for a list of variables or groups of a variable.

A **Cross Tabulation** table displays the frequencies and percentages for each combination of variable values.

The **Scatter Plot** is used to graph a two dimensional scatter diagram. Up to five pairs of X and Y variables can be plotted on the same graph.

The **Breakdown** procedure computes the sum, mean, sample size, and standard deviation for a variable broken down in a nested fashion using the levels of up to five grouping variables.

The procedures are illustrated with example data from Snedecor and Cochran (1980, p. 386). The data are the blood serum cholesterol levels and ages of 30 women, 11 from Iowa and 19 from Nebraska. The cholesterol concentrations are in the variable CONC, and the ages are in AGE. The variable STATE indicates the state, with Iowa = 1 and Nebraska = 2. Two additional categorical variables were created using

Transformations. The variable AGECLASS assigns the ages to ten-year age classes. For example, if a woman's age were within the range 50 to 59, the value of AGECLASS would be 50. AGECLASS is created using the transformation

```
AGECLASS = 10 * TRUNC (AGE/10)
```

CONCLASS is created in a similar manner, and assigns each case to a 40 mg/100 ml cholesterol concentration class. It's created as follows:

```
CONCLASS = 40 * TRUNC (CONC/40)
```

For example, CONCLASS is assigned the value 200 for any case for which the value of CONC is in the 200 to 239 range.

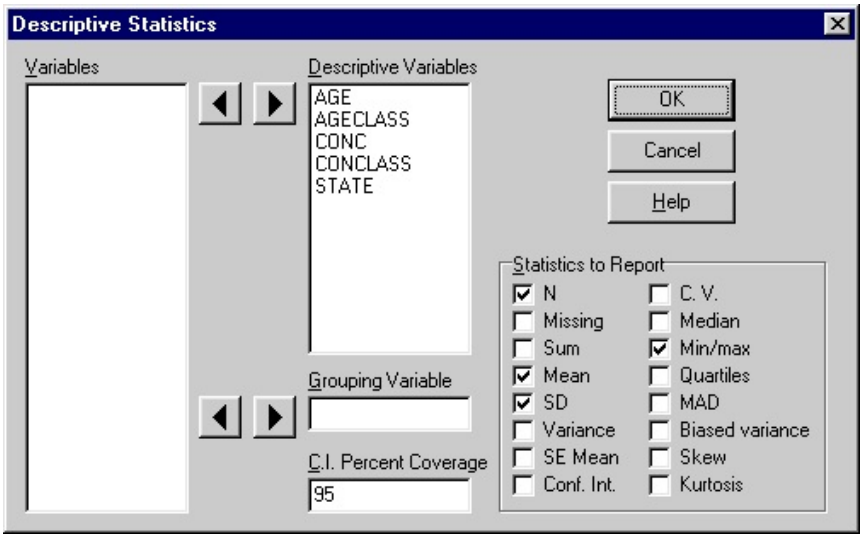
These example data are listed below. They're also distributed with the *Statistix* software in the data file Cholesterol.sx.

CASE	AGE	AGECLASS	CONC	CONCLASS	STATE
1	46	40	181	160	1
2	52	50	228	200	1
3	39	30	182	160	1
4	65	60	249	240	1
5	54	50	259	240	1
6	33	30	201	200	1
7	49	40	121	120	1
8	76	70	339	320	1
9	71	70	224	200	1
10	41	40	112	80	1
11	58	50	189	160	1
12	18	10	137	120	2
13	44	40	173	160	2
14	33	30	177	160	2
15	78	70	241	240	2
16	51	50	225	200	2
17	43	40	223	200	2
18	44	40	190	160	2
19	58	50	257	240	2
20	63	60	337	320	2
21	19	10	189	160	2
22	42	40	214	200	2
23	30	30	140	120	2
24	47	40	196	160	2
25	58	50	262	240	2
26	70	70	261	240	2
27	67	60	356	320	2
28	31	30	159	120	2
29	21	20	191	160	2
30	56	50	197	160	2

Descriptive Statistics

The **Descriptive Statistics** procedure produces a summary table of descriptive statistics for a list of variables. You can select the statistics you want tabulated from the following list: number of non-missing cases, number of missing cases, sum, mean, standard deviation, variance, standard error of the mean, confidence interval of the mean, coefficient of variation, median, minimum and maximum, first and third quartiles, median absolute deviation, biased variance, skew, and kurtosis.

Specification



Select the variables for which you want to compute descriptive statistics. Highlight the variables you want to select in the ***Variables*** list box, then press the right arrow button to move them to the ***Descriptive Variables*** list box. To highlight all variables, click on the first variable in the list, and drag the cursor to the last variable in the list.

If you select a ***Grouping Variable***, the summary statistics will be tabulated separately for each value found in the grouping variable. You can change the ***C. I. Percentage Coverage*** for mean confidence intervals. Select the statistics you want reported by checking off the ***Statistics to Report*** check boxes.

Data Restrictions

The grouping variable can be of any data type. Real values will be truncated to whole numbers. Strings will be truncated to ten characters.

Example

The data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. In the dialog box on the preceding page, all five variable names have been moved from the Variables list box to the Descriptive Variables list box. No grouping variable was specified. The results are:

Descriptive Statistics

Variable	N	Mean	SD	Minimum	Maximum
AGE	30	48.567	16.347	18.000	78.000
AGECLASS	30	44.000	16.316	10.000	70.000
CONC	30	213.67	59.751	112.00	356.00
CONCLASS	30	192.00	60.708	80.000	320.00
STATE	30	1.6333	0.4901	1.0000	2.0000

If you select more than five statistics, the table is presented with the variable names along the top. For example, the report below lists all of the statistics available.

Descriptive Statistics

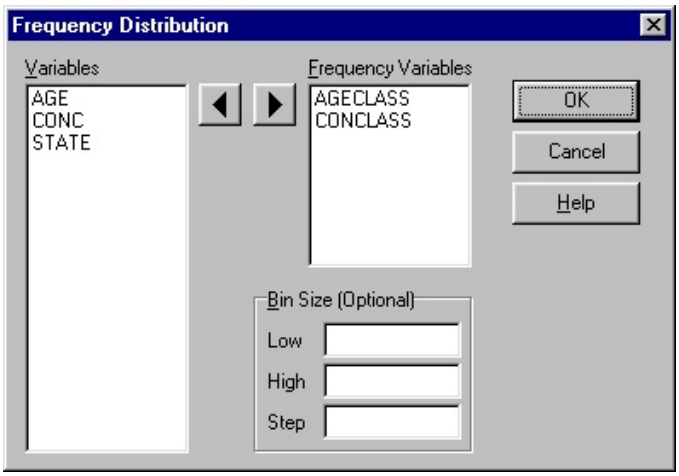
	AGE	AGECLASS	CONC	CONCLASS	STATE
N	30	30	30	30	30
Missing	0	0	0	0	0
Sum	1457	1320	6410	5760	49
Lo 95% CI	42.463	37.908	191.36	169.33	1.4503
Mean	48.567	44.000	213.67	192.00	1.6333
Up 95% CI	54.671	50.092	235.98	214.67	1.8164
SD	16.347	16.316	59.751	60.708	0.4901
Variance	267.22	266.21	3570.2	3685.5	0.2402
SE Mean	2.9845	2.9789	10.909	11.084	0.0895
C.V.	33.659	37.081	27.965	31.619	30.008
Minimum	18.000	10.000	112.00	80.000	1.0000
1st Quarti	37.500	30.000	180.00	160.00	1.0000
Median	48.000	40.000	199.00	180.00	2.0000
3rd Quarti	59.250	52.500	251.00	240.00	2.0000
Maximum	78.000	70.000	356.00	320.00	2.0000
MAD	10.000	10.000	27.500	20.000	0.0000
Biased Var	258.31	257.33	3451.2	3562.7	0.2322
Skew	-0.1009	-0.1822	0.6711	0.5851	-0.5534
Kurtosis	-0.7008	-0.3663	0.2762	-0.0783	-1.6938

The median absolute deviation (MAD) is the median value of the absolute differences among the individual values and the sample median. See Snedecor and Cochran (1980, pp. 78-81) for definitions of the other statistics.

Frequency Distribution

The **Frequency Distribution** procedure produces a frequency tabulation for discrete or continuous data. It computes the frequency, relative frequency (percentage of total), and cumulative and relative frequencies of data.

Specification



Select the variables for which you want to display frequency tables. Highlight the variables you want to select in the ***Variables*** list box, then press the right arrow button to move them to the ***Frequency Variables*** list box. To highlight all variables, click on the first variable in the list, and while holding the mouse button down, drag the cursor to the last variable in the list. You can also move a variable by double-clicking its name. To delete variables from the Frequency Variables list, highlight the variables you want to delete, then press the left arrow button.

You can specify ***Low***, ***High***, and ***Step*** values to control the number of bins and the width of each bin. The value you enter for low is the lowest value for the first bin. The value you enter for high is the highest value of the last bin. The value you enter for a step is the width of each bin. If you don't specify these values, frequencies are reported for each discrete value.

Data Restrictions

Variables of any data type can be specified. There can be no more than 500 unique values for each discrete variable and no more than 500 bins if low, high, and step values are specified.

Example

The data, from Snedecor and Cochran (1980, p. 386), are described at the beginning of this chapter. Frequencies are produced for the variable AGECLASS, which is the ten-year age class for each of the 30 female subjects, and for STATE, which indicates the state where the subject resides.

The dialog box on the preceding page illustrates what variables and options are selected. The results are:

Frequency Distribution of AGECLASS

Value	Freq	Percent	Cumulative Freq	Cumulative Percent
10	2	6.7	2	6.7
20	1	3.3	3	10.0
30	5	16.7	8	26.7
40	8	26.7	16	53.3
50	7	23.3	23	76.7
60	3	10.0	26	86.7
70	4	13.3	30	100.0
Total	30	100.0		

Frequency Distribution of STATE

Value	Freq	Percent	Cumulative Freq	Cumulative Percent
Iowa	11	36.7	11	36.7
Nebraska	19	63.3	30	100.0
Total	30	100.0		

The frequencies of each discrete value for a continuous variable, such as CONC in our example data, are not of interest. For a continuous variable, you need to establish intervals that span the range of the data and then count the number of times data values fall within the bounds of the intervals. You do this by entering low, high, and step values in the bottom of the dialog box. For example, you can enter the values 80, 360, and 40 for the low, high, and step values for CONC. The results are displayed below.

Frequency Distribution of CONC Cholesterol Concentration

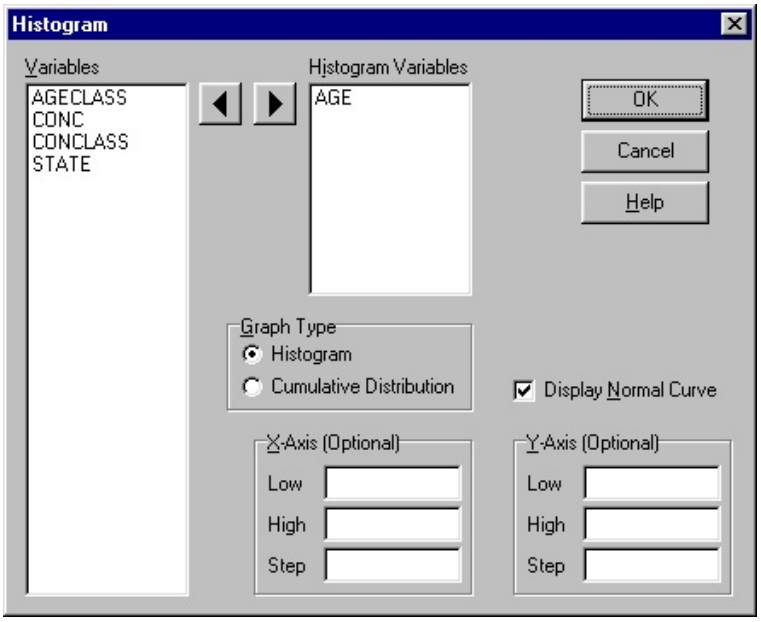
Low	High	Freq	Percent	Cumulative Freq	Cumulative Percent
80.0	120.0	1	3.3	1	3.3
120.0	160.0	4	13.3	5	16.7
160.0	200.0	10	33.3	15	50.0
200.0	240.0	6	20.0	21	70.0
240.0	280.0	6	20.0	27	90.0
280.0	320.0	0	0.0	27	90.0
320.0	360.0	3	10.0	30	100.0
Total		30	100.0		

If a value falls on an interval boundary, the value is counted in the lower of the two intervals.

Histogram

The **Histogram** procedure produces a bar graph frequency distribution for discrete or continuous variables. A histogram can summarize large amounts of data in a single visual image. You can have a normal curve superimposed over the histogram. This procedure can also produce a graph of the cumulative frequency distribution of a variable.

Specification



Highlight the variables you want to use to make a histogram in the ***Variables*** list and press the right arrow key to move them to the ***Histogram Variables*** list. You can only produce one histogram at a time. If you select more than one variable, the values of the variables will be combined to produce one plot.

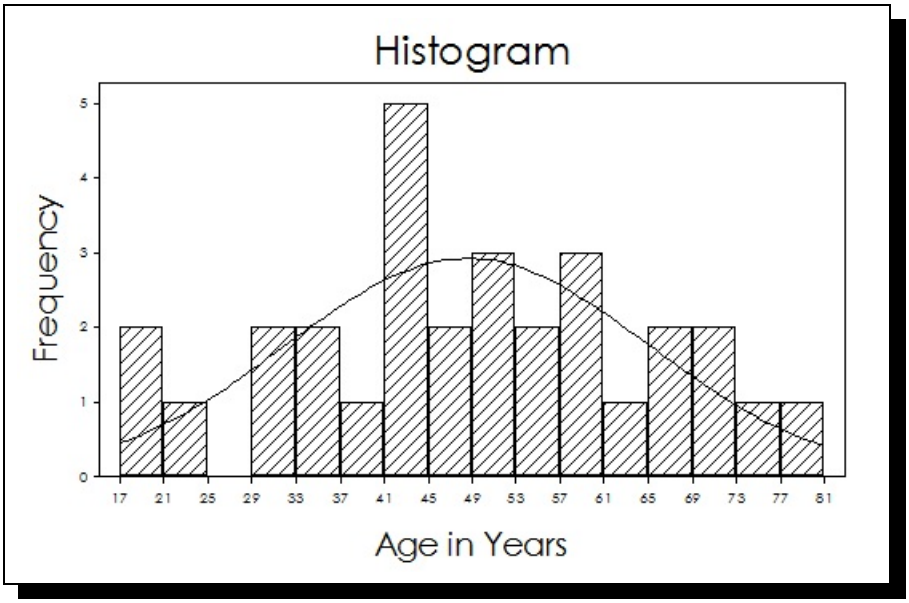
Select the graph type. Select ***Histogram*** to display the traditional histogram consisting of vertical bars. Select ***Cumulative Distribution*** to plot a curve representing cumulative percent.

Check the ***Display Normal Curve*** check box to superimpose a normal curve over the bars of the histogram.

You can enter ***Low***, ***High***, and ***Step*** values to control the X and Y axis scales. You can use this feature to create a meaningful interval width and interval boundaries. You can also use it to limit the range of data for the specified variable in order to eliminate outliers or to concentrate the plot on a particular range of values.

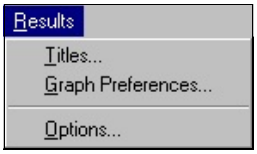
Example

The data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. The dialog box on the preceding page is used to graph a histogram with normal curve for the variable AGE. The results are:



Results Menu

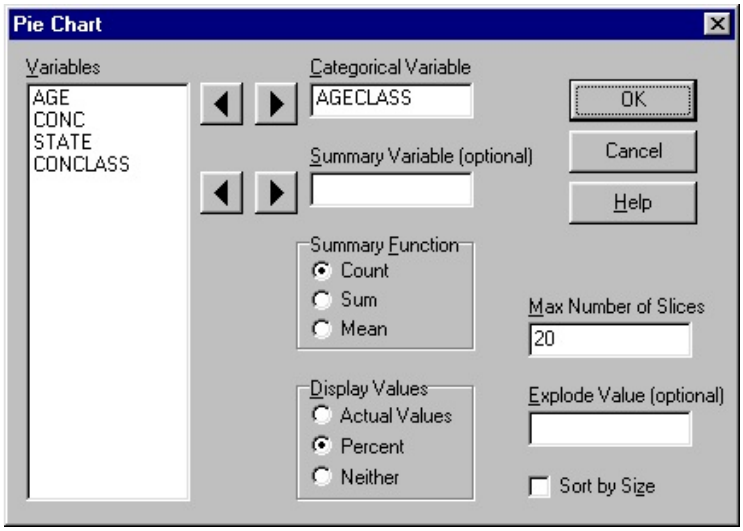
The results menu for the Histogram procedure includes a **Titles** procedure that lets you change the titles that appear on the plot, and a **Graph Preferences** procedure that lets you change fonts, colors, and fill patterns. Please see Chapter 1 for details.



Pie Chart

This procedure plots a pie chart. Pie charts are often used to graphically display a frequency distribution, but you can also apply the chart to sums or means.

Specification



First select a ***Categorical Variable***. The pie chart will have one pie slice for each value found for this variable.

If you want to base the size of pie slices on sums or means, then select the variable containing the data to use to compute the sums or means and move it to the ***Summary Variable*** box. Don't select a summary variable if you want pie slices to represent counts.

Pie slices will be labeled using the values found in the categorical variable. These labels can also include the summarized values (counts, sums, or means), or the percent of the total. Make your choice by selecting one of the ***Display Values*** radio buttons.

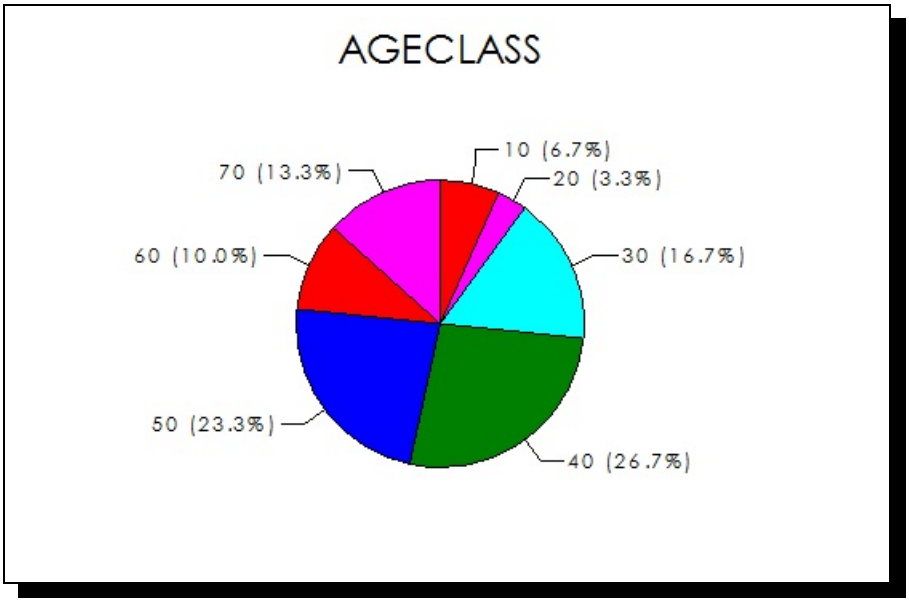
You can enter a value for the ***Max Number of Slices*** to limit the number of pie slices. If the categorical variable has more levels than the maximum number you indicate, the extra levels are grouped into one pie slice labeled "Other".

To emphasize a slice, you can explode it (pull it away from the rest of the circle). Enter the value for the categorical variable corresponding to the slice you want to emphasize in the ***Explode Value*** box.

The pie slices are normally ordered clockwise by the levels of the categorical variable. Check the ***Sort by Size*** check box to have the slices ordered by the size of the slices instead.

Example

The data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. The dialog box on the preceding page is used to graph a pie chart for the variable AGECLASS. The results are:



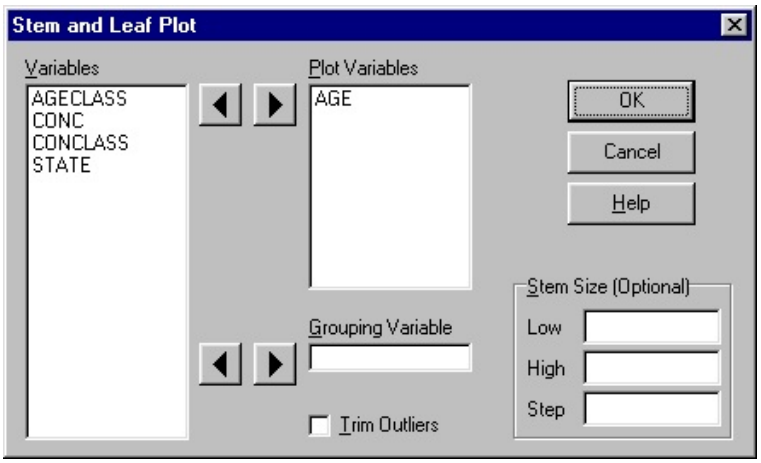
The example pie chart above is a simple frequency distribution based on counts. Pie charts are also useful for showing how sums are proportioned. Suppose you have a data set of transactions for a business that include the dollar amount of the transaction and the expense category for each transaction. By selecting the variable for expense category as your Categorical Variable, the variable for the dollar amount as your Summary Variable, and selecting Sum for the Summary Function, you could create a pie chart that shows how expenses are distributed among categories.

If the data you want to chart is already tabulated, create a data set with one case for each pie slice and select Sum for the Summary Function.

Stem and Leaf Plot

The **Stem and Leaf Plot** is a simple but handy way to organize numerical data. The digits of the individual values are ordered in a table of "stems" and "leaves" that resembles a histogram when turned sideways. Unlike a histogram, each original measurement can be read from the plot.

Specification



Select the variables you want to use to produce the stem and leaf plots from the ***Variables*** list box. If you select a ***Grouping Variable,*** a separate plot is produced for each value of the grouping variable.

Extreme values can affect the scale of the stem and leaf plot. If you see extreme values causing a scaling problem, check the ***Trim Outliers*** check box. The extreme values won't be omitted completely but will be listed individually outside the scale.

You can enter ***Low***, ***High***, and ***Step*** values to control the scale of the plot. This is useful when you want to compare two different plots using the same scale. Since stem boundaries must fall on whole digits, the step value must be 1, 2, 5, 10, or multiples of 0.1 or 10 (i.e., 0.1, 0.2, 0.5, 100, 200, etc.).

Example

The data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. The dialog box above specifies a plot for AGE, which is the age of 30 subjects. The results are given on the next page.

Stem And Leaf Plot of Age

```
   Leaf Digit Unit = 1                 Minimum  18.000
   1  8  Represents 18.                Median   48.000
                                       Maximum  78.000
          Stem  Leaves
       2    1  89
       3    2  1
       3    2
       7    3  0133
       8    3  9
      13    4  12344
     (3)    4  679
      14    5  124
      11    5  6888
       7    6  3
       6    6  57
       4    7  01
       2    7  68

30 Cases Included   0 Missing Cases
```

This plot contains all the information of a histogram. In addition, it preserves information about the fine structure of the data. Each number in your data is divided into two parts, the stem and the leaf. The stem indicates the values of the most significant digits of an observation, while the leaf gives the least significant digit. Each digit in the Leaves column is a separate leaf, so there is one leaf for each case.

For example, consider the first row of the plot “1 89". The stem value is 1, and the leaves are 8 and 9, so you know the digits for the first subject are 1 and 8 and the digits for the second subject are 1 and 9. You don’t know yet where to put the decimal point. That is, the numbers could be 1.8, 1.9, or perhaps 18, 19, or even 0.018, 0.019, etc. The message above the body of the plot “1 8 Represents 18.” is telling you that a stem value of 1 and a leaf value of 8 represents the number 18. So the first two values in our example are 18 and 19.

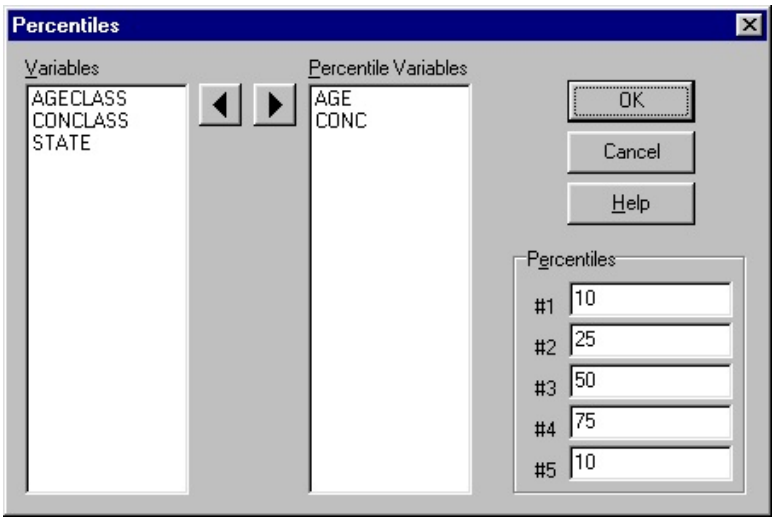
The first column in a stem and leaf plot is a cumulative frequency column that starts at both ends of the data and meets in the middle. The row that contains the median of the data is marked with parentheses around the count of observations for that row. For rows above the median, the number in the first column is the number of items in that row plus the number of items in all the rows above. Rows below the median are just the opposite. If the number of cases is even and the two middle values fall in different rows, there is no “median row”.

Further details of how to interpret stem and leaf plots can be found in Velleman and Hoaglin (1981).

Percentiles

The **Percentile** procedure computes the percentiles you specify for a list of variables. A percentile is a value such that a specified percent of the data falls at or below that value. The median is the 50th percentile. The lower and upper quartiles are the 25th and 75th percentiles.

Specification



Select the variables for which you want to compute percentiles from the ***Variables*** list and move them to the ***Percentiles Variables*** list. Enter one or more ***Percentile*** values you want computed in the space provided.

Example

The data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. The 10th, 25th, 50th (the median), 75th, and 90th percentiles are computed for the variables AGE and CONC, the age and cholesterol level for a sample of female subjects. The analysis is specified in the dialog box above. The results are given below.

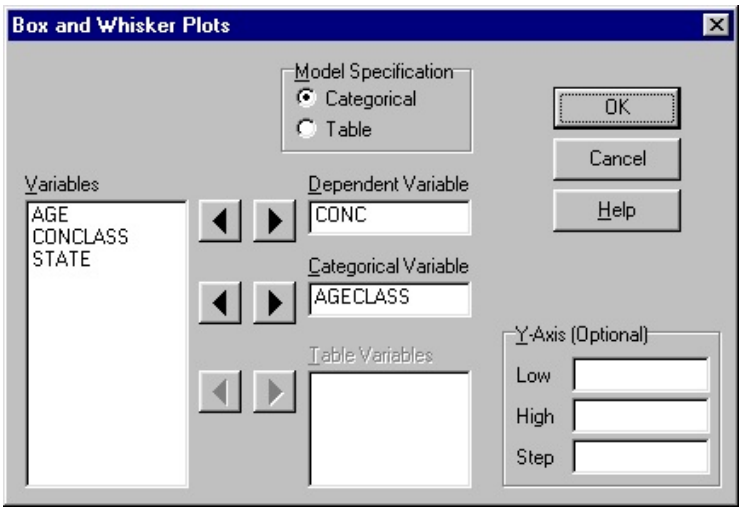
Percentiles

Variable	Cases	10.0	25.0	50.0	75.0	90.0
AGE	30	21.900	37.500	48.000	59.250	70.900
CONC	30	137.30	180.00	199.00	251.00	329.50

Box and Whisker Plot

The **Box and Whisker Plot** procedure computes box plots that graphically present measurements of central tendency and variability. A series of box plots can be displayed side by side, which can dramatically illustrate differences between groups.

Specification



First you select the method of specifying the analysis by selecting one of the ***Model Specification*** radio buttons. The method you choose depends on how you've organized the data you want to plot. Select the ***Categorical*** method if you want to plot the data of a single variable (the ***Dependent Variable***) using a second classifying variable that identifies groups (the ***Categorical Variable***). This will produce a series of box plots, one for each level of the classifying variable.

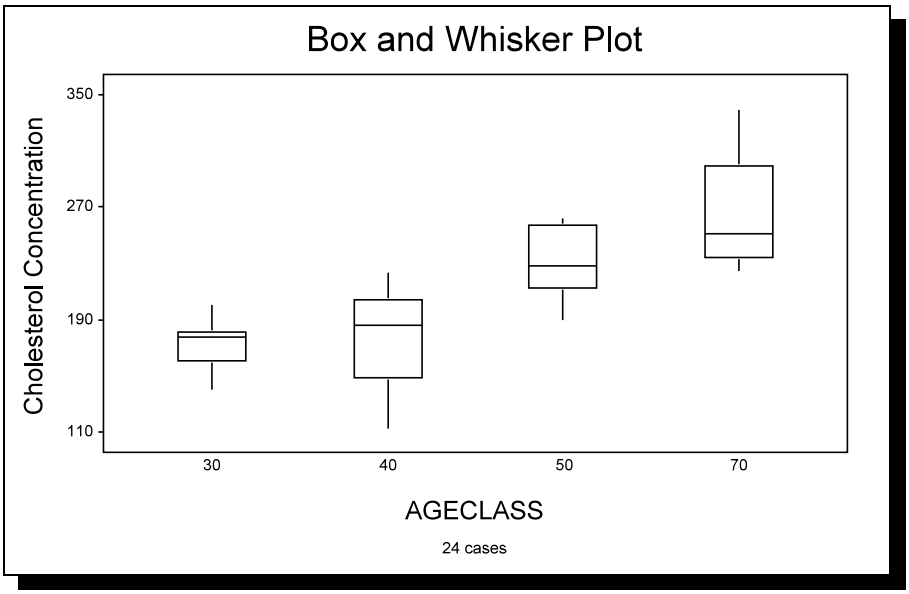
Select the ***Table*** method if you want to plot the data of several variables side by side. Move the names of the variables you want to plot from the ***Variables*** list to the ***Table Variables*** list.

Data Restrictions

Data values can't exceed 99,999. No more than 30 box plots can be displayed at once. The categorical variable can be of any type. Real values for the categorical variable will be truncated to whole numbers.

Example

The original data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. The dialog box on the preceding page specifies box plots for CONC (cholesterol concentration) grouped by AGECLASS. The resulting graph is displayed below.



These box plots powerfully illustrate that cholesterol concentration increases with age. Each box plot is composed of a box and two whiskers. The box encloses the middle half of the data. The box is bisected by a line at the value for the median. The vertical lines at the top and the bottom of the box are called the whiskers, and they indicate the range of "typical" data values. Whiskers always end at the value of an actual data point and can't be longer than 1½ times the size of the box.

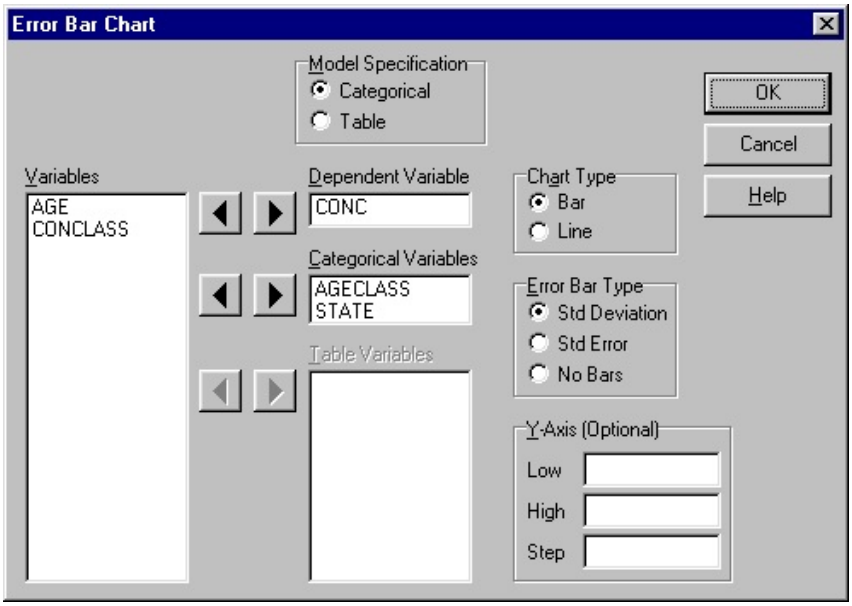
Extreme values are displayed as "∗" for possible outliers and "O" for probable outliers. Possible outliers are values that are outside the box boundaries by more than 1½ times the size of the box. Probable outliers are values that are outside the box boundaries by more than 3 times the size of the box.

More precise details of the concepts of middle half, typical values, and possible and probable outliers can be found in Velleman and Hoaglin (1981).

Error Bar Chart

The **Error Bar Chart** graphically displays the means and standard deviations (or standard errors) for a list of variables, or for one variable broken into groups by one or two grouping variables. The means are represented using vertical bars or circles, and the standard deviations are represented using a vertical line centered on the mean.

Specification



First you select the method of specifying the analysis, using either the ***Categorical*** method or the ***Table*** method. Select the ***Categorical*** method if you want to plot the data of a single variable (the ***Dependent Variable***) divided into groups by one or two classifying variables (the ***Categorical Variables***). Select a dependent variable and either one or two categorical variables. Using one categorical variable produces a series of bars, one for each level of the categorical variable. If you enter two categorical variables, the first is used to define the X axis and the second is used to further subdivide the data into sub-bars (see the example chart on the next page).

Select the ***Table*** method to plot the data of several variables side by side. Highlight the variables you want plotted in the ***Variables*** list and press the right-arrow button to move them to the ***Table Variables*** list.

Next select the ***Chart Type***, either bar chart or line chart. The bar chart uses vertical bars to represent the means. The line chart uses circles to mark the means, and the circles are connected sequentially with lines. The line chart is sometimes used when the grouping variable is ordered in a meaningful way, such as months of the year.

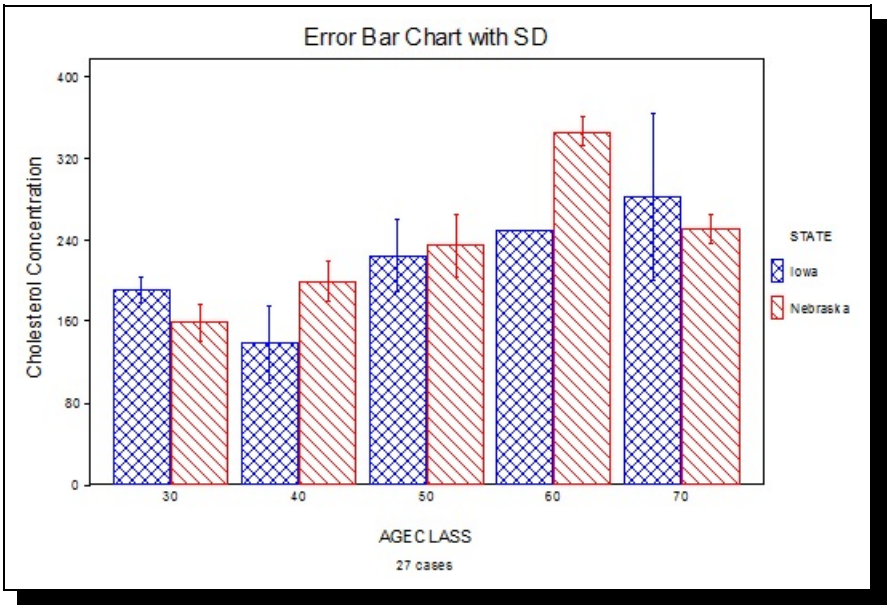
You have three choices for ***Error Bar Type***: the standard deviation, standard error of the mean, and no error bars.

Data Restrictions

When using the table method, you can select up to 20 variables. When using the categorical method, the first categorical variable can have up to 20 levels and the second categorical variable, if any, can have up to five levels.

Example

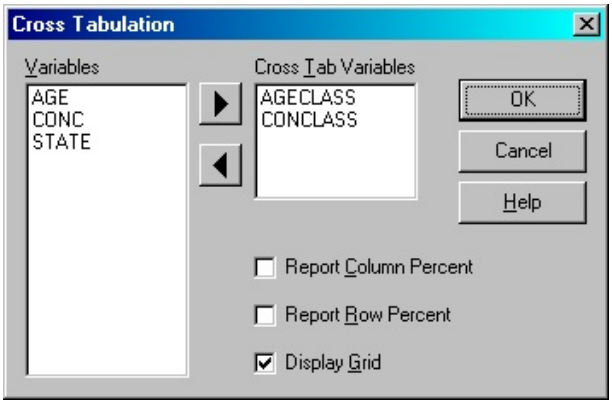
The original data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. CONC is the cholesterol concentration, AGECLASS indicates the age class, and STATE indicates in which of two states the subject lives. The three observations associated with age classes 10 and 20 were omitted for this analysis using the Omit/Select Cases procedure. The analysis is specified on the preceding page. The resulting chart is shown below.



Cross Tabulation

The **Cross Tabulation** procedure forms a cross tabulation table (also called a contingency table) for up to five classifying variables. The number of classifying variables determines the dimension of the table. There's a table cell for all unique combinations of values of the classifying variables. A cross tabulation table displays the number of cases that fall into each of the cross-classified table cells. In statistical terms, such a table represents the joint frequency distribution of the classifying variables.

Specification



To perform a cross tabulation, you simply select the classifying variables from the ***Variables*** list and press the right-arrow button to move them to the ***Cross Tab Variables*** list. The last variable becomes the column classifier and the second-to-last variable becomes the row classifier. If more than two variables are selected, the earlier variables become "control" variables. A separate table is produced for each unique combination of control variable values. These tables are produced in dictionary order; the levels of the rightmost control variables increment most rapidly.

A cross tabulation table always contains the counts for each cell. If you want to have column and row percentages reported for each cell as well, check the ***Report Column Percent*** and ***Report Row Percent*** check boxes.

Data Restrictions

There can be up to five classifying variables. Each classifying variable can have up to 500 levels. Classifying variables can have any data type (real, integer, date, and string). Numerical values of classifying variables must be whole numbers no larger than 99,999. Strings are truncated to ten

characters.

Example

The original data are from Snedecor and Cochran (1980, p. 386), described in the beginning of this chapter. AGECLASS indicates the age class a person was in (for example, AGECLASS = 60 means the person was in the 60 to 69 year age class). CONCLASS indicates the cholesterol concentration class a person was in (for example, CONCLASS = 160 means the person's cholesterol level was in the 160 to 199 mg/100 ml range).

The dialog box on the preceding page illustrates how to request cross tabulation of AGECLASS by CONCLASS. The results are displayed below:

Cross Tabulation of AGECLASS by CONCLASS

	CONCLASS						
AGECLASS	**80**	**120**	**160**	**200**	**240**	**320**	
10	0	1	1	0	0	0	2
20	0	0	1	0	0	0	1
30	0	2	2	1	0	0	5
40	1	1	4	2	0	0	8
50	0	0	2	2	3	0	7
60	0	0	0	0	1	2	3
70	0	0	0	1	2	1	4
	1	4	10	6	6	3	30

Cases Included 30 Missing Cases 0

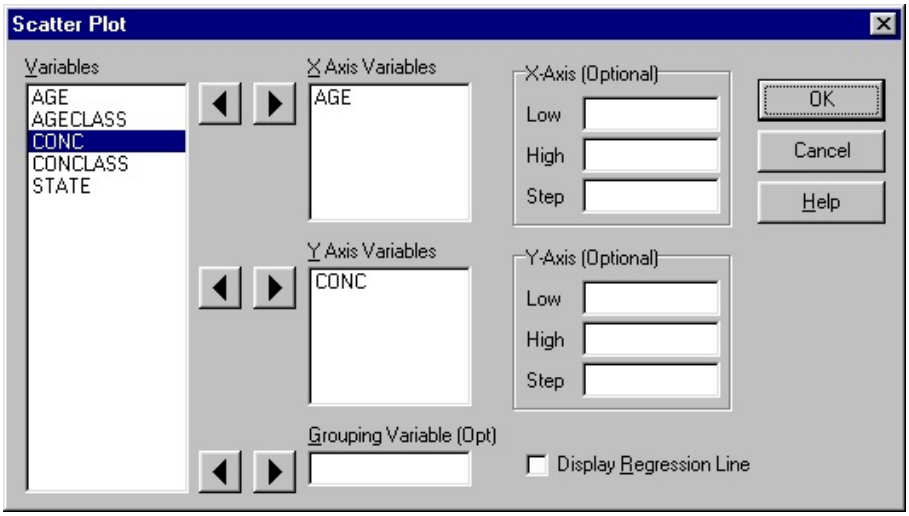
Note the diagonal pattern of nonzero cells in the table above. This suggests a relationship between age and cholesterol level.

Scatter Plot

The **Scatter Plot** procedure is used to produce a bivariate scatter diagram. Pairs of numbers are plotted as points on a X-Y graph. You can also have a fitted regression line displayed.

When investigating possible relationships between variables, plotting the data should be one of your first steps. Visual inspection of the data is invaluable and often reveals features of the data that would be overlooked if you proceeded directly with your statistical analyses.

Specification



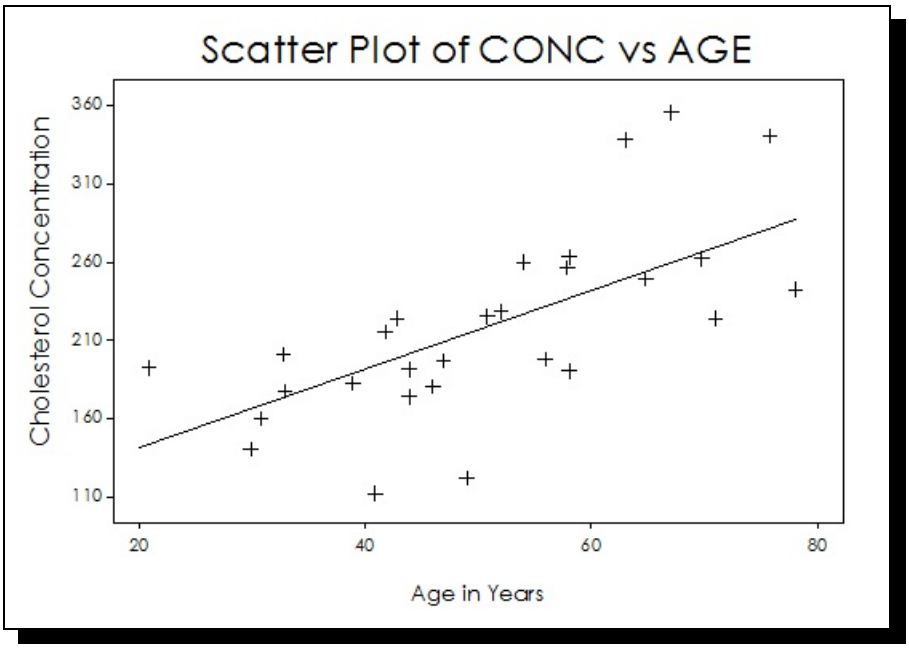
Select an ***X Axis Variable*** and a ***Y Axis Variable***. The arrow buttons to the left of each list are used to select and deselect variables for the respective lists. You can select additional variable name pairs. All pairs are displayed on the same graph using different symbols.

You can enter low, high, and step values to control the scales of either the ***X Axis*** or the ***Y Axis***. If you enter low and high values for either the X or the Y axis, only points that fall between these values will be plotted. This option is useful for eliminating outliers from the plot or zooming in on a particular portion of the plot.

Check the ***Display Regression Line*** box to have a fitted regression drawn through the points on the scatter plot.

Example

The data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. The dialog box on the preceding page is used to request a scatter plot for CONC vs. AGE, the blood cholesterol level and age of 30 female subjects. The results are displayed below.



The fitted linear regression line in the graph above makes it easier to see the linear relationship between age and cholesterol concentration. A fitted line can also be useful as a reference line to spot nonlinear relationships between two variables.

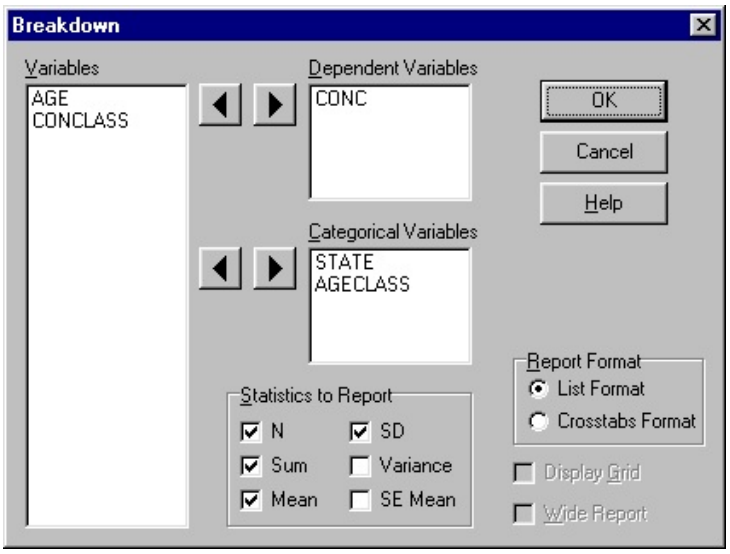
Results Menu

The results menu for the Scatter Plot procedure includes a Titles procedure that lets you change the titles that appear on the plot, and a Graph Preferences procedure that lets you change plot symbols and colors. Please see Chapter 1 for details.

Breakdown

Breakdown computes summary statistics for a variable broken into groups and subgroups in a nested fashion. The dependent variable can be classified by up to five categorical variables. The summary statistics are displayed for all levels of nesting. You can select the statistics to be reported from the following list: number of cases, sum, mean, standard deviation, variance, and standard error of the mean.

Specification



Select the variable you want to compute the summary statistics for and move it to the ***Dependent Variable*** box. Then select up to five ***Categorical Variables*** that will be used to "break down" the data into groups. The order in which the categorical variables are selected determines the order of nesting, with the value of the last variable changing most rapidly.

Check the ***Statistics to Report*** boxes for the statistics you want to include in the report. Select a report format. An example of the ***List Format*** is displayed on the next page. The ***Crosstabs Format*** presents the same information in two-dimensional tables.

Data Restrictions

There can be up to five categorical variables. The categorical variables can be of any data type. Real values will be truncated to whole numbers and

must be no larger than 99,999. Strings will be truncated to ten characters.

Example

The original data are from Snedecor and Cochran (1980, p. 386), described at the beginning of this chapter. The dependent variable CONC is the cholesterol concentration of 30 female subjects. The two categorical variables used are STATE and AGECLASS. For example, suppose you're interested in the cholesterol concentration means by states as well as the age-specific means within states. The variables selected in the dialog box on the preceding page specify that AGECLASS be nested within STATE. The results are presented below.

Breakdown for CONC Cholesterol Concentration

Variable	Level	N	Sum	Mean	SD
AGECLASS	30	2	383	191.50	13.435
AGECLASS	40	3	414	138.00	37.510
AGECLASS	50	3	676	225.33	35.076
AGECLASS	60	1	249	249.00	
AGECLASS	70	2	563	281.50	81.317
STATE	Iowa	11	2285	207.73	63.795
AGECLASS	10	2	326	163.00	36.770
AGECLASS	20	1	191	191.00	
AGECLASS	30	3	476	158.67	18.502
AGECLASS	40	5	996	199.20	19.791
AGECLASS	50	4	941	235.25	30.314
AGECLASS	60	2	693	346.50	13.435
AGECLASS	70	2	502	251.00	14.142
STATE	Nebraska	19	4125	217.11	58.796
Overall		30	6410	213.67	59.751

Cases Included 30 Missing Cases 0

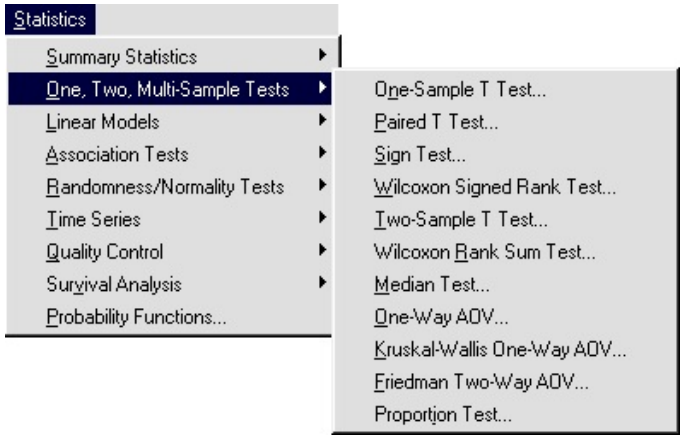
The indentations of the first two columns (the variable names and their values) depict the nesting structure. Any variable X indented with respect to another variable Z means the statistics for the levels of X are nested within the levels of Z. For example, AGECLASS is nested within STATE. The order of nesting is consistent with the order in which the classifying variables are listed in the ***Categorical Variables*** list box.

Note that the outer levels of nesting summarize the inner levels. In the example above, the line labeled "Iowa" summarizes the data for the five age classes listed above it.

C H A P T E R

5

One, Two, & Multi-Sample Tests



Statistix offers a number of procedures to test hypotheses about the central values of the population distributions from which the samples are drawn. These procedures are often referred to as tests of location. Several of these tests are parametric and require the assumption that the data are normally distributed. Nonparametric tests are provided for situations where the assumption of normality is not appropriate. When their assumptions are appropriate, parametric tests are generally more powerful than their nonparametric equivalents, although nonparametric tests often compare quite well in performance. The parametric versions test hypotheses concerning the group means. The nonparametric procedures test central value hypotheses based on measures other than the mean.

The **One-Sample T Test** is used to test hypotheses about sample means.

The **Paired T Test** is a parametric test used to test for differences between means of two groups when the samples are made in pairs.

The **Sign Test** and **Wilcoxon Signed Rank Test** are nonparametric alternatives to the Paired T Test.

The **Two-Sample T Test** is a parametric test that tests for a difference in the means of two groups when the samples are drawn independently from two normally distributed populations.

The **Wilcoxon Rank Sum Test** and **Median Test** are nonparametric alternatives to the Two-sample T Test.

The **One-Way AOV** is a multi-sample test that tests for differences among the means of several groups.

The **Kruskal-Wallis One-Way AOV** is a nonparametric alternative to the One-Way AOV.

The **Friedman Two-Way AOV** is a nonparametric alternative to the two-way analysis of variance. The **General AOV/AOCV** procedure, which is discussed in Chapter 6, performs parametric tests with two or more classifying attributes.

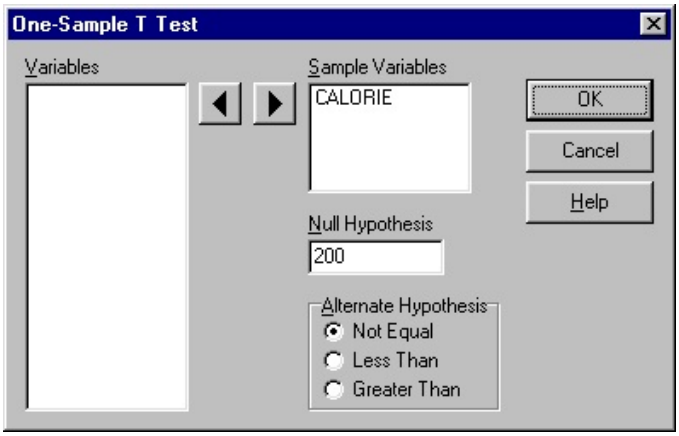
The **Proportion Test** is used to perform one- and two-sample hypothesis tests and compute confidence intervals for proportions.

Background on the parametric tests can be found in Snedecor and Cochran (1980). Hollander and Wolfe (1973), Lehmann (1975), and Siegel and Castellan (1988) are good references for the nonparametric procedures.

One-Sample T Test

The **One-Sample T Test** is used to test whether the mean of a sample drawn from a normal population differs from a hypothesized value.

Specification



Select the variable from the ***Variables*** list that contains the sample values. Double-clicking a variable will move it to the ***Sample Variable*** box. Enter a value for the null hypothesis. Select the two-sided alternative hypothesis "not equal", or a one-sided alternative "less than" or "greater than".

Example

Ten frozen dinners labeled "200 calories" were randomly selected from a day's production at a factory. The caloric content of the dinners were measured at 198, 203, 223, 196, 202, 189, 208, 215, 218, 207. The dialog box above illustrates how to test the hypothesis that the average number of calories in a dinner is 200. The results are:

One-Sample T Test

Null Hypothesis: Mu = 200
Alternative Hyp: Mu <> 200

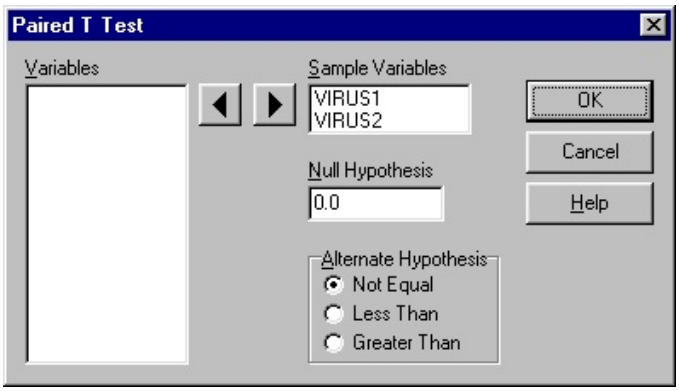
			95% Conf Interval				
Variable	**Mean**	**SE**	**Lower**	**Upper**	**T**	**DF**	**P**
CALORIES	205.90	3.3282	198.37	213.43	1.77	9	0.1100

The report includes the mean, standard error, confidence interval, t-test, and the p-value. Since the p-value of 0.1100 is larger than the typical value of 0.05 for the rejection level, the null hypothesis is not rejected in this example.

Paired T Test

The **Paired T Test**, a parametric procedure, is useful for testing whether the means of two groups are different, where the samples were drawn in pairs. The test is actually testing whether the mean of the differences of the pairs is different from zero, or from some other hypothesized value.

Specification



Select the two variables that contain the paired samples. Highlight the variables you want to select in the ***Variables*** list, then press the right-arrow button to move them to the ***Sample Variables*** list box. The typical ***Null Hypothesis*** is that the difference is zero, but you can enter a different value. You can also select an ***Alternative Hypothesis***: "not equal", "less than", or "greater than".

Example

The data for this example (Snedecor and Cochran, 1980, p. 87) concern the number of lesions produced on a tobacco leaf by the application of two different viral preparations. The halves of a leaf constitute a pair. The data for the first preparation are in variable VIRUS1, and that for the second preparation are in VIRUS2 (see Sample Data\tobacco.sx).

CASE	VIRUS1	VIRUS2
1	31	18
2	20	17
3	18	14
4	17	11
5	9	10
6	8	7
7	10	5
8	7	6

The analysis is specified on the preceding page. The results are presented below.

```
Paired T Test for VIRUS1 - VIRUS2

Null Hypothesis: difference =  0
Alternative Hyp: difference <> 0

Mean            4.0000
Std Error       1.5236
Mean - H0       4.0000
Lower 95% CI    0.3972
Upper 95% CI    7.6028
T                 2.63
DF                   7
P               0.0341

Cases Included 8    Missing Cases 0
```

The null hypothesis being examined is that the mean of the differences is zero. If the assumption of normality is appropriate, the small p-value (0.0341) suggests that the mean of the differences is not zero, i.e., the two different viral preparations do cause lesions at different rates. (The **Shapiro-Wilk Test** and **Normal Probability Plot** can be used to examine the assumption of normality.) The p-value is for a two-tailed test; halving it produces a one-tailed p-value.

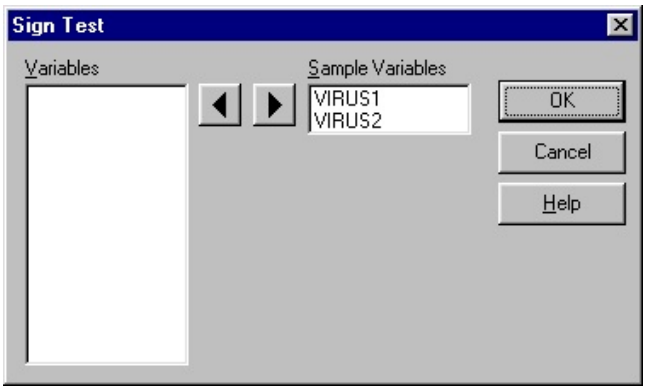
It's not appropriate to use this test if the data are not paired. We'll call the unit from which the two members of the pair were drawn a block. For example, the blocks may be individuals and the two members of the data pair are reaction times before and after ingestion of some test medication. The advantage of a paired test is that it removes variation in the data due to blocks; the data used for the test are the pair differences within the blocks. The "noise" in the data due to the fact that some individuals have naturally faster or slower reaction times regardless of the medication would thus be eliminated. (The paired t test is a special case of a randomized block design analysis of variance.) This analysis is not very efficient if the pair members are not correlated within blocks; in this case a **Two-Sample T Test** should be considered instead. Snedecor and Cochran (1980, p. 99-102) give further detail on paired versus independent sampling.

Sign Test

The **Sign Test** is a nonparametric alternative to the **Paired T Test**. It requires virtually no assumptions about the paired samples other than that they are random and independent. On the negative side, it's not as powerful as the **Paired T Test** or **Wilcoxon Signed Rank Test**. However, it's especially useful for situations where quantitative measures are difficult to obtain but where a member of the pair can be judged "greater than" or "less than" the other member of the pair.

As with other paired t tests, it assumes that you have two groups and that you have drawn your samples in pairs. The only information in the data which the sign test uses is whether, within a pair, the item from the first group was greater than ("+") or less than ("-") the item in the second group. If there is no consistent difference between the groups, there should be an equal number of "+"s and "-"s in the data except for random variation.

Specification



Select the two variables that contain the paired samples. Highlight the variables in the ***Variables*** list, then press the right-arrow button to move them to the ***Sample Variables*** list box. You can also move a variable by double-clicking on the variable name.

Example

The data for this example (Snedecor and Cochran, 1980, p. 87) concern the number of lesions produced on a tobacco leaf by the application of two different viral preparations. The halves of a leaf constitute a pair. The data for the first preparation are in variable VIRUS1, and that for the second preparation are in VIRUS2 (see Sample Data\tobacco.sx).

CASE	VIRUS1	VIRUS2
1	31	18
2	20	17
3	18	14
4	17	11
5	9	10
6	8	7
7	10	5
8	7	6

The analysis is specified on the preceding page. The results are displayed below.

```
Sign Test for VIRUS1 - VIRUS2

Number of Negative Differences              1
Number of Positive Differences              7
Number of Zero Differences (ignored)        0

Probability of a result as or more
extreme than observed (one-tailed p-value)  0.0352

A value is counted as a zero if its
absolute value is less than 0.00001

Cases Included 8    Missing Cases 0
```

The null hypothesis tested by the sign test is that the median of the differences is zero. The calculated probability is the binomial probability of observing as few or fewer of the less abundant sign, given that an individual difference is equally likely to be of either sign.

For the virus example, the calculated probability is the probability of observing one or fewer negative differences in a random sample of eight. This is a one-tailed probability; doubling it produces the correct two-tailed value. So the two-tailed p-value for the example is 0.0704, somewhat larger than the p-value observed with the **Paired T Test**.

Computational Notes

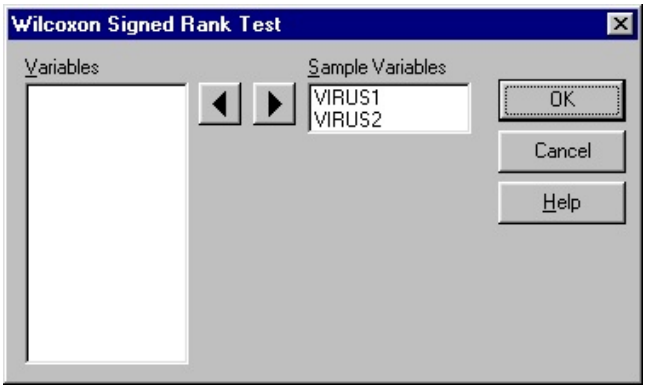
The probability is calculated using the same routine as in the binomial function in **Probability Functions**. The parameter P is set to 0.5.

Wilcoxon Signed Rank Test

The **Wilcoxon Signed Rank Test** is a nonparametric alternative to the **Paired T Test**. It's generally more powerful than the **Sign Test**. As with other paired tests, it assumes that you have two groups and that you have drawn your sample in pairs. Each pair contains an item from the first group and an item from the second group. This procedure tests the hypothesis that the frequency distributions for the two groups are identical. Exact p-values are computed for small sample sizes.

Specification

Select the two variables containing the paired samples. Highlight the variables in the ***Variables*** list, then press the right-arrow button to move them to the ***Sample Variables*** list box.



Example

The data for this example (Snedecor and Cochran, 1980, p. 87) concern the number of lesions produced on a tobacco leaf by the application of two different viral preparations. The halves of a leaf constitute a pair. The data for the first preparation are in variable VIRUS1, and that for the second preparation are in VIRUS2. See **Paired T Test** on page 128 for the data listing, or open the data file Sample Data\tobacco.sx.

The differences are first ranked by absolute value. Tied values are given a mean rank (Hollander and Wolfe 1973). Differences are considered to be tied if they are within 0.00001 of one another. The ranks are given the same signs that the original differences had. The negative and positive signed ranks are then summed separately.

Wilcoxon Signed Rank Test for VIRUS1 - VIRUS2

Sum of Negative Ranks	-2.0000
Sum of Positive Ranks	34.000
Exact probability of a result as or more extreme than the observed ranks (one-tailed p-value)	0.0117
Normal Approximation with Continuity Correction	2.170
Two-tailed P-value for Normal Approximation	0.0300
Total number of values that were tied	3
Number of zero differences dropped	0
Max. diff. allowed between ties	0.00001

Cases Included 8 Missing Cases 0

Suppose the frequency distributions for groups one and two were the same. The frequency distribution of the differences of the pairs would then be symmetrical and have a median of zero. In this instance, the absolute values of the sums of negative and positive signed ranks would be expected to be "similar". The signed rank test tests the null hypothesis that the median of the differences equals zero.

The exact p-values for the Wilcoxon signed rank test are computed for small to moderate sample sizes (20 or fewer cases). The exact one-tailed p-value is computed; doubling this yields the exact two-tailed p-value. When ties are found to be present, the "exact probability" is no longer exact but will usually be a good approximation. When sample sizes are moderate to large, the normal approximation statistic gives reliable results. The p-value for the normal approximation is two-tailed. The normal approximation includes a correction for continuity; its use is described in Snedecor and Cochran (1980, p. 142).

In the example, the exact p-value is 0.0117, which when doubled gives the two-tailed value of 0.0234. This is fairly close to the p-value of 0.0300 using the normal approximation. As with the t test, these results suggest that the preparations do produce lesions at different rates. While the paired t test is a more powerful test than the signed rank test, the difference in power is often not great. The signed rank test is a popular alternative because it requires much less restrictive assumptions about the data.

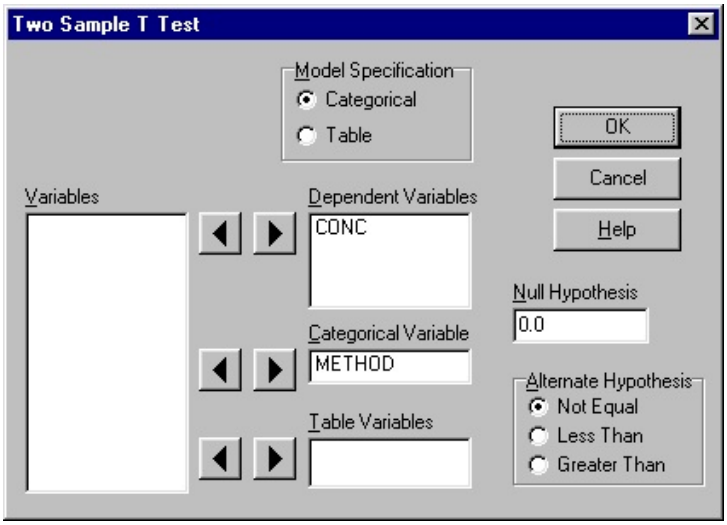
The exact p-value routine is based on the p-value routine for the **Wilcoxon Rank Sum Test**. It exploits the fact that the null distribution of the signed rank statistic can be factored as a product of a binomial distribution and the null distribution of the rank sum statistic (Bickel and Doksum 1977).

Two-Sample T Test

This procedure computes two-sample t tests, which test for differences between the means of two independent samples. It's applicable in situations where samples are drawn independently from two normally distributed groups. Two t tests are computed; one assumes equal group variances, and the other assumes different group variances. A test for equality of variances is also performed.

Specification

The analysis can be specified in one of two ways, depending on how the data are stored. If the data from both groups are entered into a single variable, and a second categorical variable is used to identify the two groups, use the ***Categorical*** method, as illustrated below. Move the variable containing the observed data into the ***Dependent Variable*** box. Move the variable that identifies the two groups into the ***Categorical Variable*** box. To move a variable, highlight the variable in the ***Variables*** box, then press the right-arrow button next to the box to which you want to move it. When using the categorical method, you can specify more than one dependent variable, in which case a separate report is displayed for each variable.



If the two groups are entered into *Statistix* as two variables, one for each group, use the ***Table*** method as illustrated on page 137. Select the two variables and move them to the ***Table Variables*** box.

Typically the null hypothesis is that the means are equal, or that the difference is zero. You can enter a different value for the null hypothesis. You can also select the alternative hypothesis: the two-sided alternative "not equal", or a one-sided alternative "less than" or "greater than".

Data Restrictions

The grouping variable used with the categorical method can be of any data type (i.e., real, integer, date, or string). Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters.

Example

The data for this example come from Snedecor and Cochran (1980). The goal is to compare the results of a standard, but slow, chemical analysis procedure with a quicker, but potentially less precise, procedure. The variable CONC is used to store the chemical concentrations determined by both methods. The variable METHOD is used to identify the method used (1 = standard, 2 = quick) to determine the concentration.

CASE	CONC	METHOD
1	25	1
2	24	1
3	25	1
4	26	1
5	23	2
6	18	2
7	22	2
8	28	2
9	17	2
10	25	2
11	19	2
12	16	2

These data are available in the file Sample Data\concentrations.sx. The analysis is specified on the preceding page. The results are displayed on the next page.

Summary statistics for the two groups are given first, including the group means, sample sizes, standard deviations, and standard errors. The t-statistics and associated information are given next. The t test labeled "Equal Variances" is testing the null hypothesis that means for the two groups are equal given that the two groups have the same variances. The t test labeled "Unequal Variances" tests the same null hypothesis except that it does not require the assumption that the variances of the two groups are equal. A discussion of such tests is given in Snedecor and Cochran (1980, p. 96-98). Note that the degrees of freedom for unequal variances are expressed as a decimal number. It's computed using Satterthwaite's

approximation, described in Snedecor and Cochran. An F test for the equality of the group variances is given after the t tests.

Two-Sample T Tests for CONC by METHOD

METHOD	Mean	N	SD	SE
Standard	25.000	4	0.8165	0.4082
Quick	21.000	8	4.2088	1.4880
Difference	4.0000			

Null Hypothesis: difference = 0
Alternative Hyp: difference <> 0

				95% CI for Difference	
Assumption	**T**	**DF**	**P**	**Lower**	**Upper**
Equal Variances	1.84	10	0.0956	-0.8433	8.8433
Unequal Variances	2.59	8.0	0.0320	0.4408	7.5592

Test for Equality of Variances	F	DF	P
	26.57	7,3	0.0106

Cases Included 12 Missing Cases 0

Snedecor and Cochran use the above example to illustrate how unequal variances can influence the analysis. Evidence for a difference between two chemical analyses is considerably weaker when equal variances are assumed (p=0.0956) than when unequal variances are assumed (p=0.0320). When in doubt, it's safer to assume the variances are unequal. In our example, the F test for equality of variances lends strong support for assuming the variances are unequal (p=0.0106).

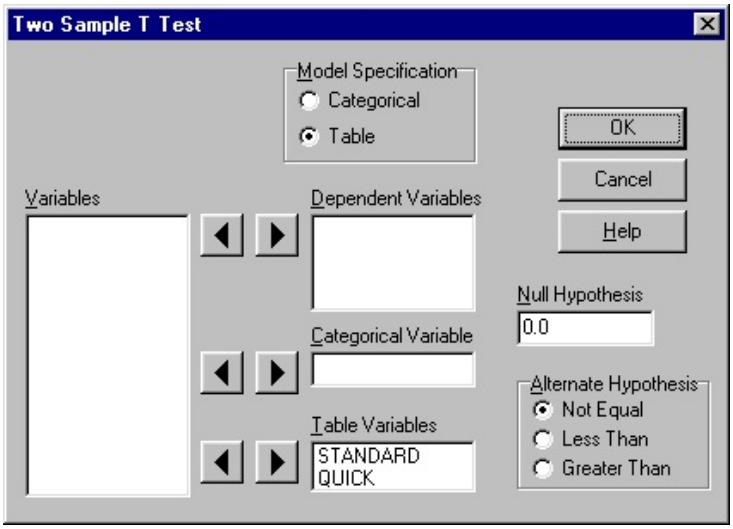
These t tests, as well as the F-test for equality of variances, are based on the assumption that the data are normally distributed. The **Shapiro-Wilk Test** and **Normal Probability Plot** are useful for examining this assumption. You should consider the **Median Test** or the **Wilcoxon Rank Sum Test** if non-normality is a problem.

To illustrate the ***Table*** method of model specification, suppose that the data from the two methods were entered as two separate variables.

CASE	STANDARD	QUICK
1	25	23
2	24	18
3	25	22
4	26	28
5	M	17
6	M	25
7	M	19
8	M	16

To specify the analysis, first select ***Table*** from the Model Specification radio buttons. Then move the two variable names STANDARD and

QUICK from the ***Variables*** list to the ***Table Variables*** list box. The dialog box below illustrates this method.



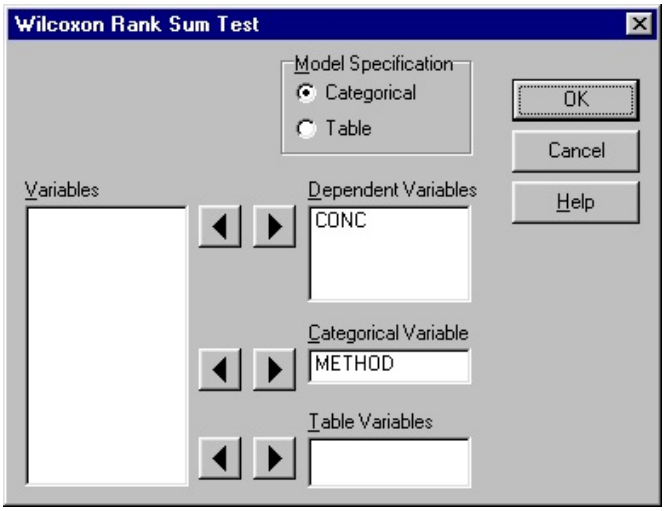
Wilcoxon Rank Sum Test

Statistix computes the **Wilcoxon Rank Sum Test**, a nonparametric procedure that tests for differences in the central values of samples from two independent samples. This test can be performed with either of two statistics—the Wilcoxon rank sum statistic or the Mann-Whitney U statistic. *Statistix* computes both statistics. Both of these statistics are mathematically equivalent and always lead to identical results. Exact p-values are given for small sample sizes. This test is often almost as powerful as the **Two-Sample T Test**, and is usually more powerful than the **Median Test**.

Specification

The analysis can be specified in one of two ways, depending on how the data are stored. If the two groups are entered into *Statistix* as two variables, use the ***Table*** method and move the two variables to the ***Table Variables*** list box.

If the data from both groups are entered into a single variable and a second categorical variable is used to identify the two groups, use the ***Categorical*** method as illustrated in the dialog box below. Move the variable containing the observed data into the ***Dependent Variable*** box, and the variable that identifies the two groups into the ***Categorical Variable*** box.



Example

The data for this example come from Snedecor and Cochran (1980). The variable CONC is used to store the chemical concentrations determined by two methods. The variable METHOD is used to identify the method used (1 = standard, 2 = quick) to determine the concentration.

CASE	CONC	METHOD
1	25	1
2	24	1
3	25	1
4	26	1
5	23	2
6	18	2
7	22	2
8	28	2
9	17	2
10	25	2
11	19	2
12	16	2

The analysis is specified on the preceding page. The results are shown below.

Wilcoxon Rank Sum Test for CONC by METHOD

METHOD	Rank Sum	N	U Stat	Mean Rank
Standard	36.000	4	26.000	9.0
Quick	42.000	8	6.0000	5.3
Total	78.000	12		

Exact Permutation Test Two-tailed P-value 0.2869

Normal Approximation with Corrections for Continuity and Ties 1.625
Two-tailed P-value for Normal Approximation 0.1042

Total number of values that were tied 3
Maximum difference allowed between ties 0.00001

Cases Included 12 Missing Cases 0

All the data are combined and converted to ranks. Tied scores are assigned mean ranks (Hollander and Wolfe 1973). Values are considered to be tied if they are within 0.00001 of one another. The ranks for each group are then summed to get the rank sum statistic for each group. If the distributions for the two groups are the same, the average ranks should be "similar" for each group. The null hypothesis being tested by the rank sum test is that the distributions for the two groups are the same. Rejecting this hypothesis usually leads to the conclusion that the central values for the two groups differ, although strictly you can only conclude that the distributions for the two groups differ in some way (Bradley 1968).

The Mann-Whitney U statistic corresponding to the rank sum is also given. When sample sizes are small to moderate, exact p-values are calculated and displayed. (Exact p-values are computed for total sample sizes of 26 or smaller.) For moderate to large samples, the traditional normal approximation statistic and associated two-tailed p-value is displayed.

The exact p-value and normal approximation p-value are quite different in this example. The exact p-value, when reported, should always be used in preference to the normal approximation. The exact p-value of 0.5859 does not provide any evidence that the two chemical techniques are different.

Computational Notes

The algorithm for the exact p-value is given in Manly (1991). The corrections for continuity and ties for the normal approximation are given in Siegel and Castellan (1988).

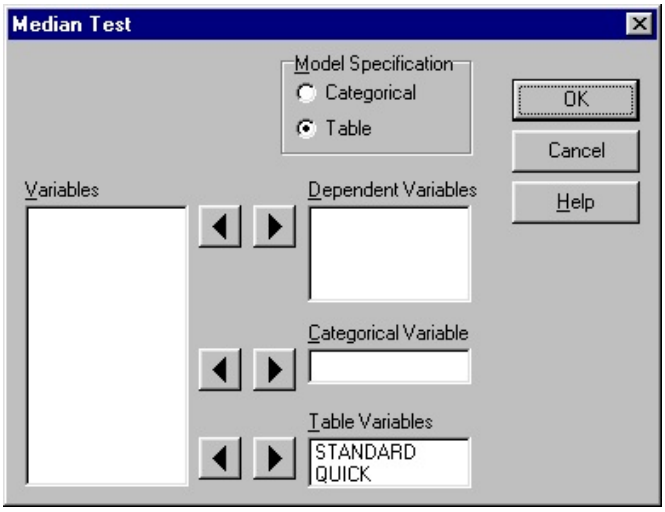
Median Test

The **Median Test** is a nonparametric two-sample test. It tests the hypothesis that the medians for the two groups from which the samples were drawn are equal.

Specification

The analysis can be specified in one of two ways, depending on how the data are stored. If the data from both groups are entered into a single variable and a second categorical variable is used to identify the two groups, use the ***Categorical*** method. Move the variable containing the observed data into the ***Dependent Variable*** box, and the variable that identifies the two groups into the ***Categorical Variable*** box.

If the two groups are entered into *Statistix* as two variables, select the ***Table*** method as illustrated in the dialog box below. Move the two variables to the ***Table Variables*** list box.



Data Restrictions

The chi-square value is not computed for sample sizes less than ten. It is recommended that you use the **Two By Two** procedure to compute Fisher's exact method in such cases.

Example

The data for this example come from Snedecor and Cochran (1980). The goal is to compare the results of a standard, but slow, chemical analysis procedure with a quicker, but potentially less precise, procedure. The variable STANDARD stores the chemical concentrations determined using the standard method, and the variable QUICK stores the concentrations determined using the quicker method.

CASE	STANDARD	QUICK
1	25	23
2	24	18
3	25	22
4	26	28
5	M	17
6	M	25
7	M	19
8	M	16

The analysis is specified on the preceding page. The results are as follows:

Median Test for QUICK - STANDARD

	QUICK	STANDARD	Total
Above Median	2	4	6
Below Median	6	0	6
Total	8	4	12
Ties with Median	0	0	0

Median Value 23.500

Chi-Square 6.00 DF 1 P-value 0.0143

Max. diff. allowed between a tie 0.00001

Cases Included 12 Missing Cases 4

The first step is to find the median for all of the data, which for our example was 23.5. The number of values above and below the median in each sample is tallied, and the two by two table displayed above is created. The number of ties with the median is also displayed, but this information isn't used in the calculations. A value is considered to be tied with the median if it differs by no more than 0.00001.

If the medians for the two groups are equal, we would expect "similar" numbers of values within each group to fall above and below the median. The null hypothesis being tested is that the medians for the two groups are equal. The test amounts to a typical chi-square test for independence or heterogeneity, performed on the two by two table.

The chi-square value in the example is 6.00, which results in a p-value of 0.0143. This supports the idea that the chemical analysis procedures are different.

One-Way AOV

This procedure performs a one-way analysis of variance (Snedecor and Cochran, Chapter 12). The **One-Way AOV** provides statistics for both the fixed effects (Type I) model and the random effects (Type II) model. It also tests for equality of variances between levels, and there are options to perform multiple comparisons of means, contrasts, residual plots, and save fitted values and residuals. The one-way AOV is equivalent to the **Completely Randomized Design** discussed in Chapter 7.

Specification

To use the One-Way AOV procedure, you can organize your data in one of two ways. In the Table method, you create one variable for each of the treatments, then enter the responses observed for each treatment in its own variable. Your second option is to create a single dependent variable and enter all of the responses observed for all of the treatments. Then create a second variable with categorical values (e.g., 1, 2, 3 ...) that represent the treatments. This is called the Categorical method. Both of these methods are illustrated below.

Data Restrictions

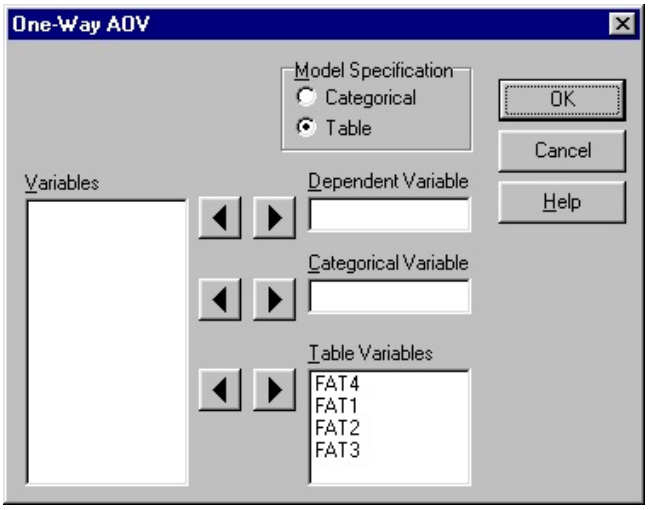
Sample sizes within treatment levels can be unequal. The maximum number of treatment levels is 500. The treatment variable used with the categorical method can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters.

Example

The example below is from Snedecor and Cochran (1980, p. 216). The grams of fat absorbed by batches of doughnuts was measured using four types of fat. The fat absorbed is the response, the fat types are the treatments. To illustrate the ***Table*** method of model specification first, suppose we entered the responses using four variables, FAT1, FAT2, FAT3, and FAT4, each representing one of the four treatments (see data file Sample Data\doughnuts.sx).

CASE	FAT1	FAT2	FAT3	FAT4
1	64	78	75	55
2	72	91	93	66
3	68	97	78	49
4	77	82	71	64
5	56	85	63	70
6	95	77	76	68

The model is specified in the dialog box below. The Table method was selected and the four variables moved to the ***Table Variables*** box.



The results are displayed below.

```
One-Way AOV for: FAT1 FAT2 FAT3 FAT4

Source    DF         SS         MS        F         P
Between    3    1636.50    545.500     5.41    0.0069
Within    20    2018.00    100.900
Total     23    3654.50

Grand Mean 73.750    CV 13.62
                                        Chi-Sq    DF         P
Bartlett's Test of Equal Variances        1.75     3    0.6258
Cochran's Q                   0.4410
Largets Var / Smallest Var    2.9470

Component of variance for between groups    74.1000
Effective cell size                            6.0

Variable    Mean
FAT1       72.000
FAT2       85.000
FAT3       76.000
FAT4       62.000
Observations per Mean             6
Standard Error of a Mean     2.0250
Std Error (Diff of 2 Means) 2.4082
```

A standard analysis of variance table is displayed first. Note that the F test suggests a substantial between-groups (fat types) effect, with a p-value of 0.0069. The F test assumes that the within-group variances are the same for all groups. Bartlett's test for equality of variances tests this assumption; it is shown below the analysis of variance table. The p-value of 0.6258 doesn't suggest that the variances are unequal. Bartlett's test is described in

Snedecor and Cochran (1980, p. 252). Another test of equality of variances, Cochran's Q, is given below Bartlett's test. Cochran's Q statistic is the ratio of the largest within-group variance over the sum of all within-group variances. The ratio of the largest within-group variance over the smallest has also been a popular test for equal variances and is displayed under Cochran's Q; tables are given in Pearson and Hartley (1954).

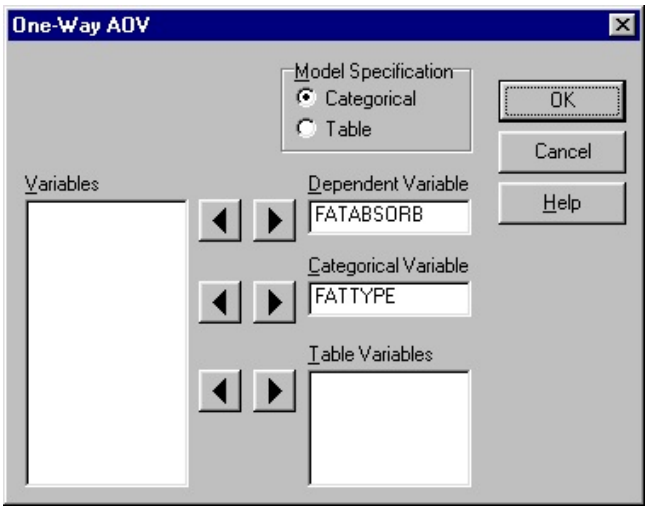
A fixed-effects model (Type I) is appropriate for these data. If a random-effects model were appropriate (Type II), the component of variance for between groups may be of interest (see Snedecor and Cochran, chap. 13). The between-groups variance component and effective cell sample size are displayed below the equality of variance tests. The computation of effective cell size is described on page 246 of Snedecor and Cochran.

The bottom portion of the report lists a table of within-group means, sample sizes, and standard errors of the means. The standard error of the difference of two means is reported when the sample sizes are equal.

We'll use the same analysis to illustrate the ***Categorical*** method of model specification. We now create two variables, a dependent variable FATABSORB and the treatment variable FATTYPE (see data file Sample Data\doughnuts.sx).

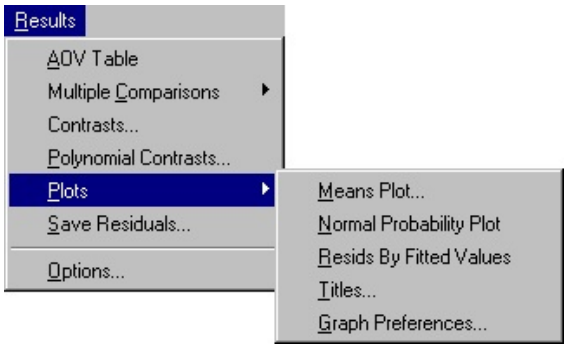
CASE	FATABSORB	FATTYPE
1	64	1
2	72	1
3	68	1
4	77	1
5	56	1
6	95	1
7	78	2
8	91	2
9	97	2
10	82	2
11	85	2
12	77	2
13	75	3
14	93	3
15	78	3
16	71	3
17	63	3
18	76	3
19	55	4
20	66	4
21	49	4
22	64	4
23	70	4
24	68	4

The model is specified in the dialog on the next page.



One-Way AOV Results Menu

Once the one-way AOV is computed and displayed, a ***Results*** pull-down menu appears in the main menu at the top of the *Statistix* window. Click on the Results menu to display the One-Way AOV results menu:



Select AOV Table to have the AOV table and means displayed again. Select Options to display the original One-Way AOV dialog box you used to specify the model. You can change the details of the model you specified and recompute the analysis. The remaining procedures are discussed briefly below; see Chapter 7 for a thorough discussion.

Multiple Comparisons

The **Multiple Comparisons** procedures are used to compare the means of the different groups. Other names for these procedures are mean separation tests, multiple range tests, and tests for homogeneity of means. The

multiple comparisons tests performed by *Statistix* are divided into three categories: all-pairwise comparisons, comparisons with a control, and comparisons with the best. See Chapter 7 for a complete discussion of these procedures and examples of each.

Plots

The Plots submenu offers three plots. The **Means Plot** produces a line-plot or a bar-chart of the means for each group. See Chapter 7 for an example.

The **Normal Probability Plot** plots the residuals against the rankits. Plots for normal data form a straight line. The Shapiro-Wilk statistic for normality is also reported on the plot. See Chapter 9 for details.

The **Resids By Fitted Values** plot is useful for examining whether the variances are equal among the groups. If the order of the groups is meaningful, then systematic departures from equality can be seen in the plot.

The **Titles** procedure is used to changes the titles of the plot displayed. The **Graph Preferences** procedure is used to change details of the plot, such as font and symbol type. See Chapter 1 for details.

Contrasts

This procedure lets you compute a test for any linear contrast of the group means. Linear contrasts are linear combinations of the means, and they're value for examining the "fine structure" of the data after the overall F test indicates that the treatment effect is significant. The test computes the contrast value, sums of squares for the contrast, Scheffe's F, and Student's t-statistic. See Chapter 7 for details and an example.

Polynomial Contrasts

This procedure computes the polynomial decomposition of the treatment sums of squares. This is useful for determining the existence and nature of trends (i.e., linear, quadratic, etc.) in the treatment level means. See Chapter 7 for details and an example.

Save Residuals

The Save Residuals procedure is used to save the fitted values and residuals in new or existing variables for later analysis. This option is only available if you use the Categorical method of model specification. Simply enter variable names in the spaces provided for fitted values and residuals. The fitted value for an observation in a one-way AOV is the class mean. The residuals are computed as the observed value minus the fitted value.

Kruskal-Wallis One-Way AOV

The **Kruskal-Wallis One-Way AOV** procedure performs a nonparametric one-way analysis of variance. The Kruskal-Wallis statistic is computed as are the results of a parametric one-way analysis of variance applied to the ranks.

Specification

To use the **Kruskal-Wallis One-Way AOV** procedure, you can organize your data in one of two ways. The Table method is where you enter the responses observed for each of the treatments in its own variable. The Categorical method is where you enter all of the observed responses in a single dependent variable and enter the treatment levels in a second grouping variable. Both the Table method and the Categorical method are illustrated in the example below.

Data Restrictions

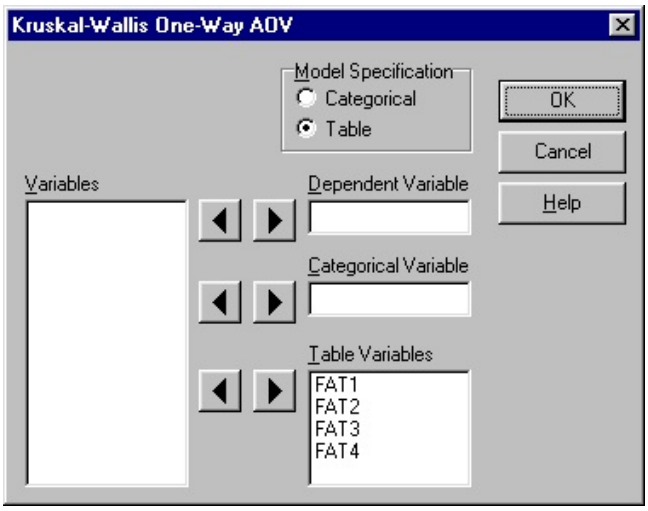
Sample sizes within treatments can be unequal. The maximum number of treatment levels is 500.

Example

The example data are from Snedecor and Cochran (1980, p. 216) and are also used as example data for the **One-Way AOV** procedure. The grams of fat absorbed by batches of doughnuts were measured using four types of fat. The fat types are the treatments, and the fat absorbed is the response. We'll illustrate the ***Table*** method of model specification first. Suppose we enter the responses using four variables—FAT1, FAT2, FAT3, and FAT4. Each variable represents one of the four treatments.

CASE	FAT1	FAT2	FAT3	FAT4
1	64	78	75	55
2	72	91	93	66
3	68	97	78	49
4	77	82	71	64
5	56	85	63	70
6	95	77	76	68

The model is specified in the dialog box on the next page.



The Table method was selected and the four variables moved to the ***Table Variables*** box. The results are displayed on the next page.

The Kruskal-Wallis test is a generalization of the Wilcoxon rank sum test. The data are first ranked irrespective of group. Tied values are assigned their average rank (Hollander and Wolfe 1973). Values are assumed to be tied if they are within 0.00001 of one another. If each of the groups had similar distributions, the mean ranks for all groups would be expected to be "similar". The null hypothesis being tested is that each of the groups has the same distribution. Strictly speaking, if the null hypothesis is rejected, the alternative is that the distributions for the groups differ, although in practice it's typical to assume that the differences are due to differences in the central values of the groups.

Kruskal-Wallis One-Way Nonparametric AOV

Variable	Mean Rank	Sample Size
FAT1	11.3	6
FAT2	19.5	6
FAT3	13.6	6
FAT4	5.7	6
Total	12.5	24

Kruskal-Wallis Statistic 11.8322
P-Value, Using Chi-Squared Approximation 0.0080

Parametric AOV Applied to Ranks

Source	DF	SS	MS	F	P
Between	3	590.58	196.861	7.06	0.0020
Within	20	557.42	27.871		
Total	23	1148.00			

Total number of values that were tied 8
Max. diff. allowed between ties 0.00001

Cases Included 24 Missing Cases 0

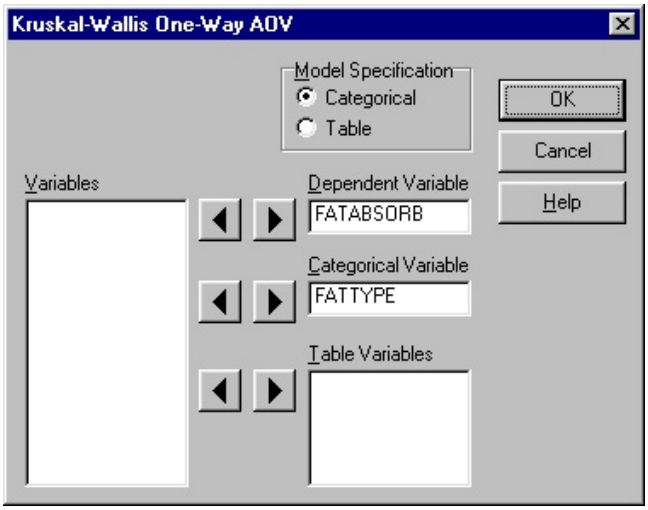
The analysis of the example is consistent with the parametric **One-Way AOV**. The p-value of 0.0080 suggests that the mean ranks for the groups are dissimilar enough to conclude that the fat types differ.

Conover and Iman (1981) proposed first ranking the data and then applying the usual parametric procedure for computing a one-way analysis of variance. The results of this procedure are displayed underneath the Kruskal-Wallis test. The results are interpreted in the same way as the usual analysis of variance, comparing the within-group variance to the between-group variance. Please note that the usual F test is generally anti-conservative, giving significant results more often than it should (Iman and Davenport 1976, 1980). This is perhaps the case in this example. Here the p-value is smaller than that observed with the parametric analysis of variance or the Kruskal-Wallis test.

We can use the same analysis to illustrate the ***Categorical*** method of model specification. The data are entered in two variables—a dependent variable FATABSORB and a categorical variable FATTYPE. The data are listed in the table on the next page.

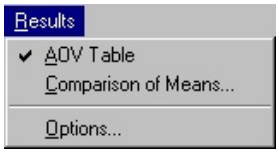
CASE	FATABSORB	FATTYPE	CASE	FATABSORB	FATTYPE
1	64	1	13	75	3
2	72	1	14	93	3
3	68	1	15	78	3
4	77	1	16	71	3
5	56	1	17	63	3
6	96	1	18	76	3
7	78	2	19	55	4
8	91	2	20	66	4
9	97	2	21	49	4
10	82	2	22	64	4
11	85	2	23	70	4
12	77	2	24	68	4

The model is specified as follows:



Kruskal-Wallis Results Menu

Once the Kruskal-Wallis AOV is computed and displayed, a ***Results*** pull-down menu appears in the menu at the top of the *Statistix* window.

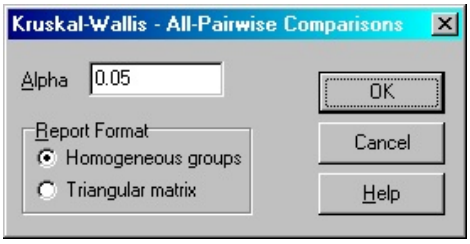


Select AOV Table to have the Kruskal-Wallis results displayed again. Select Options to display the original dialog box you used to specify the model. You can change the details of the model you specified and

recompute the analysis. The Comparison of Mean Ranks procedure is discussed below.

All-Pairwise Comparisons

The **All-Pairwise Comparisons** option is used to compare the mean ranks of the different groups. This procedure identifies subsets of similar (homogeneous) means.



To use the All-Pairwise Comparisons procedure, you enter a value for alpha, the rejection level, and select the report format. The results for the doughnut fat absorption example described on page 147 using the Homogeneous Groups report format are given below.

```
Kruskal-Wallis All-Pairwise Comparisons Test

Variable     Mean   Homogeneous Groups
FAT2       19.500   A
FAT3       13.583   AB
FAT1       11.250   AB
FAT4       5.6667    B

Alpha              0.05
Critical Z Value   2.638     Critical Value for Comparison   10.771
There are 2 groups (A and B) in which the means
are not significantly different from one another.
```

The mean ranks are sorted in descending order so the largest one is listed in the first row. The columns of letters under the heading "Homogeneous Groups" indicate which means are not significantly different from one another. Group A contains the mean ranks for FAT2, FAT3, and FAT1. Group B contains the mean ranks for FAT3, FAT1, and FAT4. We conclude that FAT2 is different from FAT4 since neither group A or B contains both fat types.

The Triangle Matrix report format makes it easier to identify pairs of means that are different. An example for the same data appears on the next page. The numbers in the body of the triangular shaped table are differences between mean ranks. Significant differences are indicated with an asterisk.

Kruskal-Wallis All-Pairwise Comparisons Test

Variable	Mean	FAT1	FAT2	FAT3
FAT1	11.250			
FAT2	19.500	8.250		
FAT3	13.583	2.333	5.917	
FAT4	5.6667	5.583	13.833*	7.917

Alpha	0.05		
Critical Z Value	2.638	Critical Value for Comparison	10.771

The comparison procedure controls the experimentwise error rate. Like the Bonferroni comparison of means procedure for the parametric AOV (see Chapter 7), the test becomes increasingly conservative as the number of means increases. A larger than normal rejection level (e.g., 0.10-0.25) is often used when testing large numbers of means. See Daniel (1990) for details.

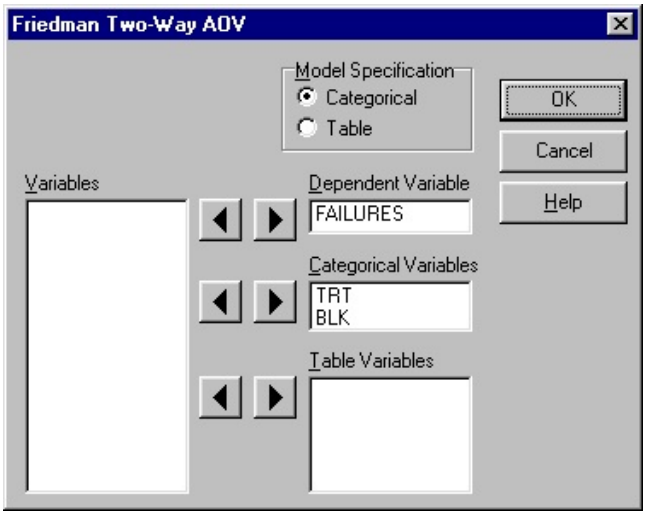
Friedman Two-Way AOV

The Friedman nonparametric two-way analysis of variance is used to analyze two-way designs without replication. The results are equivalent to Kendall's coefficient of concordance (Conover 1980).

Specification

The analysis can be specified in one of two ways, depending on how the data are arranged in the variables. A two-way analysis of variance requires that each observation be classified by two factors. For the ***Categorical*** method, all observations of the ***Dependent Variable*** are in one variable. The two factors are indicated by two ***Categorical Variables***. This is illustrated in the dialog box on the next page.

In the ***Table*** method, the levels for the one factor are represented by the variables themselves. The levels for the other factor are then represented by the cases. To specify the model, move the names of the variables that represent the column factor to the ***Table Variables*** box (see the example on pages 155-156).



Data Restrictions

There can be only one observation per cell; no replication is permitted. Missing values can't be included. You can have up to 500 levels in each of the two treatment factors.

Example

This example is a randomized block design used in Snedecor and Cochran (1980, sect. 14.2). The same data are used for as an example in the **Randomized Block Design** section of Chapter 7, where a parametric two-way analysis of variance is computed. The dependent variable is the number of soybeans out of 100 that failed to emerge, and the treatments were various fungicides (the first treatment level was a no-fungicide control).

To illustrate the Categorical method of model specification, we entered the observed counts into a single variable named Y. The fungicides were numbered 1 through 5 and entered into a variable named TRT. The blocks (replicates) were numbered 1 through 5 and entered into a variable named BLK.

CASE	Y	TRT	BLK	CASE	Y	TRT	BLK
1	8	1	1	14	8	3	4
2	10	1	2	15	10	3	5
3	12	1	3	16	3	4	1
4	13	1	4	17	5	4	2
5	11	1	5	18	9	4	3
6	2	2	1	19	10	4	4
7	6	2	2	20	6	4	5
8	7	2	3	21	9	5	1
9	11	2	4	22	7	5	2
10	5	2	5	23	5	5	3
11	4	3	1	24	5	5	4
12	10	3	2	25	3	5	5
13	9	3	3				

The **Transformation** CAT function can be used to generate repetitive sequences, like those seen for TRT and BLK. After entering the 25 values for Y, we can use the **Transformation** expressions TRT = CAT (5 5) and BLK = CAT (5 1) to create these variables. The data are available in the file Sample Data\soybeans.sx.

The analysis is specified on the preceding page. The results are presented on the next page.

For the first factor, which in this case is TRT, the observations are first ranked within the second factor (BLK). If there were no differences between the treatment levels, the mean ranks (averaged across blocks) for the different treatment levels would be expected to be "similar". Tied observations are given a mean rank (Hollander and Wolfe 1973). Values are considered to be tied if they are within 0.00001 of one another. "Corrected for Ties" appears in the display when the Friedman statistic is based on data that contain ties. For this example, ties within blocks were found. It appears that there are definite treatment effects because the p-value is fairly small (0.0530).

Friedman Two-Way Nonparametric AOV for FAILURES = TRT BLK

TRT	Mean Rank	Sample Size
Control	4.70	5
Fung #1	2.20	5
Fung #2	3.40	5
Fung #3	2.50	5
Fung #4	2.20	5

Friedman Statistic, Corrected for Ties	9.3469
P-value, Chi-Squared Approximation	0.0530
Degrees of Freedom	4

BLK	Mean Rank	Sample Size
1	1.80	5
2	3.10	5
3	3.50	5
4	3.90	5
5	2.70	5

Friedman Statistic, Corrected for Ties	5.3061
P-value, Chi-Squared Approximation	0.2573
Degrees of Freedom	4

Max. diff. allowed between ties 0.00001

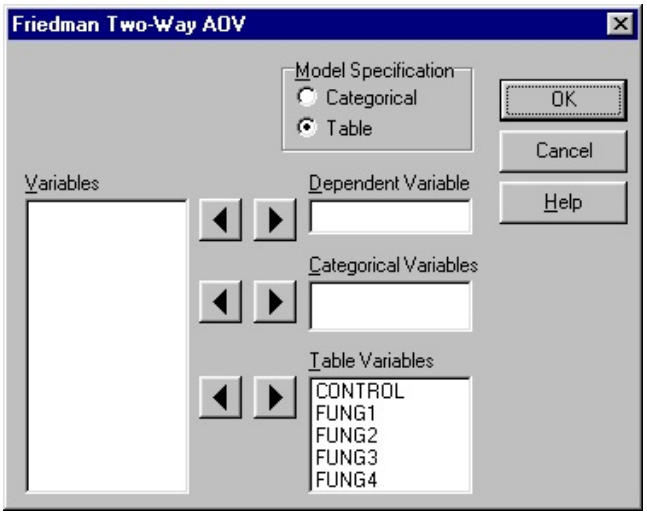
Cases Included 25 Missing Cases 0

To examine block effects, the role of the variables is reversed. The observations are now ranked within treatment levels. Ties among the observations within treatment levels were found, as indicated by the message with the Friedman statistic. There appears to be little evidence of block effects (0.2573). As with parametric analysis of variance, testing the block effect will generally not be of much interest.

To use the ***Table*** method, we'd choose one factor, say fungicide, to represent columns and enter the data for each fungicide into a separate variable. The cases then represent the blocks (replicates).

CASE	CONTROL	FUNG1	FUNG2	FUNG3	FUNG4
1	8	2	4	3	9
2	10	6	10	5	7
3	12	7	9	9	5
4	13	11	8	10	5
5	11	5	10	6	3

The order of the cases is very important with this format. The first case corresponds to the first experimental block, the second case corresponds to the second block, and so on. The model is then specified in the dialog box displayed on the next page.



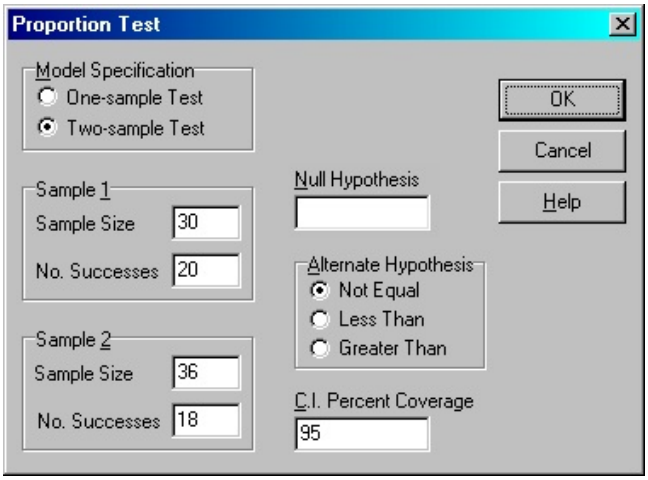
Here the variables that represent the fungicide treatments are moved to the ***Table Variables*** box.

The Friedman test is often performed as a companion analysis to the parametric two-way analysis of variance, especially when the assumption of normality in parametric analysis of variance is suspect. It's not as powerful as the parametric analysis, but it usually performs quite well.

Proportion Test

The **Proportion Test** procedure is used to perform one- and two-sample hypothesis tests and compute confidence intervals for proportions.

Specification



The one-sample test is used to test whether a proportion differs from a hypothesized value. The two-sample test is used to compare proportions in two independent samples. Select either the One-sample Test or the Two-sample Test.

The tests requires that you enter values for the sample sizes and the number of successes (the number of times the event of interest was observed). For a one-sample test, you must enter a value for the ***Null Hypothesis***. The null hypothesis for the two-sample test is always that the two proportions are equal. Select the two-sided alternative hypothesis "not equal", or a one-sided alternative "less than" or "greater than". For the one-sample test, the alternative hypothesis "less than" means the proportion is less than the null hypothesis. For the two-sample test, the alternative hypothesis "less than" means that the proportion from the first sample is less than the second. Enter a value for the percent coverage for confidence intervals.

Example

The dialog box above specifies a two-sample test using the two-sided alternative hypothesis. The results are displayed on the next page.

Two-Sample Proportion Test

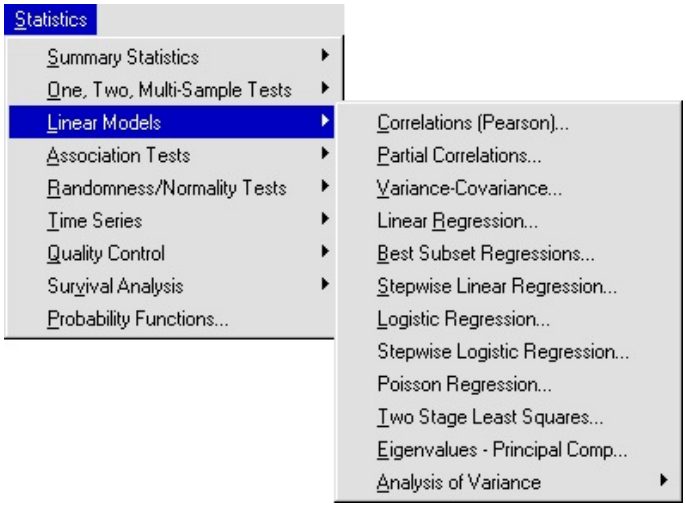
	Sample 1	Sample 2	
Sample Size	30	36	
Successes	20	18	
Proportion	0.66667	0.50000	
Null Hypothesis:	P1 = P2		
Alternative Hyp:	P1 <> P2		
Difference	0.16667		
SE (diff)	0.12218		
Z (uncorrected)	1.36	P	0.1725
Z (corrected)	1.11	P	0.2653
Fisher's Exact			0.2152
95% Confidence Interval of Difference			
Lower Limit	-0.07279		
Upper Limit	0.40613		

The report displays the computed proportions for the two samples and the difference. Two versions of the normal approximation tests are provided: the uncorrected test and the test corrected for continuity. P-values for both tests are given. When the combined sample size is less than 500, Fisher's exact test is also displayed. Fisher's exact test, when available, is the preferred test. In the example, the p-values for all three tests agree: there is no evidence that the two proportions are different.

C H A P T E R

6

Linear Models



Statistix offers you a comprehensive selection of linear model procedures, which include regression, analysis of variance, and analysis of covariance, Linear models are among the most powerful and popular tools available for data analysis.

The **Correlations (Pearson)** procedure displays the correlation matrix for a set of variables.

The **Partial Correlations** procedure computes the correlations of a set of

independent variables with a dependent variable after adjusting for the effects of another set of independent variables.

The **Variance-Covariance** procedure displays the sample variance-covariance matrix for a set of variables.

The **Linear Regression** procedure performs simple and multiple linear regression. You can compute predicted values and prediction intervals for any set of independent variable values. Extensive residual analysis options are available for model evaluation. The sensitivity of the regression coefficients to errors in the independent variables can be examined.

The **Best Subset Regressions** procedure generates a list of "best" subset models for a specified regression model.

The **Stepwise Linear Regression** procedure performs forward and backward stepwise linear regression in search of good subset regression models.

The **Logistic Regression** procedure is appropriate for a situation where the dependent variable consists of "binary" data. Common examples of binary data are yes or no responses and success or failure outcomes.

The **Stepwise Logistic Regression** procedure performs forward and backward stepwise logistic regression.

The **Poisson Regression** procedure is appropriate for situations where the dependent variable consists of discrete counts. See the section titled **Analyzing Proportions and Counts** on page 211 for more background on when and why Poisson Regression and Logistic Regression are useful.

The **Two Stage Least Squares Regression** procedure is used to develop a prediction equation when one or more of the predictor variables, or right hand side variables, is an endogenous variable. An endogenous variable is one that is determined by the system of equations being solved. The model also requires at least one exogenous variable. An exogenous variable is one whose value is determined outside the system of equations.

The **Eigenvalues-Principal Components** procedure computes the eigenvectors and eigenvalues and the principal components for a set of variables. It's often used in regression when the independent variables are highly correlated, and it's also a useful multivariate analysis in its own

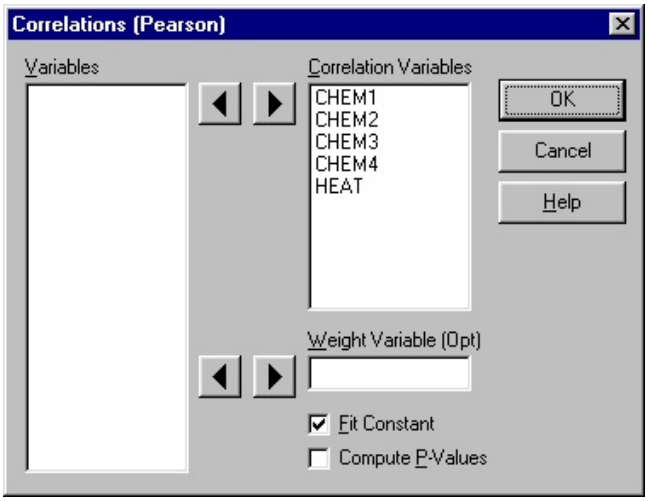
right.

The **Analysis of Variance** submenu offers a long list of AOV designs: completely randomized, complete block, Latin square, balanced lattice, full and fractional factorial, split-plot, strip-plot, split-split-plot, split-strip-plot, and repeated measures designs. All of the analysis of variance procedures contain a results menu offering numerous powerful options including multiple comparisons, linear contrasts, polynomial contrasts, means plot, and residual plots. There are so many AOV procedures and options that we've devoted Chapter 7 to them.

Correlations (Pearson)

The **Correlations** procedure computes a correlation matrix for a list of variables. Correlations, also called Pearson or product-moment correlations, indicate the degree of linear association between variables.

Specification



Select the variables for which you want correlations computed. Highlight the variables you want in the ***Variables*** list, then press the right-arrow button to move the highlighted variables to the ***Correlations Variables*** list box. If you want to use a weighting factor, move the name of the variable

containing the weights to the ***Weight Variable*** box. Use the ***Fit Constant*** check box to specify a model with a constant fitted (checked) or a model forced through the origin (not checked). Check the ***Compute P-Values*** check box to have p-values for the correlation coefficients displayed.

Data Restrictions

You can select up to 50 variables. If a case in your data has missing values for any of the variables selected, the entire case is deleted (listwise deletion). Negative weights are not allowed.

Example

We'll use the Hald data from Draper and Smith (1966) for our example. The variable HEAT is the cumulative heat of hardening for cement after 180 days. The variables CHEM1, CHEM2, CHEM3, and CHEM4 are the percentages of four chemical compounds measured in batches of cement. The data are listed below (see also Sample Data\Hald.sx).

CASE	HEAT	CHEM1	CHEM2	CHEM3	CHEM4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

The model specification is displayed on the preceding page. The results are as follows:

Correlations (Pearson)

	CHEM1	CHEM2	CHEM3	CHEM4
CHEM2	0.2286			
CHEM3	-0.8241	-0.1392		
CHEM4	-0.2454	-0.9730	0.0295	
HEAT	0.7307	0.8163	-0.5347	-0.8213

Cases Included 13 Missing Cases 0

Computational Notes

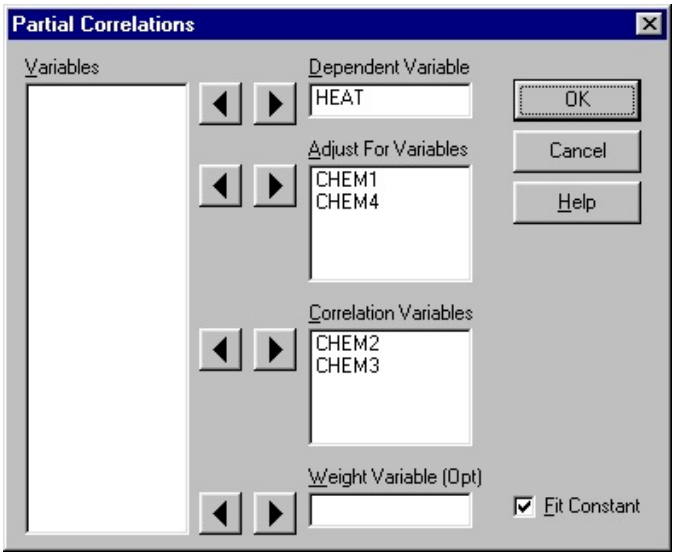
First, the matrix of sums of squares and cross products corrected for the means is calculated using the method of updating, also known as the method

of provisional means. The results are more accurate than if the usual "computational equations" were used. If the model is fit without a constant, the matrix corrected for the means is first computed and then "uncorrected".

Partial Correlations

The **Partial Correlations** procedure computes the "residual" correlation between variables after adjustment for correlations with another set of variables. Partial correlations are often used in manual stepwise regression procedures to help you decide which variable should be included next in the regression.

Specification



First select the ***Dependent Variable***. Next select the independent variables for which you want the correlations adjusted and move them to the ***Adjust For Variables*** list. Then select the independent variables for which you want to compute partial correlations and move them to the ***Correlation Variables*** list. The resulting table displays a correlation coefficient between the Dependent Variable and each of the Correlation Variables, adjusted for the Adjust For Variables.

If you want a weighting factor, move the variable containing the weights to the ***Weight Variable*** box. Use the ***Fit Constant*** check box to specify a constant fitted model (checked) or a model forced through the origin (not checked).

Data Restrictions

The total number of independent variables can't exceed 50. Missing values or zero weights cause the entire case to be dropped. Negative weights aren't allowed.

Example

We'll use the Hald data from Draper and Smith (1966) for our example. The variable HEAT is the cumulative heat of hardening for cement after 180 days. The variables CHEM1, CHEM2, CHEM3, and CHEM4 are the percentages of four chemical compounds measured in batches of cement. The data are listed on page 162 and can be obtained from the file Sample Data\Hald.sx.

The model specified in the dialog box on the preceding page is used to compute the partial correlations for CHEM2 and CHEM3 on HEAT, adjusted for CHEM1 and CHEM4. The results are as follows:

```
Partial Correlations with HEAT  Cumulative Heat of Hardening For Cement
Controlled for CHEM1 CHEM4

CHEM2         0.5986
CHEM3        -0.5657

Cases Included 13    Missing Cases 0
```

The partial correlation of CHEM2 with HEAT after the effects of CHEM1 and CHEM4 have been removed is 0.5986. The partial correlation of CHEM3 with HEAT, adjusted for CHEM1 and CHEM4, is -0.5657.

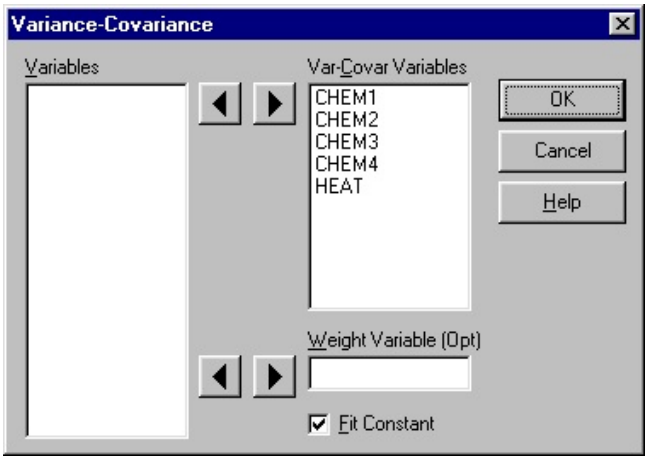
Computational Notes

First the matrix of sums of squares and cross products corrected for the means is calculated by using the method of updating, also known as the method of provisional means. The results are more accurate than if the usual "computational equations" were used. If you fit a model without a constant, the matrix corrected for the means is first computed and then "uncorrected". The matrix is ordered so that the "variables adjusted for" are on the left, and these variables are then "swept" over (Seber 1977) to produce the partial correlations.

Variance-Covariance

The **Variance-Covariance** procedure computes the variances and covariances for a list of variables.

Specification



Select the variables for which you want variances and covariances computed. Highlight the variables you want in the ***Variables*** list, then press the right-arrow button to move them to the ***Var-Covar Variables*** list box. Select the name of a weighting variable for weighted variances-covariances. If you specify a model without a constant (uncheck the ***Fit Constant*** check box), *Statistix* computes the sums of squares and cross products uncorrected for the means.

Data Restrictions

Up to 50 variables can be specified. If a case in your data has missing values for any of the variables, the entire case is deleted (listwise deletion). Negative weights aren't permitted.

Example

We'll use the Hald data from Draper and Smith (1966) for our example. The variable HEAT is the cumulative heat of hardening for cement after 180 days. The variables CHEM1, CHEM2, CHEM3, and CHEM4 are the percentages of four chemical compounds measured in batches of cement. The data are listed on the next page.

CASE	HEAT	CHEM1	CHEM2	CHEM3	CHEM4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

The model is specified in the dialog box on the preceding page. The variances and covariances are computed for all variables. The results are displayed below.

Variance - Covariance Matrix

	CHEM1	CHEM2	CHEM3	CHEM4	HEAT
CHEM1	34.6026				
CHEM2	20.9231	242.141			
CHEM3	-31.0513	-13.8782	41.0256		
CHEM4	-24.1667	-253.417	3.16667	280.167	
HEAT	64.6635	191.079	-51.5192	-206.808	226.314

Cases Included 13 Missing Cases 0

The values displayed on the diagonal of the matrix are the variances, the off-diagonal values are covariances.

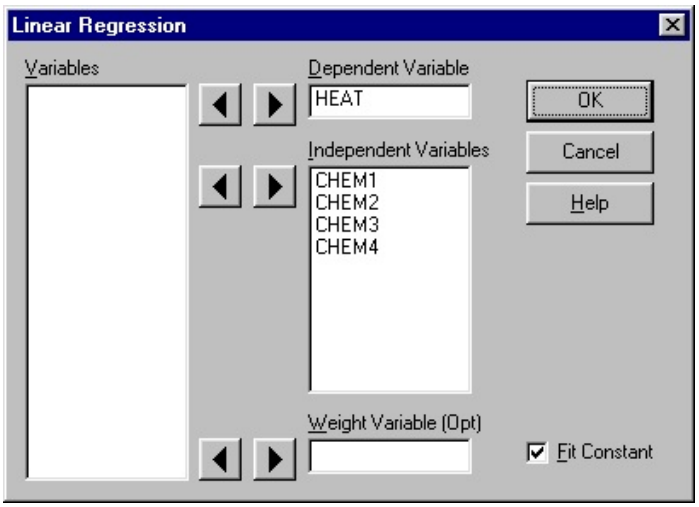
Computational Notes

The sums of squares and cross products matrix corrected for the means are calculated by using the method of updating, also known as the method of provisional means. The results are more accurate than if the usual "computational equations" were used. If you fit a model without a constant, the matrix corrected for the means is first computed and then "uncorrected".

Linear Regression

The **Linear Regression** procedure performs simple and multiple linear regression. Linear regression is a popular technique for examining linear relationships between a response (dependent) variable and one or more predictor (independent) variables. This procedure can perform both weighted and unweighted least squares fitting, and you can specify no-intercept models. Extensive analysis of variance and residual analysis options are available. You can also compute predicted values and prediction intervals. You can examine the sensitivity of the regression coefficients to measurement errors in the independent variables.

Specification



To specify a regression model, first select the dependent variable. Highlight the variable in the ***Variables*** list, then press the right-arrow button next to the ***Dependent Variable*** box to move the highlighted variable into that box. Then select one or more independent variables and move them to the ***Independent Variables*** list box. For weighted regression, move the variable containing the weights to the ***Weight Variable*** box. Use the ***Fit Constant*** check box to specify a constant fitted model (checked) or a model forced through the origin (not checked).

Data Restrictions

You can include up to 50 independent variables in the model. If any values within a case are missing for any of the variables in the model, the case is dropped (listwise deletion). If an independent variable is too highly

correlated with a linear combination of other independent variables in the model (collinearity), it's dropped from the model. If you specify a weight variable, cases with negative weights are deleted.

Example

We'll use the data from Hald (1952) for our example. Draper and Smith (1966) used the same data set to illustrate selecting the "best" regression equation. The variable HEAT is the cumulative heat of hardening for cement after 180 days. The variables CHEM1, CHEM2, CHEM3, and CHEM4 are the percentages of four chemical compounds measured in batches of cement. The goal is to relate the heat of hardening to the chemical composition. The data are listed below and can be obtained from the file Sample Data\Hald.sx.

CASE	HEAT	CHEM1	CHEM2	CHEM3	CHEM4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

The full model is specified in the dialog box on the preceding page. The results are listed below.

Unweighted Least Squares Linear Regression of HEAT

Predictor Variables	Coefficient	Std Error	T	P	VIF
Constant	62.4054	70.0710	0.89	0.3991	
CHEM1	1.55110	0.74477	2.08	0.0708	38.5
CHEM2	0.51017	0.72379	0.70	0.5009	254.4
CHEM3	0.10191	0.75471	0.14	0.8959	46.9
CHEM4	-0.14406	0.70905	-0.20	0.8441	282.5

R-Squared	0.9824	Resid. Mean Square (MSE)	5.98295
Adjusted R-Squared	0.9736	Standard Deviation	2.44601

Source	DF	SS	MS	F	P
Regression	4	2667.90	666.975	111.48	0.0000
Residual	8	47.86	5.983		
Total	12	2715.76			

Cases Included 13 Missing Cases 0

The regression coefficient table gives the regression coefficients (slopes) associated with the independent variables and their standard errors, t-statistics, associated p-values, and variance inflation factors (VIF). You

use p-values to test whether the slopes are significantly different from zero if all other variables are already in the model.

Large VIF's in a multiple regression indicate that collinearity is a problem. The VIF represents the increase in variance of a coefficient due to correlation between the independent variables. Values of 7.0 or 10.0 have been suggested for the cutoff of what constitutes a "high" value.

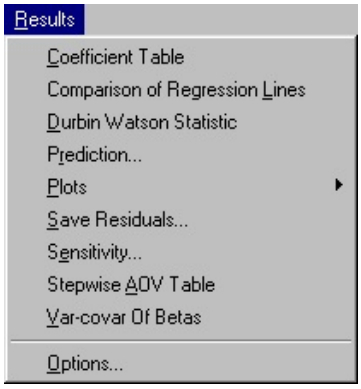
Statistix provides several other summary statistics and an analysis of variance table for the regression, including the F test and associated p-value for the significance of the overall model. In our example, the overall F is 111.48, with a p-value of 0.0000. This indicates that at least some of the independent variables are important in explaining the observed variation in HEAT. From the coefficient column, the regression equation is found to be:

```
  HEAT = 62.4 + 1.55×CHEM1 + 0.510×CHEM2 + 0.102×CHEM3 -
0.144×CHEM4
```

However, the t tests suggest that some of the coefficients are not significantly different from zero. The high VIF's warn us that collinearity among the independent variables is a problem. Selecting the best model is discussed on page 186.

Regression Results Menu

Once the regression analysis is computed and displayed, a ***Results*** pull-down menu appears on the menu at the top of the *Statistix* window. Click on the Results menu to display the regression results menu shown below.



Select Coefficient Table from the menu to redisplay the regression coefficient table. Select Options to return to the main dialog box used to

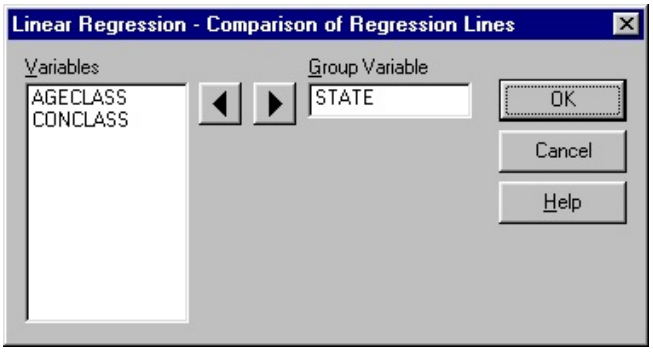
specify the regression model. The remaining options are described on the following pages.

Comparison of Regression Lines

Frequently, the relationship between two variables X and Y is studied in samples obtained by different investigators, or in different environments, or at different times. In summarizing the results, the question naturally arises: can the regression lines be regarded as the same?

This procedure compares simple regression lines for two or more groups. It tests for equality of variance, slope, and elevation. This test can only be used for simple regression, that is, when there is only one independent variable.

First, run the Linear Regression procedure. Then select the Comparison of Regression Lines from the Results menu and select a ***Group Variable*** identifying two or more groups. The group variable can be any data type (real, integer, date, or string) and can have up to 500 levels.



As an example, consider the cholesterol and age data from Chapter 4. In a survey of women from two states, cholesterol concentration and age data were collected. We first perform linear regression of cholesterol concentration on age and find a linear relationship. We now want to determine if the regression lines are the same for subjects from the two states the data were collected: Iowa and Nebraska. We select the variable STATE as show in the dialog box above. The results are shown on the next page.

Comparison of Regression Lines for CONC = AGE

STATE	N	Intercept	Slope	MSE
Iowa	11	35.8112	3.23814	2391.22
Nebraska	19	101.298	2.52044	1580.82

	F	DF	P
Equality of Variances	1.51	9, 17	0.2210
Comparison of Slopes	0.38	1, 26	0.5425
Comparison of Elevations	3.00	1, 27	0.0947

The statistics for the individual regression lines are listed first. Below that are three F tests: for equality of variance, slopes, and elevations. We first compare the variances for the two regression lines using Bartlett's test. The p-value of 0.2210 allows us to conclude that there are no real differences between the variances. Assuming homogeneity of residual variances, we now compare the two slopes, 3.24 for Iowa, and 2.52 for Nebraska. The p-value of 0.5425 does not suggest that the slopes are different. Assuming parallel lines and homogeneous variance, we now consider the test for comparison of elevations. The p-value of 0.0947 is not quite significant. We conclude that the regression lines are the same for the two states.

You can use the **Scatter Plot** procedure discussed in Chapter 4 to visually compare regression lines using a grouping variable. See Snedecor and Cochran (1980) for details of these tests.

Durbin-Watson Test

This option computes the Durbin-Watson test for autocorrelation for a particular regression model. In the example below, the regression model was specified as HEAT = CHEM2 CHEM3. The results are as follows:

Durbin-Watson Test for Autocorrelation

```
Durbin-Watson Statistic  1.7498

P-Values, using Durbin-Watson'S Beta Approximation:
  P (positive corr) = 0.2846,  P (negative corr) = 0.7154

Expected Value of Durbin-Watson Statistic     2.0359
Exact Variance of Durbin-Watson Statistic    0.23450

Cases Included 13    Missing Cases 0
```

The Durbin-Watson statistic and approximate observed significance levels (p-values) are displayed. Use the Durbin-Watson statistic to test whether the random errors about the regression line exhibit autocorrelation. In our example, there is little suggestion of either positive (p=0.2846) or negative (p=0.7154) autocorrelation.

Most tests in regression are based on the assumption that the random errors are independent. Violation of this assumption due to autocorrelation can invalidate the results of these tests. Another reason to check for autocorrelation is because it may suggest that you need additional independent variables in the model.

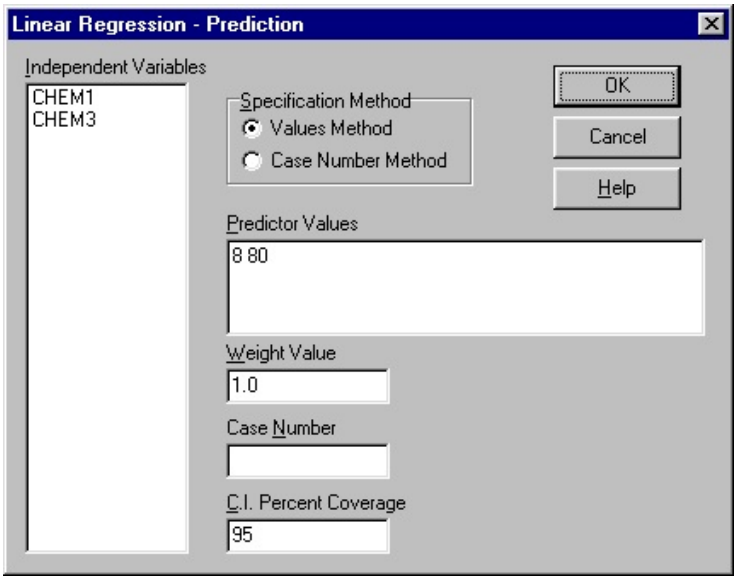
Autocorrelation can occur when the data have some natural sequence, i.e., the observations are ordered in time or space. You should always consider the possibility of autocorrelation in trend data, such as price or population levels over time. Positive autocorrelation results when large positive errors tend to be followed by large positive errors and large negative errors tend to be followed by large negative errors. Negative autocorrelation is less common and results when large errors tend to be followed by large errors with the opposite sign. Chatterjee and Price (1991) give more detail and some examples of the application of the Durbin-Watson test.

If there is neither positive nor negative autocorrelation, the Durbin-Watson statistic will be close to 2. A value close to 0 suggests positive autocorrelation, and a value close to 4 suggests negative autocorrelation. The observed significance levels (p-values) are calculated with the beta distribution approximation suggested by Durbin and Watson (1971). Their results indicated that this approximation usually works well even for relatively small sample sizes. (*Statistix* will not compute the test for samples with fewer than ten cases). The procedures to calculate the significance level, the expected value, and the variance are described in Durbin and Watson (1951). The beta approximation can't be used when the variance of the Durbin-Watson statistic is large, in which case, the p-values are not computed.

The **Runs Test** and **Shapiro-Wilk Test** are also useful for examining whether the test assumptions in regression have been violated.

Prediction

This option computes the predicted or fitted value of the dependent variable in a regression for values of the independent variable(s) you specify. The values for the independent variables can be indicated in two ways. One is simply to enter the list of desired values. The other is to enter a case number that contains the desired values for the independent variables. You choose a method by selecting one of the ***Specification Method*** radio buttons.



The ***Value Method*** is illustrated in the dialog box above. The values 8 and 80 are entered in the ***Predictor Values*** box for the variables CHEM1 and CHEM3. The default value of 1.0 is used for the weight, so the resulting prediction intervals are for a single future observation. If you specify a weight w, the prediction interval is for the mean of w future observations. You can also specify the percent coverage for confidence intervals for the fitted value. Enter a value in the ***C.I. Percent Coverage*** box. The results for the example are displayed below.

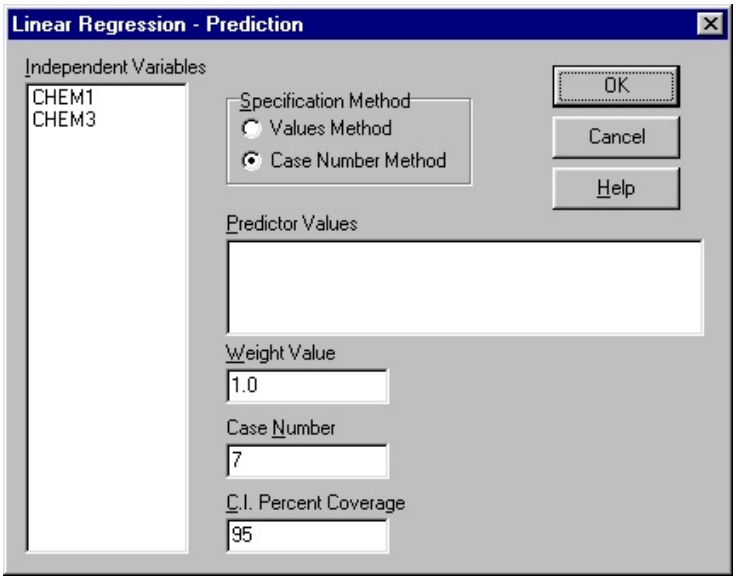
```
Predicted/Fitted Values of HEAT

Lower Predicted Bound    -6.9581     Lower Fitted Bound    -4.7225
Predicted value           130.41     Fitted Value           130.41
Upper Predicted Bound     267.77     Upper Fitted Bound     265.53
SE (Predicted Value)      61.650     SE (Fitted Value)      60.646

Unusualness (Leverage)   29.9737
Percent Coverage            95.0
Corresponding T             2.23

Predictor Values: CHEM1 = 8.0000, CHEM3 = 80.000
```

The predicted value is 130.41 with a standard error of 61.650. The 95% prediction interval is -6.9581 to 267.77. This interval is expected to contain the value of a future single observation of the dependent variable HEAT 95% of the time, given that HEAT is observed at the independent variable values CHEM1 = 8, CHEM3 = 80.

The unusualness value tells you how “close” the specified independent data

point is to the rest of the data. If the point isn't close to the rest of the data, then you're extrapolating beyond your data—and prediction abilities may be very poor. An unusualness value greater than 1.0 should be considered large, so the prediction results in the example above are clearly suspect.

We illustrate the ***Case Method*** for specifying values for the independent variables below.



Here we simply enter a case number that refers to a case in your open *Statistix* file. In the example dialog box above, we've asked for predicted values for case number 7. The results are:

```
Predicted/Fitted Values of HEAT

Lower Predicted Bound     61.404     Lower Fitted Bound     78.644
Predicted value           87.692     Fitted Value           87.692
Upper Predicted Bound     113.98     Upper Fitted Bound     96.741
SE (Predicted Value)      11.798     SE (Fitted Value)      4.0609

Unusualness (Leverage)    0.1344
Percent Coverage            95.0
Corresponding T             2.23

Case number 7 was used to estimate the regression
Predictor Values: CHEM1 = 3.0000, CHEM3 = 17.000
```

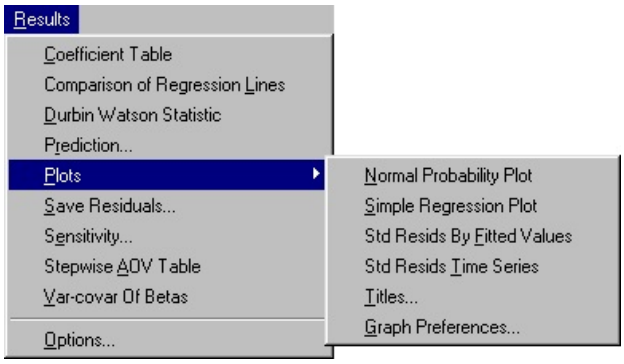
Unlike most other procedures, the prediction option lets you compute statistics for omitted cases. So you can divide the data into two subsets by omitting some cases. The regression model will be fitted using the selected

cases, and you can use this procedure to see how well the model fit cases that were not used to fit the model. Points used to fit the regression shouldn't be used to validate the regression model because least squares strives to optimize the goodness-of-fit at these points, so these points give an optimistic impression of the regression's performance.

For weighted regression models, the value for the weight variable at the case specified will be used as the weight in the prediction interval calculations. A weight w causes the prediction interval to be computed for a mean of w future observations.

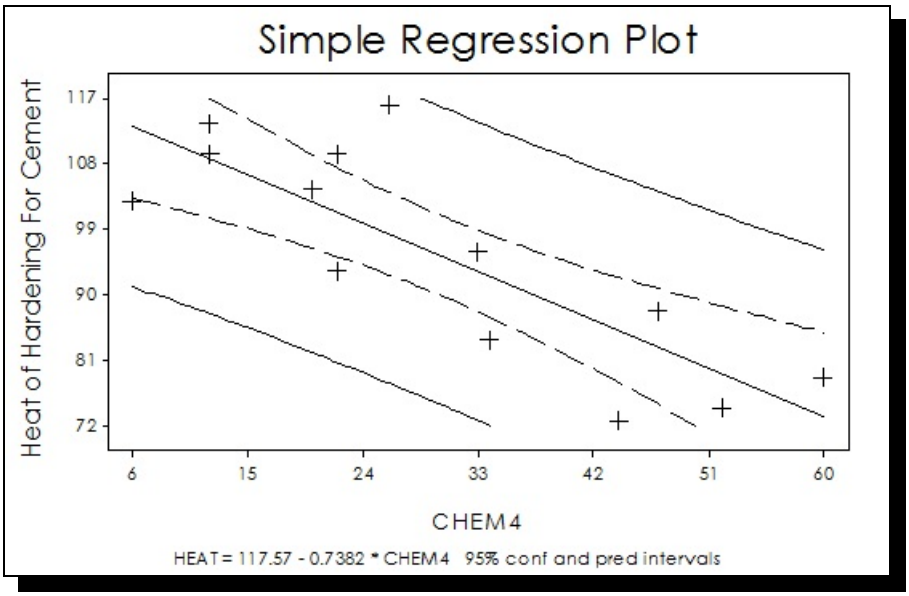
See the **Save Residuals** procedure discussed on page 176 for an alternative method of obtaining fitted values for cases in your data file.

Plots



Statistix offers four regression plots directly from the Results menu. You can also save fitted values and residuals (page 176) and use the **Scatter Plot** procedure described in Chapter 4 to graph additional residual plots.

The **Simple Regression Plot** is only available when there is one independent variable in the model. You can see an example of this on the next page. The observed values are displayed. The straight line in the center represents the fitted line. The inside curved lines mark the 95% confidence interval for the fitted line, the outside curves mark the 95% predicted interval.



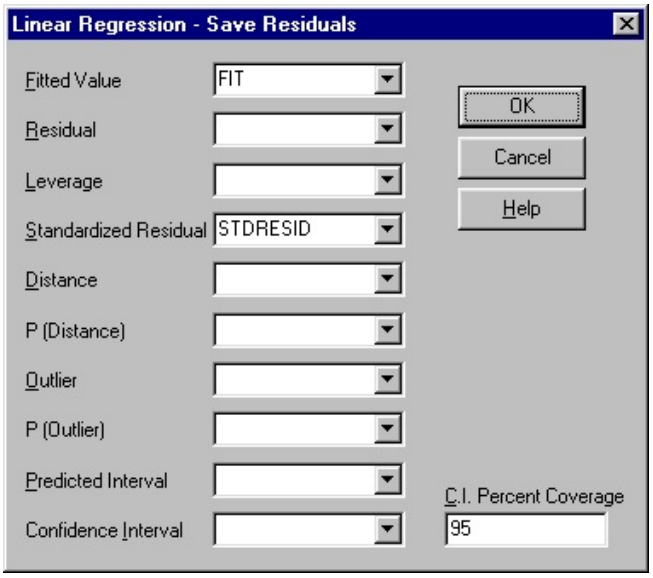
The **Std Resids by Fitted Values** plot is a scatter plot with the standardized residuals on the Y axis and the fitted values on the X axis. You use this plot to check the assumption of constant variances.

The **Std Resids Time Series** plot is a scatter plot with case numbers on the X axis. You'll find this plot useful when the data are sequentially ordered.

The **Normal Probablity Plot** is used to create a rankit plot of the standardized residuals to test the assumption of normality. The Shapiro-Wilk statistic for normality is also given (see Chapter 8).

Save Residuals

Statistix can do extensive residual analysis, which is very important for model evaluation. Systematic patterns in the residuals can suggest transformations or additional variables that would improve the model. Residuals are important for "outlier" analysis and to identify unusual observations. You can also use residuals to find those observations that are most influential in determining the values of the regression coefficient estimates. Space doesn't permit a detailed discussion here; if you're interested, consult Weisberg (1985).



Use the tab and shift-tab keys to move the cursor to the residual name you're interested in. Then type the name of a new or existing variable. You can click on the down-arrow button to display the list of current variables. You can change the value for percent coverage for predicted and confidence intervals by entering a new value in the ***C.I. Percent Coverage*** box.

When you've finished entering variable names, press the ***OK*** button to start the computations. The results are not displayed. The residuals are stored as variables and can be examined using such procedures as Scatter Plot, Shapiro-Wilk Test, Print, and Runs Test. Each of the residual options is described in more detail below.

The **Fitted Value** option saves the fitted (predicted) values for the dependent variable in a new variable. You can plot these values against the dependent variable to see how well the model fits the observed values.

The **Residual** option saves the raw residuals. You use these to examine how well the model fits the data. Systematic trends indicate that data transformations may be needed or that important independent variables were not included in the model. Large residuals may indicate observations that are "outliers". The raw residual is computed as the observed value of dependent variable minus the fitted value.

The **Leverage** option saves the leverage values in a new variable. Leverage

measures how influential a point is in determining the regression coefficients. Depending on the values of the independent variables, some data points are much more influential in determining the values of the regression coefficients than others. In general, the further the values of the independent variables are from their averages, the more influential these data points are in determining the regression coefficients. Points with "unusual" independent variable values have high leverage. If you're familiar with the matrix representation of multiple regression, leverage is calculated as the diagonal elements of the $X(X^TX)^{-1}X$ matrix (assuming no weights were used).

The **Standardized Residual** option saves the standardized residuals in a new variable. Raw residuals are very popular for examining model fit, although they do possess at least one potential problem. Data points located at "unusual" values of the independent variables will have high leverage, which means that they tend to "attract" the regression line and seldom have large residuals. Standardized residuals adjust for this problem to some extent by standardizing each raw residual with its standard error (also known as "studentizing"). The standard error of a residual is computed as the square root of the quantity MSE∗(1 - LEVERAGE), where MSE is the residual mean square for error.

The **Distance** option computes Cook's distance measure (Cook 1977, Weisberg 1985) and saves it in a new variable. It's very useful for finding influential data points, and it considers the effects of both the dependent and independent variables (LEVERAGE considers the independent variables only). Cook's distance for an observation can be thought of as a measure of how far the regression coefficients would shift if that data point was eliminated from the data set. Remember that it's a distance measure and not a test statistic. Usually, Cook's distance by itself isn't very useful because it is a function of sample size and the number of variables in the regression model. The next option, P (Distance), which "standardizes" Cook's D, is generally more useful.

P (Distance) calculates the "pseudo-significance level" of the confidence bound associated with the Cook's distance and saves it in a new variable. You can think of Cook's distance for an observation as a measure of how far the regression coefficients would shift if that data point were eliminated from the data set. It seems reasonable to standardize this distance by examining what level confidence bound the shifted regression coefficients would fall on. For example, if eliminating the I-th data point causes the new coefficients to shift to a position that corresponds to the edge of a 90%

confidence region around the original estimates, this may be cause for concern. Again, Cook's distance isn't a test statistic but rather a distance measure, and you shouldn't think of P (Distance) as an observed significance level (hence the term "pseudo-significance") but as a standardized distance.

The **Outlier** option computes the t-statistic for the outlier test and saves it in a new variable. You may suspect an observed value of the dependent data that deviates substantially from the predicted value of being an outlier. This is usually tested with a t-statistic, which is the result of this option. The computations follow Cook (1977).

P (Outlier) computes the p-value for the t-statistics resulting from the outlier test and saves it in a new variable. The computed p-value is appropriate for *a priori* tests for a single outlier. In other words, it's the appropriate observed significance level if you were interested in testing one particular case that is suspected of being an outlier before the data was observed. This is a rather unusual circumstance; more often, you notice potential outliers only after inspection of the residuals.

If you suspect a case of being an outlier after inspecting the data, you must give it special consideration. Suppose there were n cases. You now need to decide whether the value of the observed t-statistic is "unusual" given that it's the "most unusual" out of n repeated t tests. If the *a priori* p-values are used, then too many cases will be detected as "outliers", so you need to somehow "inflate" the *a priori* p-values. This inflation can be easily performed with Bonferroni's inequality (Weisberg 1985).

For example, suppose there were n cases inspected for outliers. Then, Bonferroni's equality says that the actual observed significance level will be no greater than nP, where P is the p-value returned by P (Outlier). In practice, the observed significance level may be substantially less than nP; Bonferroni's procedure has an undesirable property of becoming too conservative as n increases.

As noted earlier, you can compute Outlier and P (Outlier) for cases that were omitted and not used in the regression. Suppose m cases are used in the regression and m' are omitted. If you're interested in inspecting all m + m' cases for outliers, then you should use n = m + m'. If you're interested in just the omitted cases, then n = m'. The choice of n is your responsibility; *Statistix* always gives you the *a priori* p-value.

The **Prediction Interval** option computes the half-width confidence interval for predicted values. If the original regression model didn't include a weighting variable, the results that are saved are the half-widths of the prediction intervals for **single** future observations. Otherwise, the results are the half-widths of prediction intervals for means of w future observations, where w is the value of the weight variable. Use the ***C.I. Percent Coverage*** box to change the coverage level.

Suppose you save the prediction intervals in a variable called PI95, and the fitted values are saved in a variable called YHAT. Using **Transformations**, you can construct upper and lower prediction bounds as:

```
LOWER = YHAT - PI95
UPPER = YHAT + PI95
```

The **Confidence Interval** computes the half-width confidence interval for fitted values.

Treatment of Omitted Cases and Missing Y Values

Statistix returns values for certain residual menu options for cases that aren't used for estimation. Values are computed for predicted value, residual, leverage, outlier, and P (outlier) for omitted cases. A case that is not omitted may still not have been used for estimation if some of the values for that case are missing. If the only missing value for a case is the dependent variable, you can obtain values for predicted value and leverage.

These statistics derived from cases not used for estimation are especially useful for model validation. For omitted cases, predicted values are just what their name implies. For omitted cases, the residual option returns predicted residuals. Predicted residuals are, as you'd expect, the difference between the observed dependent variable value and predicted value (eq. 2.2.22 of Cook and Weisberg 1982). For cases that aren't used in the regression, leverage is somewhat unfortunate terminology. "Unusualness" or "generalized distance" would be a better term. For cases not used for estimation, this statistic measures how unusual the independent variable values of a case are relative to the values for cases used for estimation (see Weisberg 1985).

The interpretation of outlier is interesting and important. For cases used for estimation, outlier returns the t-statistic for the null hypothesis that the particular case conforms to the model fitted to the rest of the cases. An intuitive approach to constructing such a test is to first perform the regression without the particular case. From this regression, you can

compute the predicted residual for the particular case, say e. It's fairly easy to see that the standard error of e is simply the standard error of prediction, and the resulting statistic e/SE(e) should follow a t-distribution under the null hypothesis that E[e] = 0. This leads to an alternative interpretation of outlier as a **predicted Studentized residual**. Since the standard error used for Studentizing was derived from a data set that didn't include the particular point of interest, Cook and Weisberg (1982) refer to this as *external* Studentizing. Note that for the m cases used for estimation, the corresponding outlier values are computed from the m different data sets that did not include the case of interest. (The usual standardized residual is *internally* Studentized, which means the set of cases used to derive the standard error included the case for the residual.)

If there were m cases used in the original regression, it would appear that m separate regressions are needed to compute outlier just for the points used for estimation. However, as Cook and Weisberg (1982) show, there are some remarkable identities that permit the calculation of these statistics from quantities obtained from only the one regression using the m points. If the regression contains p parameters, the outlier statistic for a case used for estimation has m - p - 1 degrees of freedom associated with it.

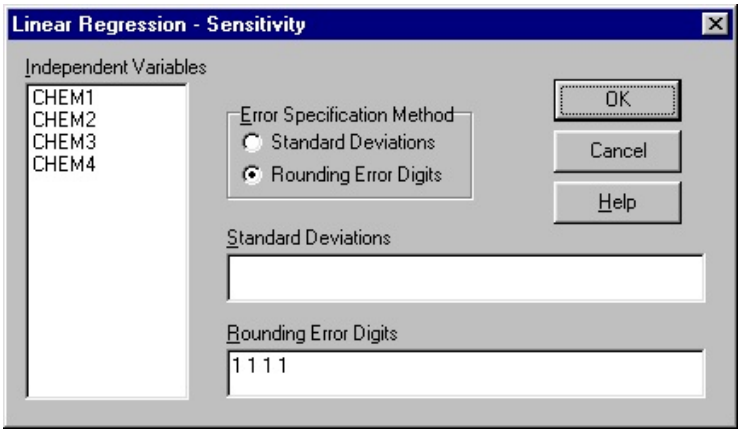
This now suggests how to handle cases not included in the regression. As before, the statistic is e/SE(e), where e is the predicted residual and SE(e) is the standard error for prediction. (The calculation of the standard error of prediction is described under Computational Notes.) For all cases not in the set of m cases used for estimation, outlier uses the same regression based on the m cases for the base comparison. (This is in contrast to the situation for Outlier for cases that were used in estimation; each such Outlier uses a different data set of m - 1 cases for base comparison.) This gains an extra case for computing the comparison regression for cases not used for estimation, so the Outlier statistic for a case not used for estimation has m - p degrees of freedom associated with it.

For cases not used for estimation, the usual standardized (internally Studentized) residuals can't be computed by definition. However, when you interpret them as predicted Studentized residuals, you can use outlier statistics for most of the purposes for which you'd want standardized residuals.

Sensitivity

Linear regression requires the assumption that the independent variables are measured without error. The **Sensitivity** procedure determines how

sensitive the estimated regression coefficients are to violations of this assumption. The user needs to specify the expected size of the errors associated with each of the independent variables, which can be done in one of two ways.



If you know the standard deviations of the errors, you can enter them directly using the ***Standard Deviations*** method. For example, if the values for an independent variable are the readings from some instrument, the standard error of the measures may be known from calibration results. However, precise information about the errors often won't be available. Sometimes the best you can do is to say, "I trust the figures down to—but not including— the d-th digit, which may be off by plus or minus 1". Often d will point to the digit at which rounding error occurred. In this case, you specify the position of the leftmost untrustworthy digit relative to the decimal point. For example, d = 2 for 9340.0 means you think the 4 may be in error. A way to model this is to assume that an additive random error is associated with 9340.0 and is uniformly distributed on the interval -5 to 5, which is exactly what *Statistix* does when treating ROUNDING error sensitivity. Negative values of d are used to indicate digits right of the decimal, for example, d = -4 points to the 8 in 0.03480. Values for d must be in the set of nonzero integers from 21 to -20.

We'll use the Hald data again to illustrate the procedure. Assume the regression was Y = CHEM1 CHEM2 CHEM3 CHEM4. Assume the figures are probably accurate except perhaps for rounding error in the least significant digit. *Statistix* assumes that this rounding error is uniformly distributed in the interval -0.5 to 0.5. The options are specified in the dialog above. The results are presented in the table on the next page.

Coefficients' Sensitivities to Errors in Independent Variables

	Variable in which Error Occurs			
	CHEM1	**CHEM2**	**CHEM3**	**CHEM4**
Error SD	**0.289**	**0.289**	**0.289**	**0.289**
Constant	0.469	0.554	0.558	0.575
CHEM1	0.829	0.943	0.923	0.962
CHEM2	0.372	0.452	0.463	0.474
CHEM3	-0.355	-0.250	-0.271	-0.230
CHEM4	-0.169	-8.75E-02	-7.89E-02	-6.74E-02

What is the interpretation of these results? When the independent variables represent continuous quantities, they'll always have some error associated with them. The question is whether the errors are large enough to seriously influence the results.

Suppose the independent variables were "perturbed" with errors of the specified size and the regression analysis rerun. We'd be happy if the slope estimates after perturbation were similar to those from before. The sensitivity coefficients indicate just how similar the new coefficients after perturbation are expected to be to the previous ones. More specifically, sensitivity coefficients give an index of how many significant figures will agree with the coefficients. For example, the estimates 0.89768 and 0.89735 agree to 3 figures, and this degree of agreement is expected when the sensitivity coefficient is near 3. Small sensitivity coefficients are undesirable. Of course "small" is subjective, but if the values are near or less than 1, such as in the example, the analysis clearly is very sensitive to errors in the independent variables and the regression coefficients can't be trusted.

It is interesting to compare these sensitivities with those for the model Y = CHEM1 CHEM4:

Coefficients' Sensitivities to Errors in Independent Variables

	Variable in which Error Occurs	
	CHEM1	**CHEM4**
Error SD	**0.289**	**0.289**
Constant	2.496	2.866
CHEM1	1.816	2.205
CHEM4	1.918	2.276

The sensitivity coefficients have increased considerably, which means the slope coefficients are now substantially less sensitive to the errors. This makes an interesting and very important point about this data set. CHEM1, CHEM2, CHEM3, and CHEM4 are nearly collinear; for all cases, the sums of these variables are between 95 and 98. When the data are nearly

collinear, the slope coefficients become very sensitive to minor changes in the independent variables. Many hours have been wasted by researchers trying to divine the interpretation of slope coefficients that were artifacts of the interaction of near-collinearity and errors in measurement. Reducing the degree of correlation in the set of independent variables has substantially reduced the sensitivities to the influence of roundoff errors.

If you're not certain how trustworthy your independent variables are, try some "worst-case scenarios". In other words, use large potential errors.

The sensitivity calculations are described in Weisberg (1982, 1985). We display the negative log (base ten) of the relative sensitivity coefficient, which in Weisberg's notation is $-\log_{10}(g_{jk})$. When you specify a placeholder d, the standard deviation of the error is calculated by assuming the error is a uniform random variable on the interval centered at 0 of width $10^{(n - sgn(n))}$ (sgn(n)=1 if n>0, sgn(n)=0 if n<0). The standard deviation is then $10^{(n-sgn(n))}/12^{1/2}$. If the absolute value of an estimated coefficient is too small (<1.0E-06) for reliable calculations, an M is displayed.

Stepwise AOV Table

This option produces a stepwise analysis of variance table for the specified regression model. The row order of the stepwise table reflects the order in which the independent variables are specified in the model. For the Hald example, the results are presented in the table below.

Stepwise Anlaysis of Variance of HEAT

Source	Individual SS	Cum DF	Cumulative SS	Cumulative MS	Adjusted R-Squared	Mallows' CP	P
Constant	118372						
CHEM1	1450.08	1	1450.08	1450.08	0.3009	202.5	2
CHEM2	1207.78	2	2657.86	1328.93	0.9680	2.7	3
CHEM3	9.79387	3	2667.65	889.217	0.9734	3.0	4
CHEM4	0.24697	4	2667.90	666.975	0.9736	5.0	5
Residual	47.8636	12	2715.76	226.314			

R-Squared	0.9824	Resid. Mean Square (MSE)	5.98295
Adjusted R-Squared	0.9736	Standard Deviation	2.44601

The table lists the individual contribution to the sums of squares, the cumulative mean squares, F test for the subset model versus the full model, and associated p-values, cumulative adjusted R^2, and Mallows' C_p statistic (see Miscellaneous Regression on the next page).

This table is useful for testing the contribution of subsets of the independent variables to the overall model. For example, you are interested in testing

whether CHEM3 and CHEM4 add anything to the model once CHEM1 and CHEM2 are already included. We first find the difference between the cumulative sums of squares for the model with all independent variables in it and one with just CHEM1 and CHEM2: 2667.90 - 2657.86 = 10.04. Because we're testing the contribution of two parameters, this sum of squares has two degrees of freedom associated with it and the resulting F test is F = (10.04/2)/5.983 = 0.839 (5.983 is the residual mean square of the full model). An F statistic this small suggests that CHEM3 and CHEM4 contribute little to the model once CHEM1 and CHEM2 are already included. However, this test says little about whether the model with just CHEM1 and CHEM2 is a "good" model.

Variance-Covariance of Betas

Select this option to obtain the variance-covariance matrix of the regression coefficient estimates. Once you've selected this option, the matrix is displayed.

Variance-Covariance Matrix for Coefficients

	Constant	CHEM1	CHEM4
Constant	4.51131		
CHEM1	-0.19254	0.01916	
CHEM4	-0.08332	0.00165	0.00237

The diagonal elements of the matrix are the variances of the regression coefficients. The off-diagonal values are the covariances; for example, 0.00165 is the covariance of the coefficient estimates of CHEM1 and CHEM4.

This matrix is most commonly used for constructing confidence regions about coefficient estimates, and for testing hypotheses about various functions of the coefficient estimates. More detail on these topics can be found in Weisberg (1985).

Miscellaneous Regression Topics

Mallows' C_p statistic, R^2, and adjusted R^2 are important criteria for evaluating and comparing regression models.

The **Mallows' C_p** statistic is useful for model selection. It's discussed in detail by Daniel and Wood (1971), Snedecor and Cochran (1980, p. 359) and Weisberg (1985). The C_p statistic is based on the fact that not including an important independent variable in the model results in the fitted response values being biased. C_p gives an index of this bias. "Good" models have C_p values near to or less than p, where p is the number of parameters in the

model. (Negative values will occasionally be observed). This statistic is most useful for eliminating variables that contribute little to the model. However, it tells you nothing about whether you started with all of the correct independent variables in the first place.

The **R^2** and **adjusted R^2** statistics measure the goodness of fit of a regression model. R^2 measures the proportion of variance in the dependent data explained by the regression. It's computed as 1 - RSS/SST, where RSS is the residual sum of squares and SST is the total sum of squares. A potential problem with R^2 is that it always increases as new independent variables are included in the model (RSS always decreases), even if they don't possess any relationship with the dependent variable. Adjusted R^2 is adjusted for the number of independent variables in the model to correct for this problem, and, therefore, will often be more interesting than the unadjusted R^2. Adjusted R^2 is computed as $1 - ((n - 1) / (n - p)) \times (1 - R^2)$, where n is the number of cases and p is the number of parameters in the regression. Adjusted R^2 is a monotonic function of the residual mean square. Unlike the unadjusted R^2, negative adjusted R^2's will occasionally be observed. Chatterjee and Price (1991), Draper and Smith (1966), and Weisberg (1985) are good references for more detail on these statistics.

For models without a constant, we use the total sums of squares adjusted for the mean to compute R^2 (Gordon, 1981). This can lead to negative values for R^2 and negative adjusted R^2 when a no constant model is a poor fit.

Best Model Selection

When there are a moderate number of independent variables, **Best Subsets Regressions** is a good way to select the best model. However, the number of subsets grows rapidly as the number of independent variables increases. If there are too many independent variables, use the **Stepwise Linear Regression** procedure. Draper and Smith (1966) present a good description of popular stepwise procedures. The potential problem with stepwise procedures is they do not necessarily result in a model that's best when judged by adjusted R^2 or Mallows' C_p. It's generally a good idea to try at least two stepwise procedures, such as backward elimination and forward inclusion, to see if they result in the same model.

Model selection is somewhat of a craft. Regression analysis is usually performed (1) to explore possible cause-effect relations, (2) to develop some predictive relationship, or (3) some combination of these. The relative importance of each of these goals may have some bearing on the model selection strategy used. Cause-effect modeling focuses on

determining the "important" independent variables. Predictive modeling focuses more on the development of a good predictor of the dependent variable than on the contribution due to any particular independent variable. Obviously, the distinction between cause-effect analysis and predictive modeling isn't a sharp one and most analyses include components of both. (As an aside, you should realize that regression analysis can't actually establish cause-effect relationships. It can examine the nature and extent of association between the dependent and potential independent variables. The interpretation of these associations as cause-effect is outside the realm of statistics, and lies in the domain of the appropriate subject-matter field.)

In addition to F tests, there are several other popular statistics for evaluating the "goodness" of a regression model. Some of these are adjusted R^2, Mallows' C_p statistic, R^2 , RSS (residual sum of squares) and RMS (residual mean square). Actually, R^2 and RSS are equivalent in the sense that they will produce the same orderings of the models. Adjusted R^2 and RMS are equivalent to one another in the same sense. It's generally best to use the adjusted R^2 (or RMS) and Mallows' C_p for model selection. R^2 is useful for comparing models with the same number of independent variables in them, but the adjusted R^2 will produce the same ordering of the models. The advantages of adjusted R^2 over the unadjusted R^2 are discussed in the previous section. There are lots of good references on the use of adjusted R^2 and Mallows' C_p in model selection, such as Weisberg (1985), Chatterjee and Price (1991), Daniel and Wood (1971), and Snedecor and Cochran (1980, p. 358).

The job isn't over when you've found the best model, as indicated by the adjusted R^2 or C_p statistic. Particularly in the case of cause-effect modeling, it will be of interest to examine whether all of the independent variables are significant (use the p-values from the coefficient table). No analysis is complete without an examination of the residuals. As a minimum, do the standardized residuals show any trend when plotted against the dependent variable? If they do, the model is not adequate.

The **Normal Probability Plot** is valuable for examining whether the assumption of normally distributed errors has been violated. If the errors are not normally distributed, the significance tests may be invalid. It's also a good idea to look at the correlations among the independent variables. High correlations may suggest problems with collinearity. If this is the case, the **Eigenvalues-Principal Components** procedure may be useful.

Computational Notes

The core computations are performed using Gentleman's square root free modification of Givens' method (Seber 1977). This is one of the most accurate methods available. An interesting feature of the method is that it exploits sparseness (zeros) in the independent variables to reduce computation time. It's therefore well suited for performing analyses of (co)variance that involve indicator variables.

The error variance is estimated as σ^{*2} = RMS, where RMS is the residual mean square from the regression. Let X be the design matrix containing all the cases used to find the coefficient estimates B*. The estimated variance of B* is $V(B^*)= \sigma^{*2} (X^TUX)^{-1}$, where U is the weight matrix if weights were used, and U = I otherwise. Suppose x^T is a specific row of X, a case used for estimation. The fitted value f corresponding to case x is $f = x^TB^*$. The estimated variance of a fitted value is then $V(f) = \sigma^{*2} x^T(X^TUX)^{-1}x$.

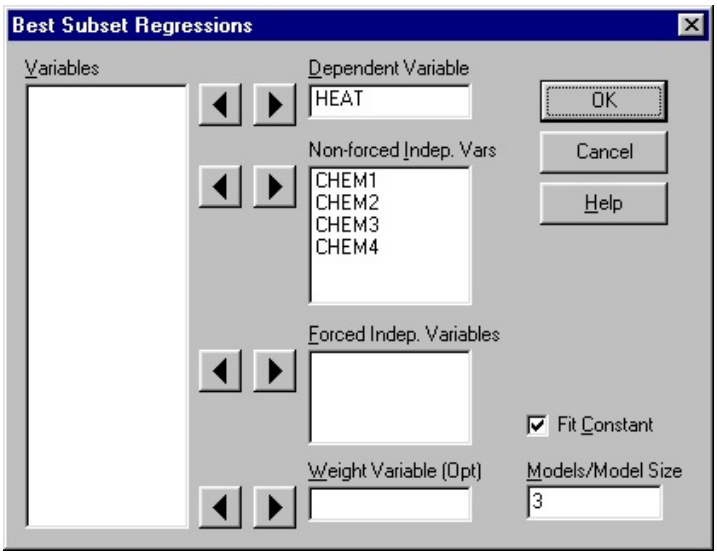
Now assume x wasn't used for estimation. The corresponding predicted value is $p = x^TB^*$. If a weight is not specified, the prediction is for a single future observation, which has estimated variance $V(p) = \sigma^{*2} + V(f)$. If there is a weight w, then the prediction is for a mean of w future observations, which has estimated variance $V(p) = \sigma^{*2}/w + V(f)$. Let SE(f) and SE(p) be the square roots of V(f) and V(p), respectively. Then, the confidence interval for the fitted value f and the prediction interval for the predicted value p are given respectively as $f \pm SE(f)\ t$ and $p \pm SE(p)\ t$, where t is the appropriate t value for the specified coverage.

If x is used for estimation, leverage is computed as $u\ x^T(X^TUX)^{-1}x$, where u is the scalar weight associated with case x. If x wasn't used for estimation, the "leverage" (unusualness) is $x^T(X^TUX)^{-1}x$. Let y be an observed value of the dependent variable (it may or may not have been used for estimation). A residual is computed as y - f, and a predicted residual is computed as y - p. An outlier t value for a case not used for estimation is (y - p) / SE(p).

Best Subset Regressions

The **Best Subset Regressions** procedure in *Statistix* computes the best subset regression models given a full model that contains all the potential predictor variables of interest. A specified number of subset models with the highest R^2 are listed for each model size.

Specification



First select the name of the dependent variable (response variable). Highlight a variable in the ***Variables*** box, then press the right-arrow button next to the ***Dependent Variable*** box to move the highlighted variable into that box.

Move candidate independent variables to the ***Non-forced Indep. Variables*** list box. These variables will be used in all possible combinations to form the subset regressions.

You can select one or more variables to be forced in the regression models. Forced variables will appear in all of the subset models. Listing some independent variables known to be important can greatly reduce the number of subset models computed. Move the variables you want forced in all models from the ***Variables*** list to the ***Forced Indep. Variables*** list.

To perform weighted least squares regression, select the name of the

variable containing the weights and move it to the ***Weight Variable*** box.

Use the ***Fit Constant*** check box to specify a model with a constant fitted (checked) or a model forced through the origin (not checked).

Enter the number of best candidate models you want listed in the results for each subset model size in the ***Models/Model Size*** edit control. You can specify as many as ten models, but this will be reduced for very large full models to limit the total number of subset models listed to 150.

Data Restrictions

Up to a total of 50 forced and unforced independent variables can be included in the model (15 is a more practical limit for the number of unforced variables because of computation time). If any values are missing for a case in the full model, the entire case is ignored (listwise deletion) for all models. If an independent variable is too highly correlated with a linear combination of other independent variables in the full model (collinearity), it's dropped from the model. Computation is reinitiated with a new full model in which the offending independent variable has been dropped. If collinearity still exists, another variable will be dropped. Variables are dropped until such collinearity has been eliminated and reliable computations can proceed. If weighted regression is specified, the variable used for the weights cannot have negative values. Zero weights are treated as missing values.

Example

Our example data are the Hald data from Draper and Smith (1966). The variable HEAT is the cumulative heat of hardening for cement after 180 days. The variables CHEM1, CHEM2, CHEM3, and CHEM4 are the percentages of four chemical compounds measured in batches of cement. The data are listed below.

CASE	HEAT	CHEM1	CHEM2	CHEM3	CHEM4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

The goal is to relate the heat of hardening to the chemical composition. The analysis is specified in the dialog box on the preceding page. The results are presented in the table below.

Best Subset Regression Models for HEAT Cumulative Heat of Hardening For Cement

Unforced Independent Variables: (A)CHEM1 (B)CHEM2 (C)CHEM3 (D)CHEM4
3 "best" models from each subset size listed.

P	CP	Adjusted R Square	R Square	Resid SS	Model Variables
1	442.9	0.0000	0.0000	2715.76	Intercept Only
2	138.7	0.6450	0.6745	883.867	D
2	142.5	0.6359	0.6663	906.336	B
2	202.5	0.4916	0.5339	1265.69	A
3	2.7	0.9744	0.9787	57.9045	A B
3	5.5	0.9670	0.9725	74.7621	A D
3	22.4	0.9223	0.9353	175.738	C D
4	3.0	0.9764	0.9823	47.9727	A B D
4	3.0	0.9764	0.9823	48.1106	A B C
4	3.5	0.9750	0.9813	50.8361	A C D
5	5.0	0.9736	0.9824	47.8636	A B C D

Cases Included 13 Missing Cases 0

Mallows' C_p statistic, unadjusted and adjusted R^2, and residual sums of squares are produced for each model. More detail on these statistics is given in **Linear Regression** under **Miscellaneous Regression Topics** (page 185).

Note: The number of possible subset models grows rapidly as the number of independent variables is increased. If there are M independent variables, then there are 2^M - 1 subset models. For example, with 10 independent variables there are 1023 subset models, while with 15 independent variables there are 32,767 subset models.

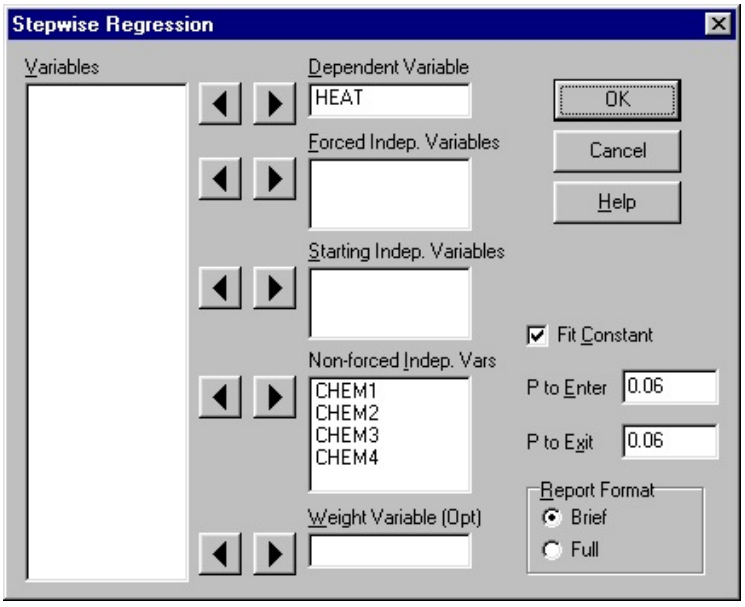
Computational Notes

The method used to generate subset statistics is patterned after Clarke (1981). An advantage of this method is that it's very accurate without requiring extended precision computing. The first regression based on the full model is performed with the method described in **Linear Regression**.

Stepwise Linear Regression

The **Stepwise Linear Regression** procedure in *Statistix* performs stepwise linear regression. You can specify an empty initial model (forward selection), a full initial model (backward elimination), or any initial model in between. Stepwise procedures are popular methods of searching for good subset models, particularly when the number of independent models is large. (See also **Best Subset Regressions** on page 189.)

Specification



First select the name of the dependent variable (response variable). Highlight a variable in the ***Variables*** box, then press the right-arrow button next to the ***Dependent Variable*** box to move the highlighted variable into that box.

Divide your independent variables between the list boxes for ***Forced***, ***Starting***, and ***Non-forced Indep. Vars***. Forced variables will appear in all steps of the stepwise procedure and will not be eliminated regardless of the elimination criteria. The starting independent variables will appear in the initial model and may be eliminated in subsequent steps. The non-forced independent variables don't appear in the initial model but will be considered for selection.

For forward selection, move all your independent variables to the ***Non-forced Indep. Vars*** box. For backwards selection, move all the independent variables to the ***Starting Indep. Vars*** box. You can also use the Starting Indep. Vars box to enter an initial model that you have previously found of interest or an initial model that includes a variable that was overlooked in a previous stepwise regression.

A stepwise regression builds a regression model by repeating a process that adds and deletes variables from a list of candidate variables. The stepwise process stops when no variables not already in the model meet the selection criterion and no variables in the model meet the elimination criterion.

At each step in the process, the variable with the lowest p-value is selected to enter the model next. A variable will not be selected unless its p-value is less than the value you enter for the ***P To Enter*** criterion. A variable's p-value tests the hypothesis that the variables' regression coefficient is zero. You can specify pure backward elimination by entering 0.0 for the P To Enter criterion to prevent eliminated variables from reentering the model.

The variable with the highest p-value is eliminated from the model at each step. A variable won't be eliminated unless its p-value is greater than the value you enter for the ***P To Exit*** criterion. You can specify pure forward selection by entering 1.0 for the P To Exit criterion to prevent selected variables from being eliminated later.

The remaining options let you choose between brief and full report formats, select a weight variable for weighted regression, and specify a model forced through the origin.

Data Restrictions

Up to 50 independent variables can be included in the model. If any values are missing for a case in the full model, the entire case is ignored (listwise deletion) for all models. If weighted regression is specified, the variable used for the weights can't have negative values. Zero weights are treated as missing values. Variables won't be selected that are found to be too highly correlated with variables already in the model (collinearity).

Example

The data used in our example is the Hald data set from Draper and Smith (1966). The variable HEAT is the cumulative heat of hardening for cement after 180 days. The variables CHEM1, CHEM2, CHEM3, and CHEM4 are the percentages of four chemical compounds measured in batches of

cement. The data are listed below.

CASE	HEAT	CHEM1	CHEM2	CHEM3	CHEM4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

The goal is to relate the heat of hardening to the chemical composition. The analysis is specified in the dialog box on page 192. The results are as follows:

```
Stepwise Linear Regression of HEAT
Unforced Variables: CHEM1 CHEM2 CHEM3 CHEM4
  P to Enter 0.0600
  P to Exit  0.0600
                                        C C C C
                                        H H H H
                                        E E E E
                                        M M M M
Step     R Sq          MSE         P    1 2 3 4
   1   0.0000      226.314               . . . .
   2   0.6745      80.3515   0.0006 +    . . . D
   3   0.9725      7.47621   0.0000 +    A . . D
   4   0.9823      5.33030   0.0517 +    A B . D
   5   0.9787      5.79045   0.2054 -    A B . .

Resulting Stepwise Model
Variable    Coefficient   Std Error         T         P      VIF
Constant        52.5773     2.28617     23.00    0.0000
CHEM1           1.46831     0.12130     12.10    0.0000      1.1
CHEM2           0.66225     0.04585     14.44    0.0000      1.1

Cases Included   13       R Squared  0.9787       MSE   5.79045
Missing Cases     0       Adj R SQ   0.9744       SD    2.40634

Variables Not in the Model
              Correlations
Variable   Multiple   Partial         T         P
CHEM3        0.8257    0.4113      1.35    0.2089
CHEM4        0.9732   -0.4141     -1.37    0.2054
```

The first part of the report is a history of the stepwise process. It lists the variables in the model for each step and presents a number of model statistics. In the example above, the intercept-only model is listed as step 1. CHEM4 is added at step 2, CHEM1 is added at step 3, CHEM2 is added at step 4, and then CHEM4 is eliminated at step 5. The R^2 and mean square error (MSE) are listed for each step. The number in the P column at each step is the p-value for the selected variable (+) or the eliminated variable (-).

At the end of the stepwise history, a complete coefficient table and model summary statistics are presented for the final model.

The final table in the report lists the variables not in the final model. It lists the multiple and partial correlations of each variable with the final model. The T and P columns list the t value and p-value for each variable were it to be added to the final model.

The full report lists a complete coefficient table for each step. The example below displays a full format report for the backward elimination stepwise regression for the Hald data.

Stepwise Linear Regression of HEAT
Unforced Variables: CHEM1 CHEM2 CHEM3 CHEM4
P to Enter 0.0500
P to Exit 0.0500

Step	Variable	Coefficient	T	P	R SQ	MSE
1	Constant	62.4054	0.89		0.9824	5.98295
	CHEM1	1.55110	2.08	0.0708		
	CHEM2	0.51017	0.70	0.5009		
	CHEM3	0.10191	0.14	0.8959		
	CHEM4	-0.14406	-0.20	0.8441		
2	Constant	71.6483	5.07		0.9823	5.33030
	CHEM1	1.45194	12.41	0.0000		
	CHEM2	0.41611	2.24	0.0517		
	CHEM4	-0.23654	-1.37	0.2054		
3	Constant	52.5773	23.00		0.9787	5.79045
	CHEM1	1.46831	12.10	0.0000		
	CHEM2	0.66225	14.44	0.0000		

Resulting Stepwise Model

Variable	Coefficient	Std Error	T	P	VIF
Constant	52.5773	2.28617	23.00	0.0000	
CHEM1	1.46831	0.12130	12.10	0.0000	1.1
CHEM2	0.66225	0.04585	14.44	0.0000	1.1

Cases Included	13	R Squared	0.9787	MSE	5.79045
Missing Cases	0	Adj R SQ	0.9744	SD	2.40634

Variables Not in the Model

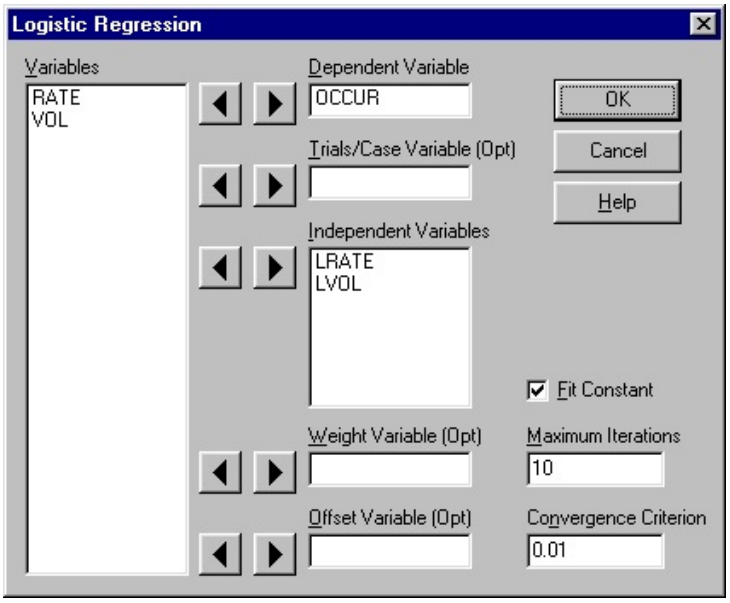
	Correlations			
Variable	Multiple	Partial	T	P
CHEM3	0.8257	0.4113	1.35	0.2089
CHEM4	0.9732	-0.4141	-1.37	0.2054

Logistic Regression

The **Logistic Regression** procedure is used when you are interested in studying how observed proportions or rates depend on particular independent variables. A direct application of linear regression to proportions is often not satisfactory because the fitted or predicted values may be less than 0 or greater than 1 (impossibilities for proportions). There may be other shortcomings as well. Logistic regression provides a convenient alternative by examining the relationships between the logistic transformation of the proportions and linear combinations of the predictor (independent) variables. The estimation method is maximum likelihood. Numerous model diagnostic options are available.

More background on logistic regression can be found in the section titled Additional Background on Logistic and Poisson Regression on page 211. In particular, you should be familiar with likelihood ratio tests (also known as analysis of deviance tests or G^2 tests) to make full use of this procedure.

Specification



Select the name of the dependent variable in the ***Variables*** list box and move it to the ***Dependent Variable*** box.

If each case in your data represents a single observation, the dependent

variable contains only zeros and ones, and the ***Trials/Case Variable*** is not used. If some or all of the cases in your data represent more than a single observation, then the dependent variable contains the sum of the zeros and ones for all trials for each case, and the Trials/Case Variable stores the total number of trials for each case. When a Trials/Case Variable is used, the logistic regression is computed on the ratio of the Dependent Variable to the Trials/Case Variable. This feature is used when indicator variables are used for the independent variables and trials can be grouped by unique combinations of values for the independent variables.

Move the independent variables from the Variables list box to the ***Independent Variables*** box. For weighted regression, move the variable containing the prior case weights to the ***Weight Variable*** box.

In some circumstances, the regression coefficient for a term in the model is known beforehand. Such a term is called an offset and can be "adjusted out" of the model. The ***Offset Variable*** is subtracted from the linear predictor, so the offset variable must be expressed on the linear predictor's scale (logit scale).

Use the ***Fit Constant*** check box to have the constant fitted in the model (checked) or have the model forced through the origin (not checked).

Logistic regression uses an iterative procedure (iterative reweighted least squares) to obtain its maximum likelihood results. You specify the maximum number of iterations performed before the procedure "gives up" if it hasn't converged in the ***Maximum Iterations*** edit control.

Iteration stops when the absolute change in the deviance between iterations is less than the deviance convergence criterion you specify in the ***Convergence Criterion*** edit control. Small values increase the estimation accuracy but may increase the number of required iterations. The value of 0.01 is usually suitable for obtaining deviances and coefficient estimates.

Data Restrictions

Up to 50 independent variables can be included in the model. If any values within a case are missing, the case is dropped (listwise deletion). If an independent variable is too highly correlated with a linear combination of other independent variables in the model (collinearity), it's dropped from the model. Computation is reinitiated with a new model in which the offending independent variable has been dropped. Variables are dropped until such collinearity has been eliminated and reliable computations can

proceed. For each case, the ratio of the dependent variable to the number of trials (success/trials) must always be bounded by 0 to 1. If weighted regression is specified, the weight variable cannot contain negative weights. Zero weights are treated as missing values.

Example

The logit transformation is ln(p/(1-p)), where p is a proportion. The ratio p/(1-p) is often interpreted as "odds"; for example, if p is the probability of success, then 1-p is the probability of failure and p/(1-p) is the odds for success. By relating ln(p/(1-p)) to a linear combination of predictors, we are assuming that the predictors act in a multiplicative fashion to influence the odds p/(1-p). Remember that logistic regression is relating the linear combination of predictors to ln(p/(1-p)) and not to p, as the analysis specification may suggest.

Finney's data from Pregibon (1981) are used for this example. The response was whether or not vasoconstriction occurred in the skin of the digits; variable OCCUR takes the values 1 or 0. There are two quantitative predictor variables—the rate and volume of air inspired by the subject. For analysis, log rate and log volume were used, variables LRATE and LVOL, respectively. The data are displayed below, and is available in the file Sample Data\ vasoconstriction.sx.

CASE	OCCUR	LRATE	LVOL	CASE	OCCUR	LRATE	LVOL
1	1	-0.1924	1.3083	21	0	0.6931	-0.9163
2	1	0.0862	1.2528	22	0	0.3075	-0.0513
3	1	0.9163	0.2231	23	0	0.3001	0.3001
4	1	0.4055	-0.2877	24	0	0.3075	0.4055
5	1	1.1632	-0.2231	25	1	0.5766	0.4700
6	1	1.2528	-0.3567	26	0	0.4055	-0.5108
7	0	-0.2877	-0.5108	27	1	0.4055	0.5878
8	0	0.5306	0.0953	28	0	0.6419	-0.0513
9	0	-0.2877	-0.1054	29	1	-0.0513	0.6419
10	0	-0.7985	-0.1054	30	0	-0.9163	0.4700
11	0	-0.5621	-0.2231	31	1	-0.2877	0.9933
12	0	1.0116	-0.5978	32	0	-3.5066	0.8544
13	0	1.0986	-0.5108	33	0	0.6043	0.0953
14	1	0.8459	0.3365	34	1	0.7885	0.0953
15	1	1.3218	-0.2877	35	1	0.6931	0.1823
16	1	0.4947	0.8329	36	1	1.2030	-0.2231
17	1	0.4700	1.1632	37	0	0.6419	-0.0513
18	1	0.3471	-0.1625	38	0	0.6419	-0.2877
19	0	0.0583	0.5306	39	1	0.4855	0.2624
20	1	0.5878	0.5878				

The data are said to be ungrouped; each case is based on a single trial, so the Trials/Case Variable is left empty. The logistic regression is specified on page 196.

Convergence is reached after the fifth iteration, and the coefficient table is

displayed as follows:

```
Unweighted Logistic Regression of OCCUR

Predictor
Variables   Coefficient   Std Error   Coef/SE         P
Constant       -2.87494     1.30650     -2.20    0.0278
LRATE           4.56100     1.81790      2.51    0.0121
LVOL            5.17862     1.84314      2.81    0.0050

Deviance               29.23
P-Value               0.7807
Degrees of Freedom        36

Convergence criterion of 0.01 met after 5 iterations

Cases Included 39    Missing Cases 0
```

The p-value of 0.7807 suggests this model fits the data fairly well.

Maybe it's not necessary to include both terms LVOL and LRATE in the model. When we run the analysis using only LRATE, we find that the deviance is 48.86 (p=0.0918). When LVOL is specified alone, the resulting deviance is 47.06 (p=0.1243). When we run the analysis with no independent variables, the resulting deviance is 54.04 (p=0.0440).

The following analysis of deviance table summarizes the results (I = intercept, R = log rate, V = log volume):

Model	Deviance	Difference	DF	Component	P-Value
I	d_I =54.04	d_I-d_{IVR} =24.81	2	V and R	0.000
I+R	d_{IR} =48.86	dI_R-d_{IRV}=19.63	1	V	0.000
I+V	d_{IV} =47.06	d_{IV}-d_{IRV}=17.83	1	R	0.000
I+R+V	d_{IRV}=29.23				

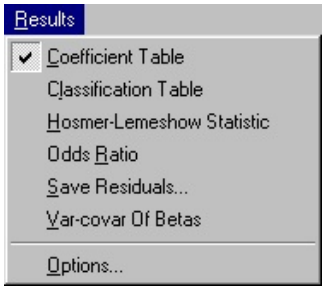
The column labeled Component shows which terms are being tested. The first row tests whether LVOL and LRATE improve the intercept-only model. The deviance for this test is the difference of the deviances for model I and model I+R+V, which is 54.04 - 29.23 = 24.81. The associated value for the degrees of freedom is the difference in the number of independent variables in the two models. The p-value displayed in the last column is computed using the chi-square function in **Probability Functions** (see Chapter 13). The second row tests whether LVOL improves the model when I and LRATE are already in the model. The deviance for this test is the difference of the deviances for model I+R and model I+R+V, which is 48.86 - 29.23 = 19.63. The LRATE term is tested in a similar manner in the

third row. The conclusion from this analysis of deviance table is that both LVOL and LRATE are needed in the model.

Before accepting this model, you should examine the regression diagnostics. You should examine the standardized residuals and Cook's D (or p-value) routinely.

Logistic Regression Results Menu

Once the regression analysis is computed and displayed, a ***Results*** pull-down menu appears on the menu at the top of the *Statistix* window. Click on the Results menu to display the regression results menu displayed below.



Select Coefficient Table from the menu to redisplay the regression coefficient table. Select Options to return to the main dialog box used to specify the model. Like the Linear Regression procedure, Logistic regression offers the options of saving various residual- and model-diagnostic statistics (see page 176) and examining the variance-covariance matrix of the regression coefficients (see page 185). The remaining options are discussed below.

Classification Table

The classification table is a 2x2 frequency table of actual and predicted responses: The fitted logistic regression model is used to obtain the estimated value of p for each case. According as the estimated value of p is less that 0.5 or greater than 0.5, the case is placed in the category 0 or 1. By then cumulating over cases, the classification table is obtained. The classification table for Finney's vasoconstriction data is shown on the next page.

Classification Table for OCCUR

Actual	Predictions 0	1	Total
0	14	5	19
1	2	18	20
TOTAL	16	23	39

Proportion of category 0 correctly classified	0.737
Proportion of category 1 correctly classified	0.900
Overall proportion correctly classified	0.821

Hosmer-Lemeshow Statistic

The Hosmer-Lemeshow statistics are goodness-of-fit tests suitable for models with a large number of covariate patterns (unique combinations of values for the independent variables). These tests are illustrated here using data from Hosmer and Lemeshow (1989, p. 92), a study whose object was to identify risk factors for low birth weights of babies (the data are available in the file Sample Data\birthwt.sx). The response variable is LOW: 0 = normal birth weight, 1 = low birth weight. The table below shows the tests for the model found on page 101, Hosmer and Lemeshow.

Hosmer-Lemeshow Goodness of Fit Tests for LOW

		Decile of Risk 1	2	3	4	5	6	7	8	9	10	
LOW		0.07	0.09	0.16	0.21	0.28	0.33	0.42	0.49	0.61	0.94	Total
1	Obs	0	1	4	2	6	6	6	10	9	15	59
	Exp	0.9	1.6	2.4	3.5	5.0	5.6	6.8	8.6	10.5	14.1	59
0	Obs	19	18	15	17	14	12	12	9	10	4	130
	Exp	18.1	17.4	16.6	15.5	15.0	12.4	11.2	10.4	8.5	4.9	130
Total		19	19	19	19	20	18	18	19	19	19	189

Hosmer-Lemeshow Statistic (C)	4.78
P-Value	0.7804
Degrees of Freedom	8

		Fixed Cut Points 0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	
LOW												Total
1	Obs	2	5	8	9	11	9	7	5	2	1	59
	Exp	2.7	4.5	8.3	8.7	10.7	8.8	6.5	5.3	1.7	1.9	59
0	Obs	38	25	25	16	13	7	3	2	0	1	130
	Exp	37.3	25.5	24.7	16.3	13.3	7.2	3.5	1.7	0.3	0.1	130
Total		40	30	33	25	24	16	10	7	2	2	189

Hosmer-Lemeshow Statistic (H)	6.17
P-Value	0.6281
Degrees of Freedom	8

The tables are constructed by grouping the observations into ten groups based on the values of the fitted values (estimated probabilities). Two grouping methods are used as follows:

The first method groups the data based on percentiles of the fitted values, resulting in a table with ten “deciles of risk”. The test statistic C = 4.78 is computed from the observed and expected frequencies within each decile of risk for each outcome. The p-value 0.7804 is computed from the chi-square distribution with 8 degrees of freedom.

The values above each of the columns in the deciles-of-risk table represent the highest fitted value for the column. Observations with the same covariate pattern are forced into the same decile, which can result in some columns with zero observed frequencies. In these cases, the statistic isn’t computed.

The second grouping method bases the groups on fixed cut points of the fitted values. Both the C and H statistics indicate a good fit. See Hosmer and Lemeshow (1989) for a detailed discussion.

Odds Ratios

Select Odds Ratios from the results menu to obtain odds ratios and 95% confidence intervals. The table of odds ratios from the logistic regression of the birth weight data from Hosmer and Lemeshow (1989, p. 94) is presented below.

Logistic Regression Odds Ratios for LOW

Predictor Variables	95% C.I. Lower Limit	Odds Ratio	95% C.I. Upper Limit
AGE	0.91	0.97	1.05
LWT	0.97	0.98	1.00
RACE1	1.26	3.54	9.91
RACE2	1.00	2.37	5.59
SMOKE	1.15	2.52	5.51
PTL	0.87	1.72	3.39
HT	1.62	6.26	24.23
UI	0.87	2.14	5.25

The odds ratios reported in *Statistix* give the change in the odds for an increase in one unit of the independent variable. For the dichotomous variable SMOKE, the odds of a low birth weight baby are 2.52 greater for a mother who smokes (SMOKE = 1) than for a mother who doesn’t smoke (SMOKE = 0). The odds ratio of 0.97 for the continuous variable AGE is for an increase of one year in age.

General Notes

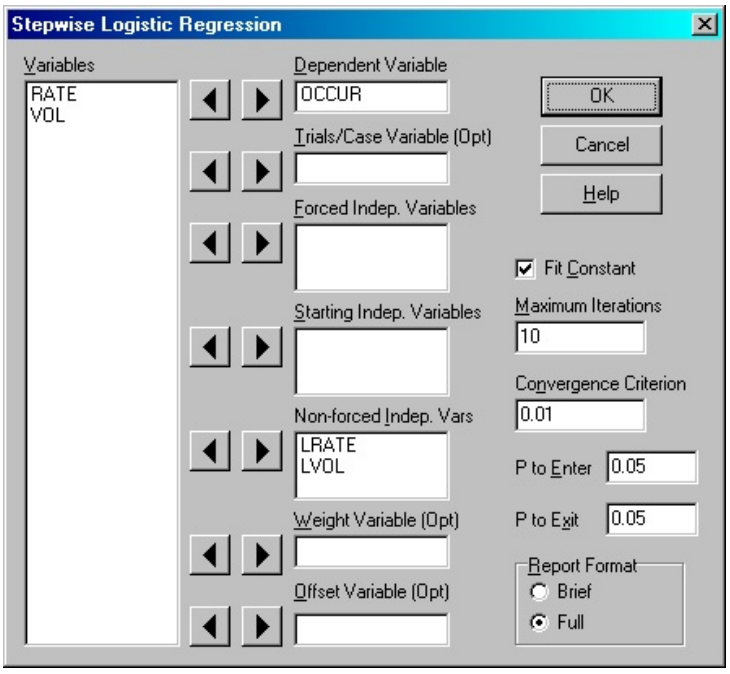
If fitted values are saved, the resulting values are the estimates of expected number of successes E[N]. If you want estimates of p, first save the fitted values and then divide them by the total number of trials per case. Estimates for the logits ln(p/(1-p)) can be obtained from the estimated p’s.

Probit regression is a method very similar to logistic regression. *Statistix* does not perform probit regression. For the vast majority of data sets, logistic and probit analyses will return virtually identical results. We prefer logistic regression because it can be calculated more efficiently and the logit transform has a simple interpretation as the log of the odds ratio.

Stepwise Logistic Regression

This procedure performs stepwise logistic regression. Logistic regression is appropriate for dependent variables that are proportions. Both forward selection and backward elimination are supported.

Specification



Select the name of the dependent variable in the ***Variables*** list box and move it to the ***Dependent Variable*** box.

If each case in your data represents a single observation, the dependent

variable contains only zeros and ones, and the ***Trials/Case Variable*** is not used. If some or all of the cases in your data represent more than a single observation, then the dependent variable contains the sum of the zeros and ones for all trials for each case, and the Trials/Case Variable stores the total number of trials for each case.

Divide your independent variables between the list boxes for ***Forced***, ***Starting***, and ***Non-forced Indep. Vars***. Forced variables will appear in all steps of the stepwise procedure and will not be eliminated regardless of the elimination criteria. The starting independent variables will appear in the initial model and may be eliminated in subsequent steps. The non-forced independent variables don't appear in the initial model but will be considered for selection.

For forward selection, move all your independent variables to the ***Non-forced Indep. Vars*** box. For backwards selection, move all the independent variables to the ***Starting Indep. Vars*** box. You can also use the Starting Indep. Vars box to enter an initial model that you have previously found of interest or an initial model that includes a variable that was overlooked in a previous stepwise regression.

A stepwise regression builds a regression model by repeating a process that adds and deletes variables from a list of candidate variables. The stepwise process stops when no variables not already in the model meet the selection criterion and no variables in the model meet the elimination criterion.

At each step in the process, the variable whose addition decreases the deviance the greatest is selected to enter the model next. A variable will not be selected unless the p-value for the deviance test (the difference of the deviance for the model excluding the candidate variable and the deviance of the model including the candidate variable) is less than the value you enter for the ***P To Enter*** criterion. You can specify pure backward elimination by entering 0.0 for the P To Enter criterion to prevent eliminated variables from reentering the model.

The variable whose removal from the model increases the deviance the least is eliminated from the model at each step. A variable won't be eliminated unless the p-value for the deviance test is greater than the value you enter for the ***P To Exit*** criterion. You can specify pure forward selection by entering 1.0 for the P To Exit criterion to prevent selected variables from being eliminated later.

In some circumstances, the regression coefficient for a term in the model is known beforehand. Such a term is called an offset and can be "adjusted out" of the model. The ***Offset Variable*** is subtracted from the linear predictor, so the offset variable must be expressed on the linear predictor's scale (logit scale).

The remaining options let you choose between brief and full report formats, select a weight variable for weighted regression, specify a model forced through the origin, specify the maximum number of iterations, and enter a value for the convergence criterion.

Data Restrictions

A total of 50 independent variables can be included in the model. If any values within a case are missing, the case is dropped (listwise deletion). For each case, the ratio of the dependent variable to the number of trials (success/trials) must always be bounded by 0 to 1.

Example

We'll use Finney's vasoconstriction data described on page 198 to illustrate forward selection. The dependent variable, OCCUR, is coded 1 if vasoconstriction occurred in the skin and 0 if not. There are two candidate independent variables, LRATE and LVOL. The analysis is specified on page 203. The results are displayed below.

```
Stepwise Logistic Regression of OCCUR
Unforced Variables: LRATE LVOL
  P to Enter 0.0500
  P to Exit  0.0500

Step  Variable   Coefficient  Std Error  Coef/SE  Deviance  Difference      P
   1  Constant       0.05129    0.32036     0.16     54.04
   2  Constant      -0.20458    0.36232    -0.56     47.06       6.98   0.0082
      LVOL           1.80788    0.77024     2.35
   3  Constant      -2.87494    1.30650    -2.20     29.23      17.83   0.0000
      LVOL           5.17862    1.84314     2.81
      LRATE          4.56100    1.81790     2.51


Resulting Stepwise Model
Variable    Coefficient   Std Error   Coef/SE         P
Constant       -2.87494     1.30650     -2.20    0.0278
LVOL            5.17862     1.84314      2.81    0.0050
LRATE           4.56100     1.81790      2.51    0.0121

Deviance               29.23
P-Value               0.7807
Degrees of Freedom        36

Cases Included 39    Missing Cases 0
```

Three steps were made in arriving at the final model. Because no starting variables were specified, the model in the first step was the intercept-only model. The variable LVOL was added in step 2. The difference of the deviances for the models in step 1 and 2 is 6.98 (p=0.0082). The variable LRATE was added in step 3. The complete coefficient table and model statistics for the final model are listed at the end of the report.

Computational Notes

See Hosmer and Lemeshow (1989) for a description of the stepwise algorithm and the deviance test. More background on logistic regression can be found in the section titled Additional Background on Logistic and Poisson Regression on page 211.

Poisson Regression

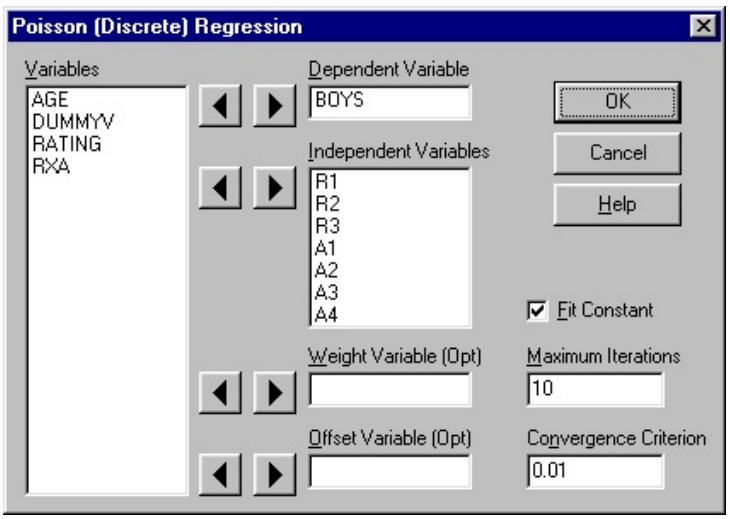
The **Poisson Regression** procedure performs Poisson regression using the maximum likelihood estimation method. It's used when you're interested in examining how observed counts depend on particular independent variables. A direct application of linear regression to counts often isn't satisfactory because the fitted or predicted values may be negative; this is impossible for counts. There may be other shortcomings as well. Poisson regression provides a convenient alternative. It examines the relationships between the log transformed counts and linear combinations of the predictor (independent) variables.

More background on Poisson regression can be found in Additional Background on Logistic and Poisson Regression on page 211. In particular, you should be familiar with likelihood ratio tests (also known as analysis of deviance tests or G2 tests) to make full use of this procedure.

Specification

A sample Poisson Regression dialog box appears on the next page. Move the dependent variable to the ***Dependent Variable*** box and the independent variables to the ***Independent Variables*** box. As with Linear Regression, you can specify a ***Weight Variable*** for prior case weights and force the model through the origin (uncheck the ***Fit Constant*** check box).

In some circumstances, the regression coefficient for a term in the model is known beforehand. Such a term is called an offset, and the offset option allows it to be "adjusted out" of the model. The offset variable is subtracted from the linear predictor, so the ***Offset Variable*** must be expressed on the linear predictor's scale (i.e., natural log of counts).



Poisson regression uses an iterative procedure (iterative reweighted least squares) to obtain the maximum likelihood estimates. You can specify the ***Maximum Iterations*** performed before the procedure "gives up" if it hasn't converged.

Iteration stops when the absolute change in the deviance between iterations reaches the deviance ***Convergence Criterion***. Decreasing the criterion will increase the estimation accuracy but may increase the number of iterations required. The default value of 0.01 is usually suitable for obtaining deviances and coefficient estimates. Decreasing the criterion appears to improve the accuracy of coefficient standard errors and the regression diagnostics.

Data Restrictions

Up to 50 independent variables can be included in the model. If any values within a case are missing, the case is dropped (listwise deletion). If an independent variable is too highly correlated with a linear combination of other independent variables in the model (collinearity), it's dropped from the model. Computation is reinitiated with a new model in which the offending independent variable has been dropped. Variables are dropped

until such collinearity has been eliminated and reliable computations can proceed. If a prior weight variable is specified, the weight variable cannot contain negative weights. Zero weights are treated as missing values.

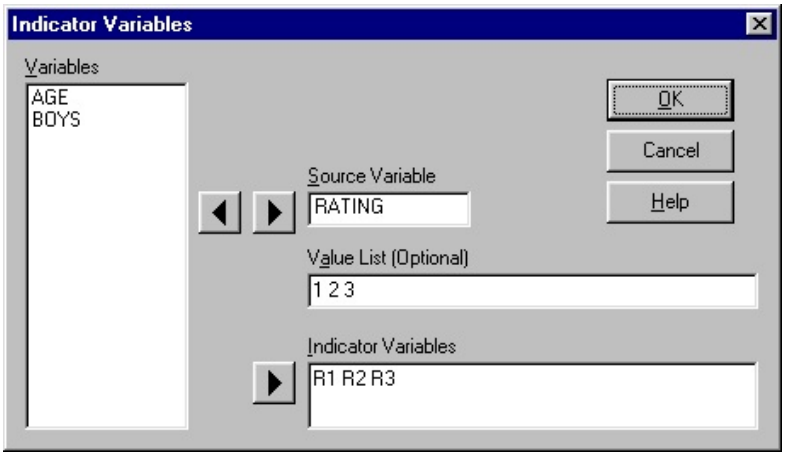
Example

Maxwell's data, presented in Nelder and Wedderburn (1972), are used. The analysis treats a 5 x 4 contingency table giving the number of boys (BOYS) with four different ratings for disturbed dreams in five different age categories:

		RATING			
Age group	AGE	4	3	2	1
5 - 7	1	7	3	4	7
8 - 9	2	13	11	15	10
10 - 11	3	7	11	9	23
12 - 13	4	10	12	9	28
14 - 15	5	3	4	5	32

AGE has the values 1 - 5, and RATING the values 1 - 4. We're interested in whether there's a linear x linear interaction of AGE and RATING.

First we create indicator variables (dummy variables) for the main effects as if we were using linear regression (see, for example, Weisberg 1985). There are 4 main effect degrees of freedom for AGE and 3 main effect degrees of freedom for RATING. The main effect indicator variables R1, R2, and R3 for RATING are created using the **Indicator Variables** procedure from the **Data** menu (discussed in Chapter 2):



The indicator variables A1, A2, A3, and A4 for AGE are created in a similar manner. The RATING X AGE interaction is computed using the

Transformation:

RXA = RATING * AGE

The data are stored in the file Sample Data\dreams.sx, and are listed below.

CASE	BOYS	AGE	RATING	A1	A2	A3	A4	R1	R2	R3	RXA	DUMMYV
1	7	1	4	0	0	0	0	0	0	0	4	0
2	3	1	3	0	0	0	0	0	0	1	3	0
3	4	1	2	0	0	0	0	0	1	0	2	0
4	7	1	1	0	0	0	0	1	0	0	1	0
5	13	2	4	1	0	0	0	0	0	0	8	0
6	11	2	3	1	0	0	0	0	0	1	6	0
7	15	2	2	1	0	0	0	0	1	0	4	0
8	10	2	1	1	0	0	0	1	0	0	2	0
9	7	3	4	0	1	0	0	0	0	0	12	0
10	11	3	3	0	1	0	0	0	0	1	9	0
11	9	3	2	0	1	0	0	0	1	0	6	0
12	23	3	1	0	1	0	0	1	0	0	3	0
13	10	4	4	0	0	1	0	0	0	0	16	0
14	12	4	3	0	0	1	0	0	0	1	12	0
15	9	4	2	0	0	1	0	0	1	0	8	0
16	28	4	1	0	0	1	0	1	0	0	4	0
17	3	5	4	0	0	0	1	0	0	0	20	0
18	4	5	3	0	0	0	1	0	0	1	15	0
19	5	5	2	0	0	0	1	0	1	0	10	0
20	32	5	1	0	0	0	1	1	0	0	5	1

We'll fit two models—one without the interaction and one with the interaction. The main effects only model is computed first. The model is specified in the dialog box on the preceding page. The results are presented in the coefficient table below.

Unweighted Poisson Regression of BOYS

Predictor Variables	Coefficient	Std Error	Coef/SE	P
Constant	1.32623	0.26102	5.08	0.0000
R1	0.91629	0.18708	4.90	0.0000
R2	0.04879	0.22093	0.22	0.8252
R3	0.02469	0.22224	0.11	0.9115
A1	0.84730	0.26082	3.25	0.0012
A2	0.86750	0.26004	3.34	0.0008
A3	1.03302	0.25410	4.07	0.0000
A4	0.73967	0.26522	2.79	0.0053

Deviance	32.46
P-Value	0.0012
Degrees of Freedom	12

Convergence criterion of 0.01 met after 4 iterations

Cases Included 20 Missing Cases 0

The deviance for this model is 32.46 with 12 degrees of freedom; this main effects only model appears to fit poorly (p = 0.0012).

Next we fit the main effects plus linear interaction model by adding the RXA term. This model fits much better (deviance = 14.08, df = 11, p =

0.2288). The contribution to the deviance due to the interaction RXA is 32.46 - 14.08 = 18.38, which is clearly significant (it is treated as a chi-square statistic with 1 df). From the coefficient table, the estimated linear x linear interaction is -0.2051. Nelder and Wedderburn conclude "that the data are adequately described by a negative linear x linear interaction (indicating that the dream rating tends to decrease with age)".

If we look at the regression diagnostics, Cook's distance calls attention to the cell in the lower right; the count in AGE = 5, RATING = 1 is a bit high, and Cook's distance indicates this is a rather influential case. Let's see what happens when this point is fitted separately. To do this, a new variable is created using the **Transformation**:

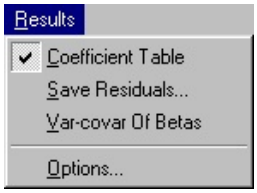
```
IF AGE=5 AND RATING=1 THEN DUMMYV = 1 ELSE DUMMYV = 0
```

When the model: BOYS = R1 R2 R3 A1 A2 A3 A4 RXA DUMMYV is fitted, the new deviance is 9.58 (10 degrees of freedom, p = 0.4781). The RXA interaction slope is now -0.13758. The change in deviance due to the DUMMYV term is 4.50 = 14.08 - 9.58, which can be treated as a chi-square statistic with 1 df. The p-value of 0.034 (from Probability Functions) is small enough to make you suspect that the lower right corner does require special attention. This is not the definitive analysis of this data set; the point is just to show the value of using the regression diagnostics to gain a better understanding of the data.

We could have fitted the main effects only model a little easier using the **Log Linear Models** procedure in Chapter 7. Log Linear Models is especially suited for fitting models with only qualitative predictors; the drawback of using Poisson regression is that you must manually create the indicator variables. However, Log Linear Models can't deal with quantitative predictors and so, for example, they could not compute the linear x linear interaction RXA. Poisson regression must be used when you need the standard errors of the coefficients.

Poisson Regression Results Menu

Once the regression analysis is computed and displayed, a ***Results*** pop-up menu is added to the main menu at the top of the *Statistix* window. Click on the Results menu to display the regression results menu shown on the next page.



Select Coefficient Table from the menu to redisplay the regression coefficient table. Select Options to return to the main dialog box used to specify the model. Poisson regression offers the options of saving various residual- and model-diagnostic statistics and examining the variance-covariance matrix of the regression coefficients. These options are performed in the same manner as in **Linear Regression** (pages 176 and 185).

General Comments

By relating ln(Y) to a linear combination of predictors, we're assuming that the predictors act in a multiplicative fashion to influence the counts Y. Remember that Poisson regression is relating the linear combination of predictors to ln(Y) and not Y, as the analysis specification might suggest.

Additional Background on Logistic and Poisson Regression

Analyzing Proportions and Counts

For those unfamiliar with logistic and Poisson regression, the following sections give you a brief background on why and when these procedures should be used.

The analysis of counts and proportions is the objective of discrete, or categorical, data analysis. The theory of analysis of variance and regression is more mature than methods for discrete data analysis. So it's natural that many of the techniques for discrete data analysis are based on ideas from analysis of variance and regression.

The typical application of least squares procedures [analysis of (co)variance or regression] usually involves the assumption that the dependent variable is some quantity that can be measured on a continuous scale, such as millimeters or grams. It's also usually assumed that the random errors in

the dependent variable are independent normal random variables with identical variances, perhaps after suitable transformations or weightings. The goal of analysis, then, is to examine whether the dependent variable is influenced by the independent variables of interest.

Problems arise when least squares is applied to data sets where the dependent variable is discrete counts or proportions that arose from discrete counts. Suppose you were to fit a simple linear regression model $p = a + bx$ to a data set, where the p's are proportions. You can always find values of x where p is less than 0 or greater than 1, a clearly undesirable situation for a model of proportions. Likewise, if you fit the model $c = a + bx$ where now c is discrete count data, you can find values of x that result in negative predicted counts, also undesirable.

You can avoid these problems by transforming the dependent data. For example, when the data are proportions, the logistic transformation $\ln [p/(1-p)]$ creates a new variable that ranges from $-\infty$ to $+\infty$. With count data, you can achieve a similar scaling by converting the data to the log scale. We are then interested in how a linear combination of the independent variables is related to the transformed data. This involves estimating the slope coefficients and testing their significance. However, even after the data have been transformed, the usual application of least squares for estimation often doesn't work very well because the errors do not closely approximate the usual assumption of identical variances.

The application of least squares to transformed discrete data can be improved by using various weighting schemes to adjust for unequal variances. A classic example of such a procedure is minimum logit C^2, described in Snedecor and Cochran (1980). These techniques should be viewed as approximations to the preferred method of estimation, which is maximum likelihood (ML) estimation. These approximate methods work well in some situations and poorly in others. The appeal of the approximate methods has been that they are easy to calculate with traditional methods.

Efficient general algorithms for ML estimation of linear models fitted to transformed discrete data are now available and are part of a larger body of theory known as generalized linear models, or GLM's. *Statistix* uses these procedures to make it possible to perform ML estimation with no more effort than required for a traditional least squares analysis. If you have count data, we suggest you use the appropriate ML procedure. This may require some background reading if you're not familiar with ML procedures, especially likelihood ratio tests. If you understand how F tests

are performed in the usual regression situation, likelihood ratio tests are easy. The ML procedures always work as well as the approximate weighted least squares procedures. More important, the ML procedures will work well in situations where least squares fails dismally.

If you're interested in proportions, **Logistic Regression** is the appropriate procedure. If you have count data, consider either **Poisson Regression** or **Log-Linear Models**. If you have count data and your independent variables include continuous variables ("covariates"), use Poisson Regression. More detail to help you decide which to use is found within each of the respective descriptions.

There are many good references on discrete data analysis. Bishop *et al*. (1975) give a thorough treatment of discrete analysis for categorical designs, the discrete analogs to analysis of variance. Fienberg (1977) gives a concise, readable account of such models. McCullagh and Nelder (1983) consider a broader class of models, including those with continuous variates, the analogs to regression or analysis of covariance. Cox (1970) and Hosmer and Lemeshow (1989) are good references for the logistic model. The references in these books can direct you to more specific areas of interest.

Generalized Linear Models

The procedures for **Logistic Regression** and **Poisson Regression** are based on the theory of generalized linear models (McCullagh and Nelder 1983), or GLM's. The class of models included in GLM's is quite rich; multiple, logistic, and Poisson regression are the most commonly encountered members of GLM's, but many others are included.

Many of the statistics in linear regression have generalized analogs in GLM's. A distinction is that many of the statistics for GLM's in general are justified by large sample approximations, while the statistics for normal theory linear regression are "exact". The performance of these approximations is an active field of investigation.

First we'll discuss the statistics displayed on the coefficient table. The standard measure of GLM fit is the deviance, also known as the G^2 statistic. Under the null hypothesis of the specified model, the deviance asymptotically follows a chi-square distribution. The deviance plays a role similar to that of the residual error in linear regression. You can think of it as a "distance measure" between the fitted model and the actual data—the smaller, the better. When used as an overall goodness-of-fit measure, the

corresponding p-value should be interpreted with caution; it is best viewed as simply a convenient way to standardize the deviance for comparison purposes. A more traditional goodness-of-fit measure is Pearson's chi-square, which *Statistix* does not display (it's easy to compute from residual results if desired). Pearson's chi-square has the same asymptotic distribution as the deviance (under the null hypothesis); the reason for preferring the deviance is that the deviance can be used to construct a stepwise analysis of deviance table similar to the stepwise analysis of variance table displayed in linear regression. Pearson's chi-square doesn't lend itself as readily to this use.

Stepwise analysis of variance tables and their uses are described in **Linear Regression**. Analogous analysis of deviance tables can be constructed for **Logistic** and **Poisson Regression** (see examples). Understanding the construction and interpretation of such tables is essential to performing regression analyses properly. *Statistix* doesn't automatically generate analysis of deviance tables because the structure of these tables often depends on the goals of the investigation imposed by the subject matter. However, such tables are easy to construct by hand from the displayed results.

In addition to the model deviance, COEF/SE statistics and associated p-values are displayed on the coefficient table. Again, these are justified by large sample theory. Better tests for the individual contributions of independent variables can be constructed as 1 degree of freedom tests using the deviance (see examples).

Direct analogies exist for most of the residual diagnostics you can select from the residuals and fitted values menu in linear regression. The computation of leverage is as described in McCullagh and Nelder (1983) and Pregibon (1981). The standardized residual, r, is described in McCullagh and Nelder (eq. 11.1); we describe its calculation at the end of this section on page 216. Leverage values, h, are the diagonal elements of the H matrix described by McCullagh and Nelder (sect. 11.3).

If you're interested in Pearson (chi-square) residuals rather than standardized residuals, you can calculate them as $r(1 - h)^{½}$ using **Transformations.** If you square the Pearson residuals and then sum them, you'll obtain the Pearson chi-square goodness-of-fit statistic for the model. Cook's distance is computed as described in McCullagh and Nelder (1983) and Pregibon (1981). Computing the "p-value" for Cook's D should simply be regarded as converting D to an alternate scale that may help make

influential points more easily recognizable. The method for computing the "p-value" of Cook's D is the same as that used in linear regression. The outlier t-statistic and its p-value should be regarded as experimental; at this point, a cautious interpretation of these statistics is that they are monotonic transforms of the standardized residuals (Weisberg 1985) and should be sensitive to outliers.

Aliasing

Parameters being estimated are said to be aliased if the associated independent variables are (nearly) linear combinations of other independent variables. Another name for this is collinearity. As in linear regression, aliased independent variables in logistic and Poisson regression are successively detected and dropped until the remaining independent variables constitute a linearly independent set.

Some special considerations come into play when this variable-dropping technique is used for logistic and Poisson regression. The first is what is called "saw-toothing". As the iterative fitting process used for GLM's proceeds, it occasionally happens that the information in a parameter diminishes to the point that the procedure detects it as being aliased, even though it really isn't. Saw-toothing can be recognized as occurring when a variable is dropped sometime after the first fitting cycle, along with a substantial increase in the deviance to the next cycle. When this is observed, the deviance from the cycle immediately before the one in which the variable is dropped should be used.

There is a related issue that you should be aware of. This is the potential interaction of dropping aliased variables and missing values. A case is not included in the analysis if any of the values for any of the variables (dependent, independents, or weight) used in the analysis are missing values. Thus, if an independent variable with missing values is detected as being aliased and is dropped, new cases may be pulled into the analysis in the next cycle. This makes use of all of the available data, but it also means the deviance may be based on an expanded set of cases as iteration continues and hence is not truly comparable from cycle to cycle. This will usually not be of concern, but it's a point you need to know. If this behavior influences the results, it can be controlled by using **Omit/Restore Cases** to select case subsets for analysis.

Infinite Parameter Estimates

A potential problem with the logistic transformation occurs when the estimate of p is very near to 0 or 1. The fitted logit ln[p/(1-p)] then approaches minus or plus infinity, which can obviously cause computational problems. A similar problem occurs in Poisson regression when the fitted count is very near 0; the fitted log count approaches minus infinity. In these cases, the parameter estimates in the linear predictors approach infinity. *Statistix* prevents potential computational problems in such cases by enforcing an upper bound on the absolute value of the fitted values of the linear predictors.

This approach appears to work quite well, and biases are generally small. However, when it's known in advance that some parameter estimates are infinite, it's better to drop the corresponding data cases from the analysis. An example would be to delete all rows and columns in two-way contingency tables that have marginal sums of zero. The major negative consequence of not deleting such cases is that the deviance may be distorted downward relative to the degrees of freedom.

Equations for Linear, Logistic, and Poisson Regression

The following section outlines some of the equations used for various quantities in multiple, logistic, and Poisson regression. The matrix X represents the design matrix of predictor variable values. The prior weights are the weights specified by the weighting variable option. RMS stands for residual mean square. In logistic regression, p is the expected proportion, or E[p] (the expected logit is ln [p/(1-p)]). The various terms are described in more detail in McCullagh and Nelder (1983). The number of cases used for estimation is n. The number of parameters in the linear model is m (m includes the intercept, if present).

Description	Linear	Logistic	Poisson
prior weight	w	w	w
s = scale parameter	σ^2 σ^2 = E[RMS]	1	1
f = variance function	1	$n\pi(1 - \pi)$ π = E[p]	M M= expected count
q = iterative weight = fw	w	$n\pi(1 - \pi)w$	Mw

Description	Linear	Logistic	Poisson
v = variance of an observation = sf/w	σ^2/w	$n\pi(1 - \pi)/w$	M/w

Let $C = (X^TQX)^{-1}$

variance of estimated coefficients $V(B^*) = sC$

variance of a fitted value of the linear predictor $V(f) = sx^TCx$

h = leverage = $q\, x^TCx$
("unusualness" or "distance" is fx^TCx)

Let e = raw residual (observation minus fitted or predicted value)

Standardized residual = $r = e[((sf) / w)(1 - h)]^{-½}$

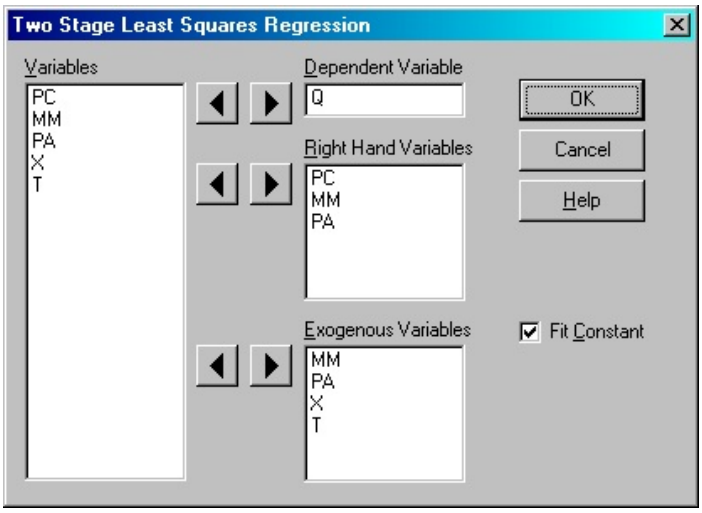
outlier t-statistic = $r [(n - m - 1) / (n - m - r^2)]^{-½}$

Cook's $D = (r^2 / m) (h / (1 - h))$

Two Stage Least Squares Regression

The **Two Stage Least Squares Regression** procedure (2SLS) is used to estimate a linear equation when one or more of the predictor variables, or right hand side variables, is an endogenous variable. An endogenous variable is one that is determined by the system of equations being solved. For example, quantity and price of a product are both endogenous variables determined by a system of two simultaneous equations: the demand curve and the supply curve. The 2SLS model also requires at least one exogenous variable. An exogenous variable is one whose value is determined outside the system of equations.

Specification



First select the name of the dependent variable (response variable) and move it to the ***Dependent Variable*** box.

Then select one or more right hand variables and copy them to the ***Right Hand Variables*** list box. The right hand variables are analogous to the independent variables in ordinary least squares regression and will be listed in the coefficient table.

Next select one or more exogenous variables and copy them to the ***Exogenous Variables*** list box. At least one of the exogenous variables must not be among those selected for the Right Hand Variables.

Use the ***Fit Constant*** check box to specify a constant fitted model (checked) or a model forced through the origin (not checked). Press the ***OK*** button to begin computing the analysis.

Data Restrictions

Up to 50 variables (endogenous plus exogenous variables) can be specified. At least one exogenous variable that is not also a right hand variable must be specified. If there are missing values for any of the variables for a case, the entire case is deleted (listwise deletion).

Example

The purpose of the example data is to estimate the world demand for copper (Maurice and Thomas, 2002). The data are listed on the next page, and are available in the file Sample Data\Copper.sx.

CASE	Q	PC	MM	PA	X	T
1	3173.0	26.56	0.70	19.76	0.97679	1
2	3281.1	27.31	0.71	20.78	1.03937	2
3	3135.7	32.95	0.72	22.55	1.05153	3
4	3359.1	33.90	0.70	23.06	0.97312	4
5	3755.1	42.70	0.74	24.93	1.02349	5
6	3875.9	46.11	0.74	26.50	1.04135	6
7	3905.7	31.70	0.74	27.24	0.97686	7
8	3957.6	27.23	0.72	26.21	0.98069	8
9	4279.1	32.89	0.75	26.09	1.02888	9
10	4627.9	33.78	0.77	27.40	1.03392	10
11	4910.2	31.66	0.76	26.94	0.97922	11
12	4908.4	32.28	0.79	25.18	0.99679	12
13	5327.9	32.38	0.83	23.94	0.96630	13
14	5878.4	33.75	0.85	25.07	1.02915	14
15	6075.2	36.25	0.89	25.37	1.07950	15
16	6312.7	36.24	0.93	24.55	1.05073	16
17	6056.8	38.23	0.95	24.98	1.02788	17
18	6375.9	40.83	0.99	24.96	1.02799	18
19	6974.3	44.62	1.00	25.52	0.99151	19
20	7101.6	52.27	1.00	26.01	1.00191	20
21	7071.7	45.16	1.02	25.46	0.95644	21
22	7754.8	42.50	1.07	22.17	0.96947	22
23	8480.3	43.70	1.12	18.56	0.98220	23
24	8105.2	47.88	1.10	21.32	1.00793	24
25	7157.2	36.33	1.07	22.75	0.93810	25

The variables are Q: quantity sold; PC: price of copper; MM: per capita income; PA: price of aluminum (an alternative to copper); X: inventory of copper; and T: time as a proxy for technology. Consider the simultaneous functions of the demand and supply of copper:

Q = f (PC, MM, PA)
Q = f (PC, X, T)

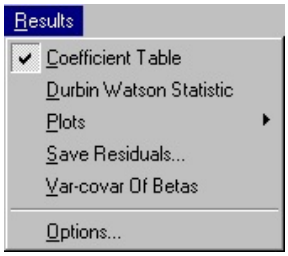
The demand is a function of PC, MM, and PA. Since price (PC) is determined by both the demand function and the supply function, it is a endogenous variable. The model is specified on the preceding page. The variables PC, MM, and PA are copied to the Right Hand Variables box. The right hand variables MM and PA are exogenous variables, so we copy them to the Exogenous Variables box also. The variables X and T are not right hand variables in the demand function, but only appear in the supply function. Thus, they are exogenous variables, so we copy them to the Exogenous Variables box. At least one such variable is required, or the model is said to be underidentified. The results are shown on the next page.

The summary statistics reported have the same interpretation as with ordinary least squares regression. The negative sign of the coefficient for PC means that the quantity of copper sold decreases with increasing price of copper, which is what we would expect.

```
Two Stage Least Squares Regreesion of Q
4 Exogenous Variables: MM, PA, X, T

Predictor
Variables   Coefficient   Std Error         T         P
Constant       -6837.83     1264.46     -5.41    0.0000
PC             -66.4950     31.5338     -2.11    0.0472
MM              13997.7     1306.34     10.72    0.0000
PA              107.662     44.5098      2.42    0.0247

R-Squared          0.9421      Redid. Mean Square (MSE)     184327
Adjusted R-Squared  0.9339     Standard Deviation          429.333

Cases Included 25    Missing Cases 0
```

2SLS Results Menu

Once the regression analysis is computed and displayed, a ***Results*** pull-down menu appears on the menu at the top of the *Statistix* window. Click on the Results menu to display the 2SLS results menu displayed below.



Select Coefficient Table from the menu to redisplay the regression coefficient table. Select Options to return to the main dialog box used to specify the model. Like the Linear Regression procedure, 2SLS regression offers the options of computing the Durbin-Watson statistic for autocorrelation, (see page 171), displaying residual plots (see page 175), saving fitted values and residuals (see page 176), and examining the variance-covariance matrix of the regression coefficients (see page 185).

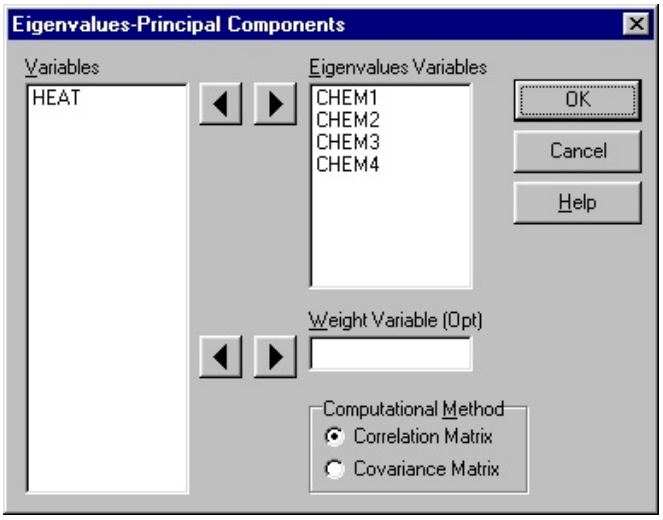
Computaional Notes

In the first stage, ordinary least squares regression is used to compute the fitted values of each endogenous variable using the full set of exogenous variables as predictor variables. The vectors of fitted values are then used in place of the original endogenous variables to estimate the final equation in the second stage, again using OLS regression. See Griffiths *et al*. (1993) for details.

Eigenvalues-Principal Components

The **Eigenvalues-Principal Components** procedure displays the eigenvectors and eigenvalues for a list of variables. You can also save principal components as new variables.

Specification



Select the variables you want to analyze and move them to the ***Eigenvalues Variables*** list box. To weight the analysis, move the name of the variable that contains the values to be used as weights to the ***Weight Variable*** box. The eigenvalues can be based on the correlation matrix or the covariance matrix. Select the method you want by clicking on one of the ***Computational Method*** radio buttons.

Data Restrictions

Up to 50 variables can be specified. If there are missing values for any of the variables for a case, the entire case is deleted (listwise deletion). Negative weights are not allowed.

Example

We use the Hald data from Draper and Smith (1966) for this example. The same data are used to illustrate **Linear Regression** and are listed on page 162. The variables CHEM1, CHEM2, CHEM3, and CHEM4 are the percentages of four chemical compounds measured in batches of cement.

The analysis is specified on the preceding page. The results are presented below.

Eigenvalues / Eigenvectors based on Correlation Matrix

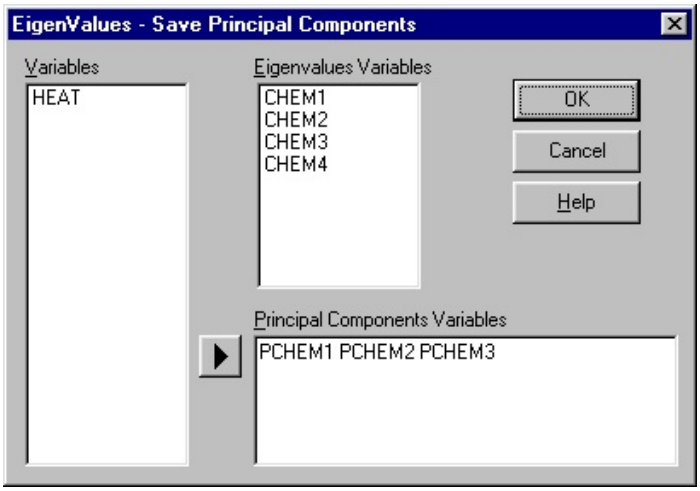
	Eigenvalues	Percent of Variance	Cumulative Percent of Variance
1	2.23570	55.9	55.9
2	1.57607	39.4	95.3
3	0.18661	4.7	100.0
4	0.00162	0.0	100.0

Vectors

Factor	1	2	3	4
CHEM1	0.4760	0.5090	0.6755	0.2411
CHEM2	0.5639	-0.4139	-0.3144	0.6418
CHEM3	-0.3941	-0.6050	0.6377	0.2685
CHEM4	-0.5479	0.4512	-0.1954	0.6767

Eigenvalue analysis is interesting in its own right as a way to analyze multivariate data structure (Morrison 1977, chap. 8). It's also an important supplement to multiple regression analysis (Chatterjee and Price 1991).

Principal Components



After viewing the resulting eigenvalues and eigenvectors, select **Principal Components** from the ***Results*** menu to display the principal components dialog box. Then enter variable names for the principal components.

The dialog box on the preceding page creates three new variables PCHEM1, PCHEM2, and PCHEM3 to store the first three principal components. The

principal component with the largest eigenvalue (variance component) comes first, the next largest second, and so on.

In some regression data sets, the independent variables may be highly correlated with one another. When such collinearity exists, estimates of the regression coefficients may be unstable and can lead to erroneous inferences. If this is the case, it's sometimes useful to perform the regression on a set of principal components. The computational advantage of using principal components rather than the original data is that they're all uncorrelated (they're said to be orthogonal). Chatterjee and Price (1991) give a nice example of the application of this method.

Typically, the principal components corresponding to the largest eigenvalues (i.e., largest variance components) are used for the regression. Jolliffe (1982) makes the point that this is not always a wise approach.

Generally, it's best to use the correlation matrix to compute the eigenvalues because in effect this assigns equal weight to each variable. If the covariance matrix is used, the results depend on the scales on which the original variables were measured.

Computational Notes

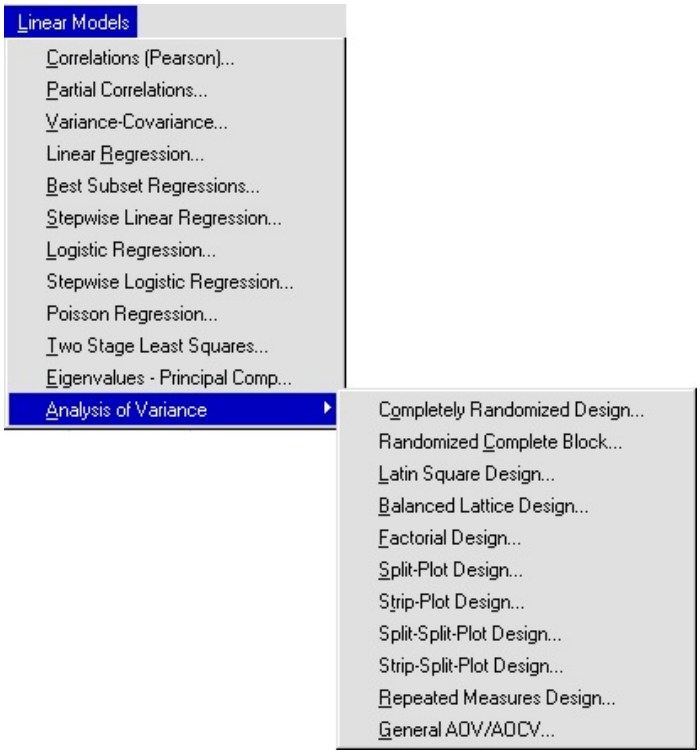
The correlation or covariance matrix is first computed using the method of updating (see **Correlations**). The resulting matrix is converted to tridiagonal form using Householder reductions. The eigenvalues and eigenvectors are then extracted using the QL decomposition. Details on these methods are described by Martin *et al*. (1968) and Bowdler *et al*. (1968).

In matrix notation, the principal components are calculated as XU, where X is an n x p matrix derived from the original data and U is a p x m matrix of eigenvectors. The number of usable cases is n, the number of variables in the original variable list is p, and m is the number of eigenvectors retained. If the calculations were based on the correlation matrix, the data in X are the original data after Studentization (the means subtracted from the values and then divided by their standard deviations). If the calculations are based on the covariance matrix, X is the original data with the means subtracted.

C H A P T E R

7

Analysis of Variance

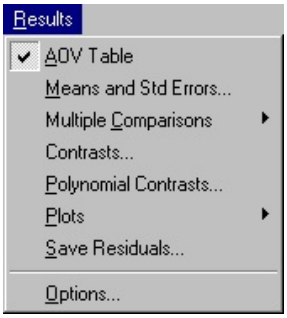


The analysis of variance menu, accessed from the Linear Models menu, offers you a wide variety of AOV designs. Each of the designs listed on the

menu provides a specific dialog box relating to that design making it easy to specify the model. The dialog box for the General AOV/AOCV procedure offers the flexibility of specifying the model directly by entering a model statement, and listing covariates for analysis of covariance.

All of the AOV procedures, with the exception of the Balanced Lattice Design, can handle unbalanced designs (data with missing values). Unbalanced designs, and designs with covariates, are solved using general linear models (GLM) techniques. The sums of squares listed in the AOV tables for unbalanced designs and designs with covariates are marginal sums of squares, also called type III sums of squares. These are the correct sums of squares to use when constructing F-tests that test the hypothesis that the means for the term in question are equal, given that the remaining terms are in the model. An important limit on the level of unbalancedness is that any interaction term included in the model must have all cells filled (no cells empty).

All of the analysis of variance procedures contain a results menu offering numerous powerful options including multiple comparisons, linear contrasts, polynomial contrasts, means plot, and residual plots. The results menu appears on the main menu when the basic analysis of variance table is displayed.



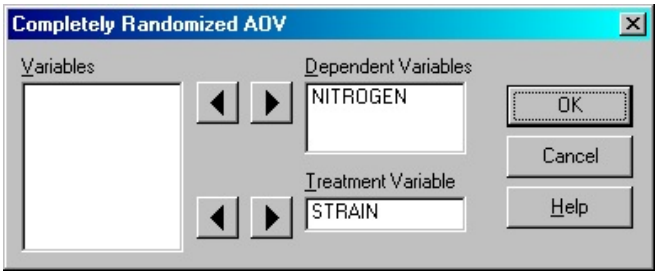
These options are discussed in detail at the end of this chapter.

Completely Randomized Design

This procedure computes the analysis of variance for the completely randomized design. As implied by the name, the allocation of treatments to the experimental units is performed completely at random. The advantages of this design are that the number of treatments is flexible, the loss of information because of missing values is relatively low, and the degrees of freedom for error is maximum. The disadvantage is that it's often inefficient because the experimental error includes all the variation between experimental units except that due to treatments.

The **Completely Randomized Design** also goes by the name One-Way Design. This procedure produces the same results as the **One-Way AOV** procedure discussed in Chapter 5.

Specification



The observed data must be entered into a single variable to use this procedure. Move the variable name containing the observed data to the ***Dependent Variables*** box. If you specify more than one dependent variable, a separate analysis is produced for each variable. A second variable identifying the different treatment groups is also required. Move the grouping variable to the ***Treatment Variable*** box. Press the ***OK*** button to start the analysis.

Data Restrictions

Up to ten dependent variables can be specified. Sample sizes within treatment levels can be unequal. The maximum number of treatment levels is 500. The treatment variable can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters.

Example

The example data are from Steel and Torrie (1980, p. 139). Nitrogen content was measured from six strains of red clover. Five plots in a greenhouse were randomly assigned to each strain. The data are shown in the table below, and is available in the file Sample Data\red clover.sx.

CASE	NITROGEN	STRAIN	CASE	NITROGEN	STRAIN
1	19.40	3DOk1	16	20.70	3DOk7
2	32.60	3DOk1	17	21.00	3DOk7
3	27.00	3DOk1	18	20.50	3DOk7
4	32.10	3DOk1	19	18.80	3DOk7
5	33.00	3DOk1	20	18.60	3DOk7
6	17.70	3DOk5	21	14.30	3DOk13
7	24.80	3DOk5	22	14.40	3DOk13
8	27.90	3DOk5	23	11.80	3DOk13
9	25.20	3DOk5	24	11.60	3DOk13
10	24.30	3DOk5	25	14.20	3DOk13
11	17.00	3DOk4	26	17.30	Composite
12	19.40	3DOk4	27	19.40	Composite
13	9.10	3DOk4	28	19.10	Composite
14	11.90	3DOk4	29	16.90	Composite
15	15.80	3DOk4	30	20.80	Composite

The analysis is specified on the preceding page. The results are shown below.

Completely Randomized AOV for NITROGEN

Source	DF	SS	MS	F	P
STRAIN	5	847.05	169.409	14.4	0.0000
Error	24	282.93	11.789		
Total	29	1129.97			

Grand Mean 19.887 CV 17.27

	Chi-Sq	DF	P
Bartlett's Test of Equal Variances	14.2	5	0.0143

Cochran's Q 0.4756
Largets Var / Smallest Var 26.345

Component of variance for between groups 31.5241
Effective cell size 5.0

STRAIN	Mean
3DOk1	28.820
3DOk5	23.980
3DOk4	14.640
3DOk7	19.920
3DOk13	13.260
Composite	18.700

Observations per Mean 5
Standard Error of a Mean 1.2391
Std Error (Diff of 2 Means) 1.4736

A standard analysis of variance table is displayed first. Note that the F test suggests a substantial between-groups (strains of red clover) effect, with a p-value less than 0.0001. The coefficient of variation (CV) expresses the experimental error as a percentage of the mean; the higher the CV value, the lower the reliability of the experiment.

The F test assumes that the within-group variances are the same for all groups. Bartlett's test for equality of variances tests this assumption; it is shown below the analysis of variance table. The p-value of 0.0143 suggest that the variances are unequal. Bartlett's test is described in Snedecor and Cochran (1980, p. 252). Another test of equality of variances, Cochran's Q, is given below Bartlett's test. Cochran's Q statistic is the ratio of the largest within-group variance over the sum of all within-group variances. The ratio of the largest within-group variance over the smallest has also been a popular test for equal variances and is displayed under Cochran's Q; tables are given in Pearson and Hartley (1954).

A fixed-effects model (Type I) is appropriate for these data. If a random-effects model were appropriate (Type II), the component of variance for between groups may be of interest (see Snedecor and Cochran, chap. 13). The between-groups variance component and effective cell sample size are displayed below the equality of variance tests. The computation of effective cell size is described on page 246 of Snedecor and Cochran.

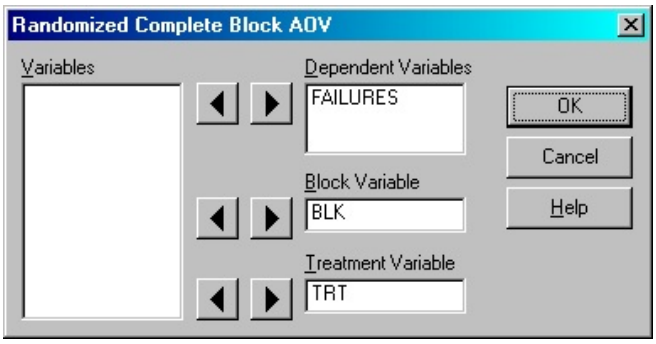
The bottom portion of the report lists a table of treatment means, sample sizes, and standard errors of the means. The standard error of the difference of two means is reported when the sample sizes are equal.

Randomized Complete Block Design

The **Randomized Complete Block** (RCB) design is used to reduce experimental error by dividing the experimental units into blocks of units that are thought to be similar. The object of blocking is to minimize the variability within the blocks, and maximize the variability between the blocks. Blocks in the RCB design are of equal size, each of which contains all the treatments. This procedure handles missing values by using the GLM technique to compute marginal sums of squares.

This procedure is used to analyze single-factor RCB designs (i.e., one treatment factor). Multiple-factor experiments in a RCB design can be analyzed using the **Factorial Design** procedure discussed on page 238.

Specification



Move the name of the variable containing the observed data to the ***Dependent Variables*** box. If you specify more than one dependent variable, a separate analysis is produced for each variable. Move the variable that identifies blocks to the ***Block Variable*** box. Move the variable that identifies treatments to the ***Treatment Variable*** box. Press the ***OK*** button to start the analysis.

Data Restrictions

Up to ten dependent variables can be specified. The maximum number of block and treatment levels are 200 each. The block and treatment variables can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. Missing values are allowed.

Example

The example data are from Snedecor and Cochran (1980, sect. 14.2). The dependent variable is the number of soybeans out of 100 that failed to emerge, and the treatments are various fungicides (the first treatment level was a no-fungicide control). The data are listed below, and are stored in the file Sample Data\soybeans.sx.

CASE	FAILURES	TRT	BLK	CASE	FAILURES	TRT	BLK
1	8	1	1	14	8	3	4
2	10	1	2	15	10	3	5
3	12	1	3	16	3	4	1
4	13	1	4	17	5	4	2
5	11	1	5	18	9	4	3
6	2	2	1	19	10	4	4
7	6	2	2	20	6	4	5
8	7	2	3	21	9	5	1
9	11	2	4	22	7	5	2
10	5	2	5	23	5	5	3
11	4	3	1	24	5	5	4
12	10	3	2	25	3	5	5
13	9	3	3				

You can use the **Transformations** CAT function to generate repetitive sequences, such as those seen for TRT and BLK. After entering the 25 values for FAILURES, we can use the Transformation expressions TRT = CAT(5,5) and BLK = CAT(5,1) to create these variables. The model is specified in the dialog box on the preceding page. The results are presented in the table below.

Randomized Complete Block AOV Table for FAILURES

Source	DF	SS	MS	F	P
BLK	4	49.840	12.4600		
TRT	4	83.840	20.9600	3.87	0.0219
Error	16	86.560	5.4100		
Total	24	220.240			

Grand Mean 7.5200 CV 30.93

Tukey's 1 Degree of Freedom Test for Nonadditivity

Source	DF	SS	MS	F	P
Nonadditivity	1	1.4957	1.49569	0.26	0.6150
Remainder	15	85.0643	5.67095		

Relative Efficiency, RCB 1.19

Means of FAILURES for TRT Fungicide treatments

TRT	Mean
Control	10.800
Fung #1	6.200
Fung #2	8.200
Fung #3	6.600
Fung #4	5.800

Observations per Mean 5
Standard Error of a Mean 1.0199
Std Error (Diff of 2 Means) 1.2129

The analysis of variance table appears first in the report. An F test and the associated p-value are listed for the treatment variable TRT. The test suggests a between-fungicides effect (p = 0.0219).

Tukey's one degree of freedom test for nonadditivity is useful when the experimental design only permits an additive model to be fitted to the data but you suspect that interaction is present. There's little suggestion of nonadditivity (p = 0.6150) in this example. If nonadditivity is present, you should consider transforming your data in an effort to remove it.

The object of using blocks is to increase efficiency by reducing the error mean square. The relative efficiency indicates the magnitude to which blocking succeeded in reducing experimental error. In this example, the relative efficiency of using the RCB design over the completely randomized design is 1.19, which is a 19% increase in precision. See Gomez and Gomez (1984) for computational details.

A table of treatment means, sample sizes, and standard errors of the means is displayed at the bottom of the report. The standard error of the difference of two means is reported when the sample sizes are equal.

Results Menu

Once the AOV table is displayed, a results menu appears on the main menu. Use the procedures on this menu to compute multiple comparisons, linear contrasts, polynomial contrasts, means plots, residual plots, and to save residuals. These options are discussed in detail at the end of this chapter.

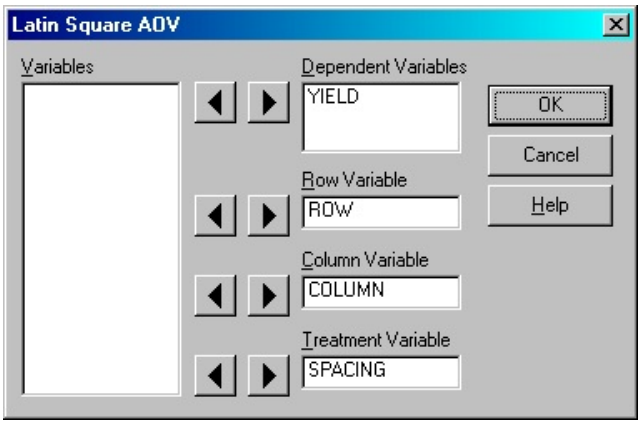
Computational Notes

Oliver's (1967) generalization of Yates' algorithm (Daniel, 1976) is used for balanced designs. Unbalanced designs are computed using general linear models (Searle, 1987; Glantz and Slinker, 1990).

Latin Square Design

This procedure computes the analysis of variance for **Latin Square Designs**. This design simultaneously handles two known sources of variation, commonly referred to as row-blocking and column-blocking.

Specification



Move the name of the variable containing the observed data to the ***Dependent Variables*** box. If you specify more than one dependent

variable, a separate analysis is produced for each variable. Move the variable that identifies row-blocking to the ***Row Variable*** box. Move the variable that identifies column-blocking to the ***Column Variable*** box. Move the variable that identifies treatments to the ***Treatment Variable*** box. Press the ***OK*** button to start the analysis.

Data Restrictions

Up to ten dependent variables can be specified. The number of treatments must be equal to the number of rows and columns. The row, column, and treatment variables can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. Missing values are allowed.

Example

The example data are from a field trial to study the effect of row spacing on the yield of millet (Snedecor and Cochran, 1980). As is common with agricultural field experiments, the row and column blocks represent fertility gradients in two directions. The data are listed below, and are stored in the file Sample Data\millet.sx.

YIELD	ROW	COLUMN	SPACING	YIELD	ROW	COLUMN	SPACING
257	1	1	4	203	4	1	2
230	1	2	10	204	4	2	6
279	1	3	2	227	4	3	8
287	1	4	6	193	4	4	10
202	1	5	8	259	4	5	4
245	2	1	8	231	5	1	6
283	2	2	2	271	5	2	8
245	2	3	10	266	5	3	4
280	2	4	4	334	5	4	2
260	2	5	6	338	5	5	10
182	3	1	10				
252	3	2	4				
280	3	3	6				
246	3	4	8				
250	3	5	2				

If you study the data above, you'll see that the five values for the treatment variable SPACING appear exactly once for each combination of ROW and COLUMN. The model is specified in the dialog box on the preceding page. The results are presented in the table on the next page.

The analysis of variance table appears first in the report. An F test and the associated p-value are listed for the treatment variable SPACING. The test for treatment effect is not significant (p = 0.4523).

Latin Square AOV Table for YIELD

Source	DF	SS	MS	F	P
ROW	4	13601.4	3400.34		
COLUMN	4	6146.2	1536.54		
SPACING	4	4156.6	1039.14	0.98	0.4523
Error	12	12667.3	1055.61		
Total	24	36571.4			

Grand Mean 252.16 CV 12.88

Tukey's 1 Degree of Freedom Test for Nonadditivity

Source	DF	SS	MS	F	P
Nonadditivity	1	142.9	142.89	0.13	0.7298
Remainder	11	12524.4	1138.58		

	Relative Efficiency
Completely Randomized Design	1.45
Randomized Complete Block, ROW	1.06
Randomized Complete Block, COLUMN	1.40

Means of YIELD for SPACING

SPACING	Mean
2	269.80
4	262.80
6	252.40
8	238.20
10	237.60

Observations per Mean 5
Standard Error of a Mean 3.8118
Std Error (Diff of 2 Means) 4.5331

Tukey's one degree of freedom test for nonadditivity is useful when the experimental design only permits an additive model to be fitted to the data but you suspect that interaction is present. There's little suggestion of nonadditivity ($p = 0.7298$) in this example. If nonadditivity is present, you should consider transforming your data in an effort to remove it. See Snedecor and Cochran (1980) for computational details.

As with the RCB design, we're interested in the efficiency of blocking. There are three relative efficiencies reported, indicating the improved efficiencies of the Latin square design compared to three alternative designs. Using a Latin square design in this experiment resulted in a 45% improvement compared to the completely randomized design, but only a 6% improvement over the RCB design using rows as blocks. See Gomez and Gomez (1984) for computational details.

A table of treatment means, sample sizes, and standard errors of the means is displayed at the bottom of the report. The standard error of the difference of two means is reported when the sample sizes are equal.

Results Menu

Once the AOV table is displayed, a results menu appears on the main menu. Use the procedures on this menu to compute multiple comparisons, linear contrasts, polynomial contrasts, means plots, residual plots, and to save residuals. These options are discussed in detail at the end of this chapter.

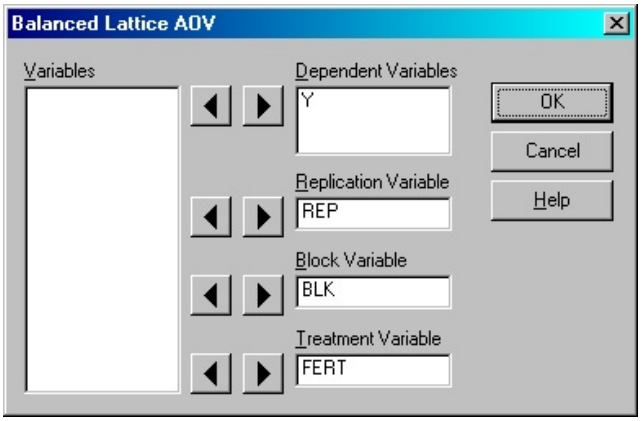
Computational Notes

The analysis of variance is computed using general linear models (Searle, 1987; Glantz and Slinker, 1990).

Balanced Lattice Design

The **Balanced Lattice Design** is an incomplete block design that's useful when the number of treatments is large, which can make a RCB design impractical. Individual blocks don't contain all treatments. This design requires that the number of treatments is a perfect square, the block size is the square root of the number of treatments, and the number of replications is one more than the block size. Unlike the other AOV procedures discussed in this chapter, this procedure doesn't allow missing values.

Specification



Move the name of the variable containing the observed data to the ***Dependent Variables*** box. If you specify more than one dependent variable, a separate analysis is produced for each variable. Move the variables that identify replicates, blocks, and treatments to the ***Replication***

Variable, Block Variable, and ***Treatment Variable*** boxes respectively. Press the ***OK*** button to start the analysis.

Data Restrictions

Up to ten dependent variables can be specified. The maximum number of treatments allowed is 196. The replication, block, and treatment variables can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. Missing values are not allowed.

Example

The example data are from Gomez and Gomez (1984, p. 45). Tiller number per square meter is recorded from 16 fertilizer treatments of rice in a 4 X 4 balanced lattice design. The data for the first of five replicates is listed below. The complete data are stored in the file Sample Data\tiller.sx.

CASE	Y	REP	BLK	FERT
1	147	1	1	1
2	152	1	1	2
3	167	1	1	3
4	150	1	1	4
5	127	1	2	5
6	155	1	2	6
7	162	1	2	7
8	172	1	2	8
9	147	1	3	9
10	100	1	3	10
11	192	1	3	11
12	177	1	3	12
13	155	1	4	13
14	195	1	4	14
15	192	1	4	15
16	205	1	4	16
. . .				
80	220	5	4	14

The model is specified in the dialog box on the preceding page. The results are presented in the table on the next page.

The analysis of variance table lists the sums of squares computed in the usual manner. The mean square for the treatment factor is adjusted to account for an unequal block effect, if one exists, since not all treatments are represented in each block. The F test for the treatment factor is computed using the adjusted mean square for treatment and the effective error. The test for treatment effect is significant (p = 0.0001).

The relative efficiency of the balanced lattice design, compared to the RCB design, for these data is 1.17, indicating a 17% improvement in efficiency.

Balanced Lattice AOV for Y

Source	DF	SS	MS	F	P
REP	4	5946.0			
FERT(unadj)	15	26994.3			
BLK*REP	15	11381.8	758.79		
Intrablock error	45	14533.3	322.96		
FERT(adj)	(15)		1600.12	4.33	0.0001
Effective error	(45)		369.34		
Total	79	58855.5			

Grand Mean 171.82 CV 11.18
Relative efficiency, RCB 1.17

Means of Y for FERT

FERT	Mean	FERT	Mean
1	165.76	9	163.00
2	161.04	10	118.82
3	183.92	11	188.19
4	175.68	12	190.54
5	162.88	13	169.51
6	173.82	14	197.23
7	168.43	15	185.67
8	176.92	16	167.78

Observations per Mean 5
Standard Error of a Mean 2.9317
Std Error (Diff of 2 Means) 3.4863

A table of treatment means is displayed at the bottom of the report. Like the treatment mean square, the treatment means are adjusted to account for an unequal block effect, if one exists, since not all treatments are represented in each block.

Results Menu

Once the AOV table is displayed, a results menu appears on the main menu. Use the procedures on this menu to compute multiple comparisons, linear contrasts, polynomial contrasts, means plots, residual plots, and to save residuals. These options are discussed in detail at the end of this chapter.

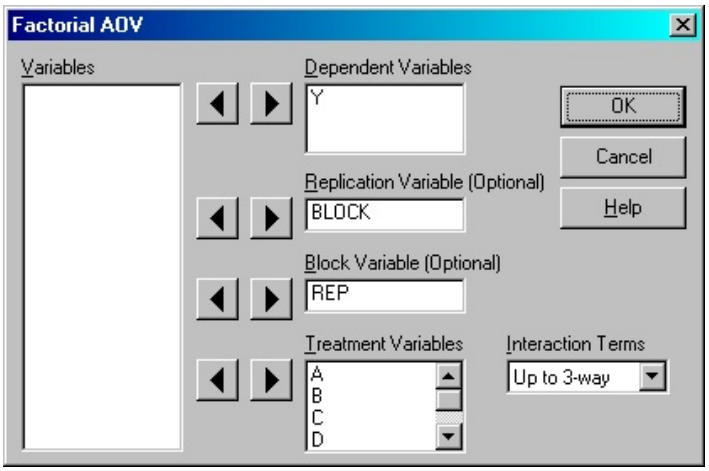
Computational Notes

The analysis of variance, relative efficiency, and adjustments to the treatment means are computed using the algorithms described by Gomez and Gomez (1984).

Factorial Design

The **Factorial Design** procedure computes the analysis of variance for complete factorial designs and fractional factorial designs. It can handle factorial experiments in a completely randomized design, randomized block design without replication, and a randomized block design with replication.

Specification



Move the name of the variable containing the observed data to the ***Dependent Variables*** box. If you specify more than one dependent variable, a separate analysis is produced for each variable.

The ***Replication Variable*** and ***Block Variable*** are both optional. If the experiment was performed in a completely randomized design without replication, then leave both boxes empty. If the experiment was performed in a randomized block design without additional replication, then leave the Replication Variable box empty and move the blocking variable to the Block Variable box. If the experiment is in a randomized block design with replication, then use both boxes the specify replication and blocking variables. You can use the General AOV/AOCV procedure to analyze factorial experiments in other designs, such as a Latin square or split-plot design.

Move the treatment factor variables to the ***Treatment Variables*** box. The ***Interaction Terms*** drop-down list lets you select the highest order of interaction terms to be included in the model: no interactions, up to 2-way interactions, up to 3-way interactions, etc. Press the ***OK*** button to start the analysis.

Data Restrictions

Up to ten dependent variables can be specified. The total number of factors (replication, block, and treatment variables) selected can't exceed ten. The maximum number of levels for each factor is 200. The factor variables can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. Missing values are allowed. For unbalanced and fractional designs, the maximum size of the GLM design matrix is 500. (The size of the design matrix is equivalent to the model degrees of freedom: total degrees of freedom minus error degrees of freedom.)

Example

The example data are from a rice yield trial in a fractional factorial design (Gomez and Gomez, 1984, p. 172). Fractional factorial designs are useful when the number of factors of interest is so large that it would be too expensive or too impractical to include the complete set of factorial treatments. The example data are from a 2^6 factorial experiment in a ½ fractional design with two blocks of 16 experimental plots each, and with two replications. A partial listing of the data are presented below. The complete data are available in the file Sample Data\fractional.sx.

CASE	YIELD	REP	BLK	A	B	C	D	E	F
1	2.92	1	1	0	0	0	0	0	0
2	3.28	1	1	0	0	0	0	1	1
3	3.34	1	1	0	0	0	1	0	1
4	3.29	1	1	0	0	0	1	1	0
5	3.16	1	1	0	1	1	0	0	0
6	3.63	1	1	0	1	1	0	1	1
7	4.00	1	1	0	1	1	1	0	1
8	4.04	1	1	0	1	1	1	1	0
9	3.65	1	1	1	0	1	0	0	0
10	3.77	1	1	1	0	1	0	1	1
11	4.37	1	1	1	0	1	1	0	1
12	4.05	1	1	1	0	1	1	1	0
13	3.45	1	1	1	1	0	0	0	0
14	3.85	1	1	1	1	0	0	1	1
15	3.95	1	1	1	1	0	1	0	1
16	3.88	1	1	1	1	0	1	1	0
.	.	.							
64	4.78	2	2	1	1	1	1	1	1

Since a fractional factorial doesn't include all treatment combinations, it's not possible to estimate all of main effects and interactions. These experiments must be designed carefully so that all the main effects, and all or most of the low-order interactions terms can be estimated. The remaining terms assumed to have zero effects.

The analysis is specified on the preceding page. The results are presented on the next page.

Analysis of Variance Table for Y

Source	DF	SS	MS	F	P
BLOCK	1	0.00391	0.00391		
REP	1	0.05641	0.05641		
BLOCK*REP	1	0.00391	0.00391		
A	1	3.00156	3.00156	324.51	0.0000
B	1	0.57760	0.57760	62.45	0.0000
C	1	2.00223	2.00223	216.47	0.0000
D	1	3.20410	3.20410	346.40	0.0000
E	1	0.50410	0.50410	54.50	0.0000
F	1	1.76226	1.76226	190.52	0.0000
A*B	1	0.03422	0.03422	3.70	0.0639
A*C	1	0.01323	0.01323	1.43	0.2412
A*D	1	0.00160	0.00160	0.17	0.6804
A*E	1	1.000E-04	0.00010	0.01	0.9179
A*F	1	0.04101	0.04101	4.43	0.0437
B*C	1	0.03516	0.03516	3.80	0.0606
B*D	1	0.04101	0.04101	4.43	0.0437
B*E	1	0.01381	0.01381	1.49	0.2313
B*F	1	0.00423	0.00423	0.46	0.5043
C*D	1	0.35701	0.35701	38.60	0.0000
C*E	1	0.01156	0.01156	1.25	0.2725
C*F	1	0.00302	0.00302	0.33	0.5717
D*E	1	0.13876	0.13876	15.00	0.0005
D*F	1	0.04000	0.04000	4.32	0.0462
E*F	1	0.05290	0.05290	5.72	0.0233
A*B*D	1	0.00456	0.00456	0.49	0.4882
A*B*E	1	0.00391	0.00391	0.42	0.5207
A*B*F	1	0.02403	0.02403	2.60	0.1175
A*C*D	1	0.09151	0.09151	9.89	0.0037
A*C*E	1	0.01756	0.01756	1.90	0.1785
A*C*F	1	9.000E-04	0.00090	0.10	0.7572
A*D*E	1	0.04951	0.04951	5.35	0.0277
A*D*F	1	0.04622	0.04622	5.00	0.0330
A*E*F	1	2.500E-05	0.00003	0.00	0.9589
Error	30	0.27749	0.00925		
Total	63	12.4193			

Grand Mean 3.6191 CV 2.66

Since up to 3-way interactions was selected in the dialog box, only main effects, 2-way interactions, and 3-way interactions are included in the AOV table. The higher-order interaction effects are assumed to be zero and the sums of squares are pooled to provide the error sums of squares.

Note that not all the possible 2-way and 3-way interactions are listed. This is a feature of fractional factorial designs. The omitted interactions are aliased with terms that are listed in the table. *Statistix* automatically selects which aliased terms to display, placing a higher priority on low-order terms, and selecting aliased terms of the same order in lexical order. If you'd like to chose different aliased terms, then you must use the General AOV/AOCV and list each term you want included.

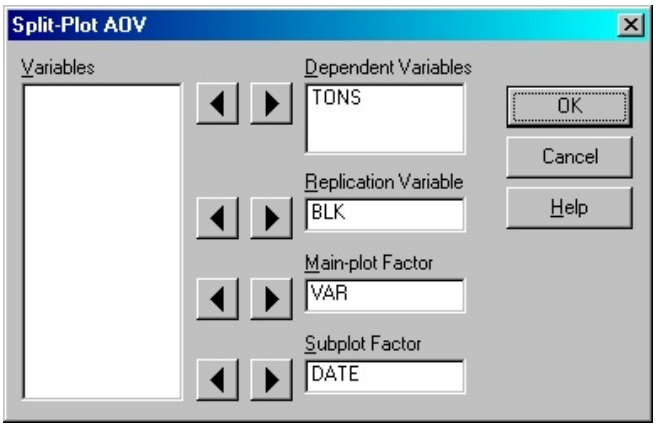
Results Menu

Use the procedures on the results menu to compute means, multiple comparisons, contrasts, plots, and save residuals. These options are discussed in detail at the end of this chapter.

Split-Plot Design

The split-plot design is a two-factor design suitable for experiments that have more treatments than can be accommodated in a complete block design. One of the factors is assigned to the main plot. The main plot is divided into subplots to which the second factor is assigned. The precision of the effects of the main-plot factor is sacrificed to improve the precision of the subplot factor.

Specification



Move the name of the variable containing the observed data to the ***Dependent Variables*** box. If you specify more than one dependent variable, a separate analysis is produced for each variable.

Move the blocking, or replication, variable to the ***Replication Variable*** box. Move the variable name of the main-plot factor to the ***Main-plot Factor box.*** Move the variable name of the subplot factor to the ***Subplot Factor*** box. Press the ***OK*** button to start the analysis.

Data Restrictions

Up to ten dependent variables can be specified. The maximum number of levels for the replication, main-plot factor, and subplot factors are 200 each. The factor variables can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. For unbalanced designs, the maximum size of the GLM design matrix is 500. (The size of the design matrix is equivalent to the model degrees of freedom: total DF minus error DF).

Example

The example is a split-plot design from Section 16.15 of Snedecor and Cochran (1980). TONS is the yield of alfalfa in tons per acre. BLK identifies the six blocks used. VAR is the variety of alfalfa. DATE is the time in days between the second and third cuttings. A partial listing of the data are presented below. The complete data are available in the file Sample Data\alfalfa.sx.

CASE	TONS	VAR	BLK	DATE
1	2.17	1	1	106
2	1.58	1	1	35
3	2.29	1	1	54
4	2.23	1	1	71
5	1.88	1	2	106
6	1.26	1	2	35
7	1.60	1	2	54
8	2.01	1	2	71
9	1.62	1	3	106
10	1.22	1	3	35
11	1.67	1	3	54
12	1.82	1	3	71
. . .				
72	1.33	3	6	71

The second cutting was on July 27. The third cuttings were on September 1, September 20, and October 7. One treatment wasn't cut a third time. We assigned this group the date November 10, intending to reflect the end of the growing season. The values for DATE are 36, 55, 72, and 106. VAR and BLK are qualitative factors, and the actual values of them have meaning only as labels. When possible, factors, such as DATE, should be represented quantitatively because response surfaces can then be examined with polynomial contrasts (see Polynomial Contrasts on page 273).

The analysis is specified in the dialog box on the preceding page. The results are shown below.

Analysis of Variance Table for TONS Tons of alfalfa

Source	DF	SS	MS	F	P
BLK	5	4.14982	0.82996		
VAR	2	0.17802	0.08901	0.65	0.5412
Error BLK*VAR	10	1.36235	0.13623		
DATE	3	1.96247	0.65416	23.39	0.0000
VAR*DATE	6	0.21056	0.03509	1.25	0.2973
Error BLK*VAR*DATE	45	1.25855	0.02797		
Total	71	9.12177			

Grand Mean 1.5968
CV(BLK*VAR) 23.11
CV(BLK*VAR*DATE) 10.47

The split-plot design has two error terms. These are labeled "Error BLK*VAR" and "Error BLK*VAR*DATE" in the AOV table above. The F test for the main–plot factor VAR is computed using the first error term.

The subplot factor DATE and the VAR*DATE interaction are tested using the second error term. The p-value of 0.5412 suggests little difference between varieties. The DATE effect appears to be very significant.

There are two coefficients of variation listed, one for each error term. The first coefficient indicates the degree of precision for the main-plot factor. The second coefficient indicates the precision for the subplot factor and the VAR*DATE interaction.

Results Menu

Use the procedures on the results menu to compute means, multiple comparisons, contrasts, plots, and save residuals. These options are discussed in detail at the end of this chapter.

Computational Notes

Oliver's (1967) generalization of Yates' algorithm (Daniel, 1976) is used for balanced designs. Unbalanced designs are computed using general linear models (Searle, 1987; Glantz and Slinker, 1990).

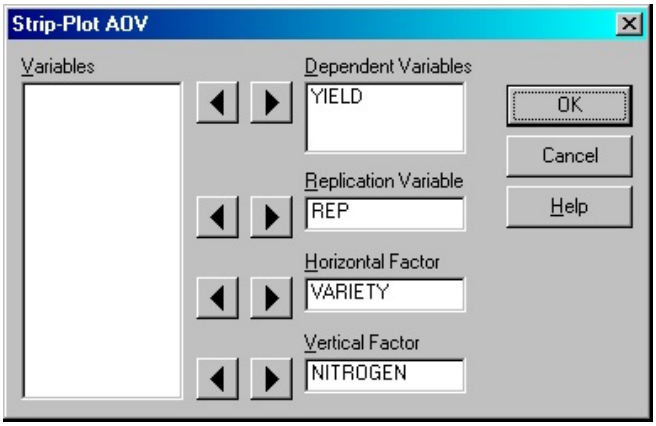
Strip-Plot Design

The strip-plot design is a two-factor design that's useful when the desired precision for the two-factor interaction is greater than that of either main effect. The design calls for a horizontal-strip factor, a vertical-strip factor, and an intersection plot for the interaction of the two factors.

Specification

The strip-plot dialog box is shown on the next page. Move the name of the variable containing the observed data to the ***Dependent Variables*** box. If you specify more than one dependent variable, a separate analysis is produced for each variable.

Specify the model by moving the variables for replication, the horizontal-strip factor, and the vertical-strip factor to the corresponding ***Replication Variable***, ***Horizontal Factor***, and ***Vertical Factor*** boxes. Press the ***OK*** button to start the analysis.



Data Restrictions

Up to ten dependent variables can be specified. The maximum number of levels for the replication, horizontal, and vertical factors are 200 each. The factor variables can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. For unbalanced designs, the maximum size of the GLM design matrix is 500.

Example

The example data are from a yield trial of six varieties of rice and three levels of nitrogen fertilizer in a strip-plot design with three replications (Gomez and Gomez, 1984). A partial listing of the data are presented below. The complete data can be found in Sample Data\strip-plot.sx.

CASE	YIELD	REP	VARIETY	NITROGEN
1	2373	1	IR8	0
2	4076	1	IR8	60
3	7254	1	IR8	120
4	4007	1	IR127-80	0
5	5630	1	IR127-80	60
6	7053	1	IR127-80	120
7	2620	1	IR305-4-12	0
8	4676	1	IR305-4-12	60
9	7666	1	IR305-4-12	120
10	2726	1	IR400-2-5	0
11	4838	1	IR400-2-5	60
12	6881	1	IR400-2-5	120
. . .				
54	3214	3	Peta	120

The analysis is specified in the dialog box above. The results are shown on the next page.

Analysis of Variance Table for YIELD Grain yield

Source	DF	SS	MS	F	P
REP	2	9220962	4610481		
VARIETY	5	5.710E+07	1.142E+07	7.65	0.0034
Error REP*VARIETY	10	1.492E+07	1492262		
NITROGEN	2	5.068E+07	2.534E+07	34.07	0.0031
Error REP*NITROGEN	4	2974908	743727		
VARIETY*NITROGEN	10	2.388E+07	2387798	5.80	0.0004
Error REP*VARIETY*NITROGEN	20	8232917	411646		
Total	53	1.670E+08			

Grand Mean 5289.9
  CV(REP*VARIETY) 23.09
  CV(REP*NITROGEN) 16.30
  CV(REP*VARIETY*NITROGEN) 12.13

The strip-plot design has three error terms. The F test for the horizontal-strip factor VARIETY is computed using the first error term REP*VARIETY. The vertical-strip factor NITROGEN is tested using the error term REP*NITROGEN. The interaction VARIETY*NITROGEN is tested using the REP*VARIETY*NITROGEN error term.

Results Menu

Use the procedures on the results menu to compute means, multiple comparisons, contrasts, plots, and save residuals. These options are discussed in detail at the end of this chapter.

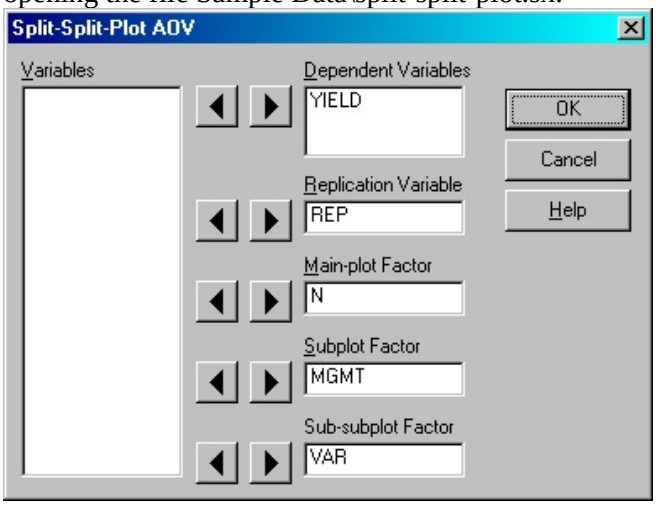
Split-Split-Plot Design

The split-split-plot design is an extension of the split-plot design to accommodate a third factor. There are three plots sizes: the main plot, the subplot, and the sub-subplot. There are three levels of precision: the main-plot factor has the lowest, and the sub-subplot factor having the highest degree of precision.

Example

The example data are from a yield trial of rice in a split-split-plot design with three replications (Gomez and Gomez, 1984, p. 143). The main-plot factor is nitrogen (N), the subplot factor is management practice (MGMT), and the sub-subplot factor is variety (VAR). You can view the data by

opening the file Sample Data\split-split-plot.sx.



The model is specified in the dialog box above. The results are shown below.

Analysis of Variance Table for YIELD

Source	DF	SS	MS	F	P
REP	2	0.732	0.366		
N	4	61.641	15.410	27.70	0.0001
Error REP*N	8	4.451	0.556		
MGMT	2	42.936	21.468	82.00	0.0000
N*MGMT	8	1.103	0.138	0.53	0.8226
Error REP*N*MGMT	20	5.236	0.262		
VAR	2	206.013	103.007	207.87	0.0000
N*VAR	8	14.145	1.768	3.57	0.0019
MGMT*VAR	4	3.852	0.963	1.94	0.1149
N*MGMT*VAR	16	3.699	0.231	0.47	0.9538
Error REP*N*MGMT*VAR	60	29.732	0.496		
Total	134	373.541			

Grand Mean 6.5544
CV(REP*N) 11.38
CV(REP*N*MGMT) 7.81
CV(REP*N*MGMT*VAR) 10.74

Note the three error terms. The terms that use the mean square of each error term as the denominator for the F test are listed immediately above them in the table.

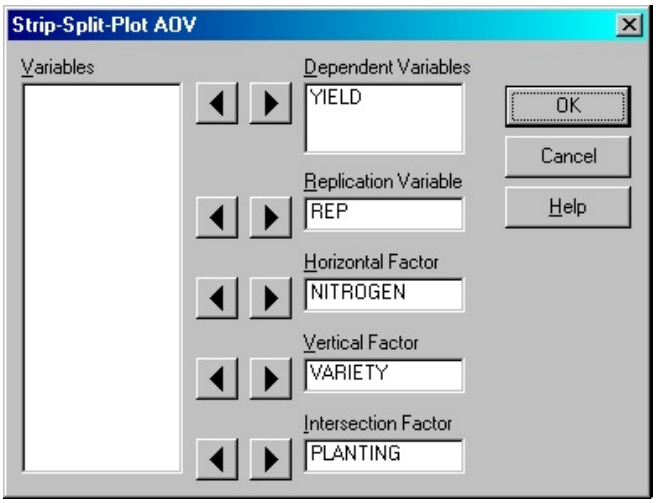
Computational Notes

Oliver's (1967) generalization of Yates' algorithm (Daniel, 1976) is used for balanced designs. Unbalanced designs are computed using general linear models (Searle, 1987; Glantz and Slinker, 1990).

Strip-Split-Plot Design

The strip-split-plot design is an extension of the strip-plot design to accommodate a third factor. The intersection plot of the strip-plot design is divided into subplots for the third factor. There are four plots sizes: the horizontal strip, the vertical strip, the intersection plot, and the subplot. There are four levels of precision with the subplot factor have the highest degree of precision.

Example



The example data are from a yield trial of rice (Gomez and Gomez, 1984, p.155). The treatment factors are nitrogen, variety, and planting method (broadcast vs. transplanted). You can view the data by opening the file Sample Data\strip-split-plot.sx. The model is specified in the dialog box above. The results are shown on the next page.

Note that there are four error terms. The terms that use the mean square of each error term as the denominator for the F test are listed immediately above them in the table.

Analysis of Variance Table for YIELD

Source	DF	SS	MS	F	P
REP (A)	2	1.530E+07	7653156		
NITROGEN (B)	2	1.165E+08	5.827E+07	36.65	0.0027
Error A*B	4	6359988	1589997		
VARIETY (C)	5	4.909E+07	9818698	3.67	0.0380
Error A*C	10	2.673E+07	2672583		
B*C	10	2.461E+07	2461442	2.58	0.0344
Error A*B*C	20	1.911E+07	955675		
PLANTING (D)	1	726028	726028	1.72	0.1982
B*D	2	2467935	1233968	2.92	0.0668
C*D	5	2.376E+07	4751761	11.25	0.0000
B*C*D	10	7513641	751364	1.78	0.1007
Error A*B*C*D	36	1.521E+07	422560		
Total	107	3.074E+08			

Grand Mean 5371.6
CV(REP*NITROGEN) 23.47
CV(REP*VARIETY) 30.43
CV(REP*NITROGEN*VARIETY) 18.20
CV(REP*NITROGEN*VARIETY*PLANTING) 12.10

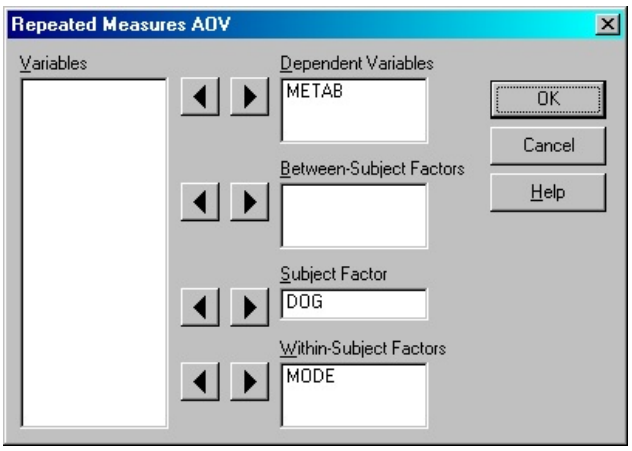
Repeated Measures Design

For the analysis of variance designs discussed earlier in the chapter, an individual experimental unit, whether it be a test animal or a plot of land, is assigned to a single treatment and the response variable is measured only once. For a repeated measures analysis of variance, an individual experimental unit, or subject, is observed under several different levels of one or more experimental factors. This procedure computes the analysis of variance for a variety of repeated measures designs.

Specification

In order to specify a repeated measures design, it's important to understand the distinction between a between-subject factor and a within-subject factor. A between-subject factor is an experimental treatment whose effect is estimated by observing differences between subjects. The pool of available subjects are divided into one group for each level of a between-subjects factor, so an individual subject has only one level of the treatment applied to it. A within-subject factor is one whose effect is estimated by observing differences within subjects. All subjects have each level of the treatment applied to them.

A repeated measures design doesn't require a between-subjects factor, but it must have at least one within-subjects factor. The simplest repeated measures design is the one-way repeated measures analysis of variance. In this design, each subject is observed over the various levels of a single experimental factor (see example I below).



Move the name of the variable containing the observed data to the ***Dependent Variables*** box. If you specify more than one dependent variable, a separate analysis is produced for each variable.

If your model includes any between-subject factors, move the variable for those factors to the ***Between-Subject Factors*** box. Move the variable that identifies subjects to the ***Subject Factor*** box. Move the variables that identify within-subject factors to the ***Within-Subject Factor*** box. Press the ***OK*** button to start the analysis.

Data Restrictions

Up to ten dependent variables can be specified. Up to three between-subject factors and three within-subject factors can be specified. All factors are limited to 200 levels each. The factor variables can be of any data type.

Example I - One-way RM AOV

The example data are from a study to examine the effects of two stages of digestion on the metabolism of dogs (Glantz and Slinker, 1990, p. 392). The first stage is the act of eating the food (smelling, chewing, and tasting). The second stage is the digestion that occurs in the stomach. The metabolic rate was observed after meals were provided to six dogs in three different manners: (1) eating normally, (2) placing the food in the mouth but

bypassing the stomach, and (3) bypassing the mouth by placing the food directly into the stomach. The data are shown in the table below, and are stored in the file Sample Data\eating.sx.

CASE	METAB	DOG	MODE	CASE	METAB	DOG	MODE
1	104	1	Normal	10	114	4	Normal
2	91	1	Mouth	11	106	4	Mouth
3	22	1	Stomach	12	15	4	Stomach
4	106	2	Normal	13	117	5	Normal
5	94	2	Mouth	14	120	5	Mouth
6	14	2	Stomach	15	18	5	Stomach
7	111	3	Normal	16	139	6	Normal
8	105	3	Mouth	17	111	6	Mouth
9	14	3	Stomach	18	8	6	Stomach

The model is specified in the dialog box on the preceding page. The results are presented below.

Analysis of Variance Table for METAB

Source	DF	SS	MS	F	P
DOG	5	572.9	114.6		
MODE	2	36188.4	18094.2	198.35	0.0000
Error	10	912.2	91.2		
Total	17	37673.6			

Grand Mean 78.278 CV 12.20

Tukey's 1 Degree of Freedom Test for Nonadditivity

Source	DF	SS	MS	F	P
Nonadditivity	1	523.522	523.522	12.12	0.0069
Remainder	9	388.700	43.189		

Means of METAB for MODE

MODE	Mean
Normal	115.17
Mouth	104.50
Stomach	15.17

Observations per Mean 6
Standard Error of a Mean 1.9746
Std Error (Diff of 2 Means) 2.3483

The F test for the effect of eating modes in the AOV table above is highly significant. Placing food directly into the stomachs of the dogs resulted in a lower metabolic rate compared to the other two modes of eating.

Example II - Two-way RM AOV

The following example is a two-way repeated measures with one between-subject factor and one-within subject factor. It compares the effects of alcohol on people diagnosed with antisocial personality disorder (ASP) and those that don't have ASP (Glantz and Slinker, 1990, p. 410). Personality type (ASP and non-ASP) is a between-subject factor (a subject has either one personality type or the other). Subjects are given alcohol to drink, and

the aggressiveness is evaluated before and after drinking. Drinking status (sober and drinking) is a within-subject factor. The data are listed below, and are stored the file Sample Data\alcohol.sx.

AGGRESS	P_TYPE	SUBJECT	DRINK	AGGRESS	P_TYPE	SUBJECT	DRINK
0.81	Non-ASP	1	Sober	0.72	ASP	1	Sober
0.59	Non-ASP	1	Drinking	0.83	ASP	1	Drinking
0.91	Non-ASP	2	Sober	0.82	ASP	2	Sober
1.04	Non-ASP	2	Drinking	0.99	ASP	2	Drinking
0.98	Non-ASP	3	Sober	0.89	ASP	3	Sober
1.11	Non-ASP	3	Drinking	1.17	ASP	3	Drinking
1.08	Non-ASP	4	Sober	1.01	ASP	4	Sober
1.13	Non-ASP	4	Drinking	1.24	ASP	4	Drinking
1.10	Non-ASP	5	Sober	1.10	ASP	5	Sober
1.15	Non-ASP	5	Drinking	1.33	ASP	5	Drinking
1.16	Non-ASP	6	Sober	1.14	ASP	6	Sober
1.16	Non-ASP	6	Drinking	1.47	ASP	6	Drinking
1.19	Non-ASP	7	Sober	1.24	ASP	7	Sober
1.25	Non-ASP	7	Drinking	1.59	ASP	7	Drinking
1.44	Non-ASP	8	Sober	1.34	ASP	8	Sober
1.70	Non-ASP	8	Drinking	1.73	ASP	8	Drinking

The dependent variable AGGRESS is a score for aggressiveness obtained from a questionnaire. Note that although there were a total of 16 subjects in the study, the values for the variable SUBJECT are numbered 1 through 8 within each personality type. The results are shown below.

Analysis of Variance Table for AGGRESS Aggressiveness

Source	DF	SS	MS	F	P
P_TYPE	1	0.02050	0.02050	0.16	0.6921
Error P_TYPE*SUBJECT	14	1.75589	0.12542		
DRINK	1	0.20320	0.20320	29.23	0.0001
P_TYPE*DRINK	1	0.08303	0.08303	11.94	0.0039
Error P_TYPE*SUBJECT*DRINK	14	0.09732	0.00695		
Total	31	2.15995			

Grand Mean 1.1378
CV(P_TYPE*SUBJECT) 31.13
CV(P_TYPE*SUBJECT*DRINK) 7.33

Note that there are two error terms listed in the AOV table. Personality type is tested using the subject within personality type term P_TYPE*SUBJECT. DRINK and the two-factor interaction are tested using the P_TYPE*SUBJECT*DRINK term.

The test for the P_TYPE*DRINK interaction is significant. Examining this interaction using the **Means Plot** available on the results menu clearly illustrates that while drinking increases aggressiveness, the increase is much more dramatic for the ASP personality type.

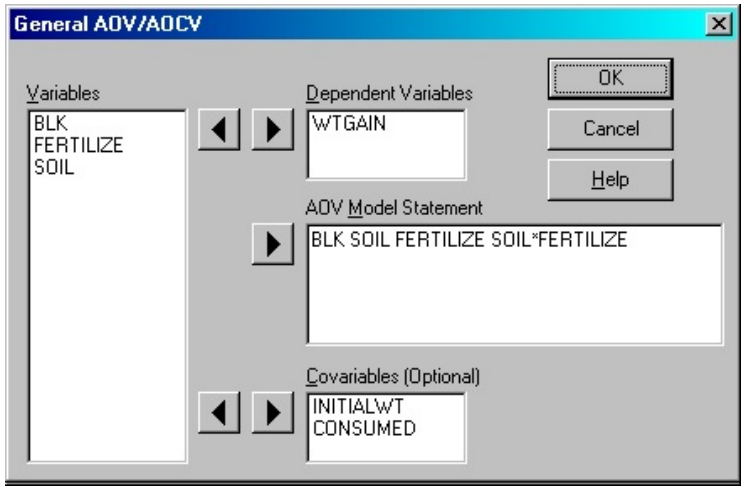
Missing Values in Repeated Measures

Missing values are allowed in one-way repeated measures analysis of variance models. The presence of missing values is limited in multi-factor repeated measures designs. *Statistix* allows the number of subjects to be different for different levels of a between-subjects factor. But individual subjects must have valid observations recorded for all levels of the within-subject factors.

General AOV/AOCV

The **General AOV/AOCV** procedure is a flexible procedure you can use to analyze many analysis of variance and covariance designs, including ones that are unbalanced. Models are specified by explicitly entering a model statement. Many options are available—mean estimation, multiple comparisons of means, general and polynomial contrasts, residual plots, and least squares estimation of missing values.

Specification



The dependent variable contains the observed data. Select the dependent variable from the ***Variables*** list box and move it to the ***Dependent Variable***s box. If you specify more than one dependent variable, a separate analysis is produced for each variable.

Models are specified in a manner similar to the usual algebraic expression of analysis of variance models, such as those illustrated in Snedecor and Cochran (1980). Use the factor variables to list the terms in your model in the ***AOV Model Statement*** box. The terms are main effects, which are entered as a single variable, and interaction terms, which are listed as a group of factor variables combined using stars (e.g., TRT∗BLK). The high order interaction term is assumed to be an error term and can often be omitted from the model statement. Other interaction terms can be indicated as an error term by typing (ERROR) or (E) after the term (see examples 5 and 6 below).

Typically, the AOV Model Statement is simply typed in manually. You can also copy variables from the ***Variables*** list box to the current cursor position of the ***AOV Model Statement*** by first highlighting one or more variables in the Variables list, then pressing the right-arrow button next to the AOV Model Statement.

For analysis of covariance, select the names of the variables you want to use as covariates and move them to the ***Covariables*** list box.

Press the ***OK*** button to begin the analysis. You'll be offered additional options once the analysis is specified and computed.

Example Model Statements

Model specification (list of main effects and interaction terms) is very flexible and best illustrated by example.

Example 1: Completely randomized design, also called the one-way design (see page 227 for a discussion of the CRD). If the treatment factor is A, the model is specified simply as:

```
A
```

Example 2: Randomized complete block design (see page 229 for a discussion of the RCB design). BLK is the factor for blocks and A if the treatment factor.

```
BLK A
```

Example 3: Single-factor Latin square design (see page 232 for a discussion of the Latin square design). The variables ROW and COL are the row- and column blocking factors, and A is the treatment factor.

```
ROW COL A
```

Example 4: Three-factor factorial in a completely randomized design with all two factor interactions (see page 238 for a discussion of factorial designs).

```
A B C A*B A*C B*C
```

Note: The example factorial design above could be entered more concisely using the ALL2 keyword discussed below.

```
ALL2(A B C)
```

Example 5: Split-plot design (see page 241 for a discussion of the split-plot design). The variable REP is the factor for replication, A is the main-plot factor, and B is the subplot factor.

```
REP A REP*A(E) B A*B
```

Note the interaction REP*A is an error term. The three factor interaction term REP*A*B is also an error term, but was omitted above because *Statistix* always adds the high order interaction term automatically.

Example 6: Strip-plot design (see page 243 for a discussion of the strip-plot design). The variable REP is the factor for replication, A is the main-plot factor, and B is the subplot factor.

```
REP A REP*A(E) B REP*B(E) A*B
```

Example 7: One-way repeated measures design (see page 248 for a discussion of repeated measures designs). SUBJ is the subjects factor and A is the within-subjects factor.

```
SUBJ A
```

Example 8: Two-factor repeated measures design with a between-subjects factor A and a within-subjects factor B.

```
A SUBJ*A(E) B A*B
```

Examples 1 - 8 above are all models that could be more easily specified using the specific AOV procedures discussed earlier in this chapter. But the General AOV/AOCV procedure allows you to specify other models, including variations of the above models.

Example 9: A two-factor nested model with factor B nested within factor A.

```
A B*A
```

Compare the nested model with the two-factor cross-classified (factorial) model:

```
A B A*B
```

Example 10: A three-factor factorial experiment in a split-plot design. The factors A and B are both main-plot factors, and C is the subplot factor.

```
REP A B A*B REP*A*B(E) C A*C B*C A*B*C
```

In most situations, neither the order in which terms are specified in the model nor the order in which variables are specified within terms has any influence on the analysis. The only exception occurs in certain models with multiple error terms, which are described next.

Error Term Specification

The (ERROR) and (E) modifiers behind a term indicates that the term is to be used as an error term for computing F tests. Note that multiple error terms can be specified, as in Examples 5, 6, 8, and 10 above. If the highest order interaction is not explicitly listed in the model, it's automatically added and is an error term. For instance, in Example 5, the results would be exactly the same if the model had been specified as:

```
REP A REP*A(E) B A*B REP*A*B(E)
```

When multiple error terms are specified, an F test for a main effect or interaction is based on the lowest order error term that includes the main effect or interaction being tested. The lines of an analysis of variance table are organized so that an error term appears directly below the group of terms that use it for the F tests. See the sections for the split-plot design and repeated measures design in this chapter for examples of analysis of variance tables for models with multiple error terms.

Occasionally, there will be error terms of equal order that contain the term for which a test is desired. For example, consider the model:

```
X1 X2 X1*X2(ERROR) X3 X2*X3(ERROR) X1*X3 X1*X2*X3(ERROR)
```

The second order error terms X1∗X2 and X2∗X3 both contain X2. Which term will be used to construct the F test for X2? *Statistix* scans the model from left to right. When it encounters the first factor (X1), it assigns it the name "A". The next factor encountered (X2) is assigned the name "B", and so on. The error terms X1∗X2 and X2∗X3 are represented as AB and BC, respectively. To decide which term to use, *Statistix* always uses the one with the lowest dictionary order, AB in this case.

ALL Term Specification

The ALL modifier is provided to simplify model specifications when you want all subset interactions. For example, the model

```
A ALL (B C D)
```

is equivalent to

```
A B C D B*C B*D C*D B*C*D
```

You can also use the ALL2 modifier to specify all terms up to and including 2-factor interaction terms. So the model statement

```
ALL2 (A B C)
```

is equivalent to

```
A B C A*B A*C B*C
```

The modifiers ALL3, ALL4, and ALL5 are defined in a similar manner.

Data Restrictions

You can specify up to ten dependent variables. Up to ten factors (control and treatment variables) can be included in a model statement. The maximum number of levels for each factor is 200. The factor variables can be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. Missing values are allowed. For unbalanced designs and designs with covariates, the maximum size of the GLM design matrix is 500. (The size of the design matrix is equivalent to the model degrees of freedom: total degrees of freedom minus error degrees of freedom.)

Pooling of Sums of Squares

Internally, the AOV is initially treated as if it's a full factorial design; sums of squares are computed for all possible terms. Then if a term isn't included in a model, the sums of squares calculated for that term is pooled in the lowest order interaction that contains that term as a subset. For example, in the model A B C A∗B∗C, the sums of squares for A∗B, A∗C, and B∗C are pooled with the A*B*C sums of squares.

Example

The example data are from a two-factor factorial in a randomized block design with two covariates (Steel and Torrie, 1980, p. 429). The object of the experiment was to study how forage fed to guinea pigs from four types

of soil at two levels of fertilization affected weight gain. Animals were selected for blocks based on initial weight. The initial weights of each subject were also used as a covariate for error control and to adjust the means. The second covariate, forage consumed, is affected by the treatments and is included to help interpret the data. The data are listed below and are stored in the file Sample Data\forage.sx.

INITIALWT	CONSUMED	WTGAIN	BLK	SOIL	FERTILIZE
220	1155	224	1	1	1
222	1326	237	1	1	2
198	1092	118	1	2	1
205	1154	82	1	2	2
213	1573	242	1	3	1
188	1381	184	1	3	2
256	1532	241	1	4	1
202	1375	239	1	4	2
246	1423	289	2	1	1
268	1559	265	2	1	2
266	1703	191	2	2	1
236	1250	117	2	2	2
236	1730	270	2	3	1
259	1363	129	2	3	2
278	1220	185	2	4	1
216	1170	207	2	4	2
262	1576	280	3	1	1
314	1528	256	3	1	2
335	1546	115	3	2	1
268	1667	117	3	2	2
288	1593	198	3	3	1
300	1564	212	3	3	2
283	1232	185	3	4	1
225	1273	227	3	4	2

The analysis is specified on page 252. The results are shown below.

Analysis of Variance Table for WTGAIN

Source	DF	SS	MS	F	P
BLK	2	395.4	197.7	0.46	0.6408
SOIL	3	59216.4	19738.8	46.12	0.0000
FERTILIZE	1	1850.6	1850.6	4.32	0.0597
SOIL*FERTILIZE	3	1136.6	378.9	0.89	0.4764
INITIALWT	1	1341.6	1341.6	3.13	0.1020
CONSUMED	1	10585.1	10585.1	24.73	0.0003
Error	12	5135.5	428.0		
Total	23				

Note: SS are marginal (type III) sums of squares

Grand Mean 200.42 CV 10.32

Covariate Summary Table

Covariate	Coefficient	Std Error	T	P
INITIALWT	-0.49430	0.27918	-1.77	0.1020
CONSUMED	0.15837	0.03184	4.97	0.0003

The F tests are significant for both soil types and fertilizer level. In an analysis with covariates, the F tests test the adjusted means. For example, the null hypothesis that the means for soil type, adjusted for initial weight and forage consumed, are equal is rejected (p = 0.0000).

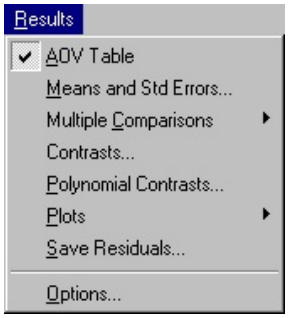
The covariates INITIALWT and CONSUMED also appear in the analysis of variance table. The F tests for the covariates test the hypothesis that the regression coefficients for the covariates are zero, given that all other analysis of variance and covariance terms are in the model. The t tests listed in the covariate summary test the same hypothesis. Initial weight is not significant, but nearly so ($p = 0.1020$). Foraged consumed is highly significant ($p = 0.0003$). The negative coefficient for INITIALWT means that guinea pigs with lower initial weights gained more weight than those with higher initial weights. Weight gain increased with forage consumed.

Computational Notes

Oliver's (1967) generalization of Yates' algorithm (Daniel, 1976) is used for balanced designs. An algorithm similar to Cooper's (1968) is used to generate orthogonal polynomials. Unbalanced designs are computed using general linear models (Searle, 1987; Glantz and Slinker, 1990).

AOV Results Menu

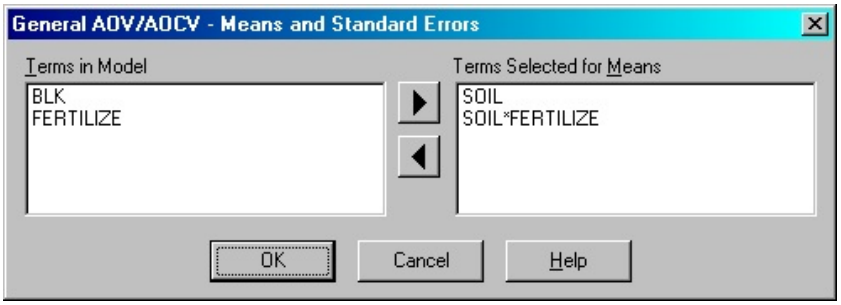
After the initial analysis of variance is completed and displayed, a ***Results*** menu appears on the menu at the top of the *Statistix* window. This menu is displayed below.



This results menu is available for all the analysis of variance procedures discussed in this chapter. Select AOV Table from the menu to redisplay the initial AOV results. Select Options to return to the dialog box used to specify the model. The remaining results options are described on the following pages.

Single-factor procedures, such as the **Completely Randomized Design** and the **Randomized Complete Block Design** display a table of treatment means along with the analysis of variance table. Select the **Means and Standard Errors** procedure from the results menu to compute least squares means for the remaining procedures.

The dialog box list terms in the model available for computing means and standard errors. Error terms aren't listed. The dialog box for the guinea pig forage example discussed on page 256 is shown below.



The main effect SOIL and the two-factor interaction SOIL*FERTILIZE have been selected and moved to the ***Terms Selected for Means*** box. The results are shown below.

```
Means of WTGAIN for SOIL

SOIL   N     Mean       SE
   1   6   259.60   8.5942
   2   6   126.55   8.4855
   3   6   186.16   9.3343
   4   6   229.36   9.1457

Means of WTGAIN for SOIL*FERTILIZE

SOIL FERTILIZE   N     Mean       SE
   1         1   3   266.01   12.079
   1         2   3   253.19   12.918
   2         1   3   144.83   12.770
   2         2   3   108.26   12.445
   3         1   3   200.65   13.987
   3         2   3   171.67   11.963
   4         1   3   228.98   14.269
   4         2   3   229.74   15.153
```

The means are least squares estimates based on the model, so they won't be the same as the arithmetic means for unbalanced designs or designs with covariates. The means for analyses with covariates, such as those in this

example, are adjusted for the covariates.

The standard error of a mean is computed using the mean square for error from the error term associated with it from the original AOV table. The calculations for the standard errors incorporate the sample sizes and the covariate means, which explains why the standard errors aren't all the same in this example. For balanced designs without covariates, the standard errors are all the same, and the report format is different from the one above.

The standard error for the difference of two means is included in the report for balanced designs without covariates.

Multiple Comparisons

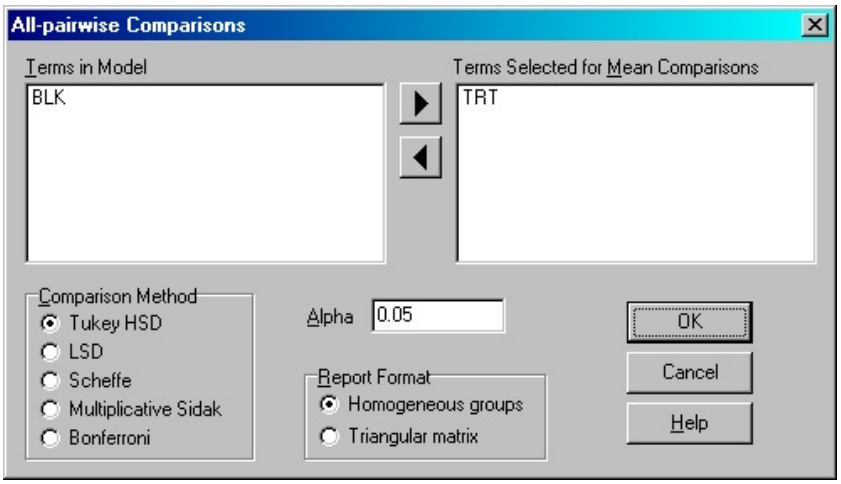
You will often be interested in comparing means for different levels of a main effect or interaction. This is the function of multiple comparisons procedures. Multiple comparisons are divided into three categories: all-pairwise comparisons, comparisons with a control, and comparisons with the best. Statistix offers tests for all three types of comparisons.

All-pairwise Multiple Comparisons

Two means are said to be similar, or homogeneous, if they're not significantly different from one another. This procedure identifies groups (subsets) of similar, or homogeneous, means. Use of the procedure is first illustrated with an example, and then some details of its application are discussed.

We'll use the randomized complete block example from page 230, where the treatment factor TRT was the type of fungicide applied to batches of 100 soybeans. TRT has five levels; the first is a no-fungicide control, and

the remaining four are different types of fungicide. The dependent variable FAILURES is the number of beans out of 100 that failed to sprout.



First select the main effects and/or interactions for which you want comparisons and move them to the ***Terms Selected for Mean Comparisons*** box. Next select a comparison method by clicking on one of the five ***Comparison Method*** radio buttons. Enter a value for the rejection level in the ***Alpha*** edit control. Select the report format (examples of both formats are given below).

The example dialog box above is used to compute Tukey's HSD comparisons for treatments using 0.05 for the rejection level. The results are shown below.

```
Tukey HSD All-Pairwise Comparisons Test of FAILURES for TRT

TRT         Mean  Homogeneous Groups
Control   10.800  A
Fung #2    8.200  AB
Fung #3    6.600  AB
Fung #1    6.200   B
Fung #4    5.800   B

Alpha              0.05     Standard Error for Comparison  1.4711
Critical Q Value   4.333    Critical Value for Comparison  4.5072
Error term used: BLK*TRT, 16 DF
There are 2 groups (A and B) in which the means
are not significantly different from one another.
```

The second column of the results shows the means of FAILURES for the levels of the factor TRT. The means have been sorted in descending order so the largest one is listed in the first row, the next to largest in the second

row, and so on.

The columns of the letters A and B under the heading "Homogenous Groups" indicate which means are not significantly different from one another. There are two columns in the example since there are two groups of similar or homogenous means. The first group contains the means for the control and fungicides 2 and 3. The second group contains the means for fungicides 1 and 4. As you see in this example, it's not unusual for the groups to overlap, although they need not. There are two pairs of means that are different in this example. The mean for the control group is not in group B, and the means for fungicides 1 and 4 is not in group A. So the mean of the control are different than the means for fungicides 1 and 4.

Many people prefer the triangular matrix report format because it's easier to identify means that are significantly different. This report format for the same data is shown below.

```
Tukey HSD All-Pairwise Comparisons Test of FAILURES for TRT

TRT         Mean Control  Fung #1  Fung #2  Fung #3
Control   10.800
Fung #1    6.200   4.600*
Fung #2    8.200   2.600    2.000
Fung #3    6.600   4.200    0.400    1.600
Fung #4    5.800   5.000*   0.400    2.400    0.800

Alpha              0.05     Standard Error for Comparison  1.4711
Critical Q Value   4.333    Critical Value for Comparison  4.5072
Error term used: BLK*TRT, 16 DF
```

The treatment levels are listed along the top and down the left side of the table. The values in the body of the table are differences between pairs of means. The critical value for a comparison, 4.5072, is the minimum difference between two means needed for significance. Pairs that are significantly different are flagged with an asterisk.

It's easy to use this figure to construct confidence intervals for the differences of any two means. Suppose you were interested in 95% confidence bounds around the difference of fungicides 2 and 4. Simple subtract and add the critical value of the comparison from the difference: 2.400 ± 4.5072 = -2.1072 to 6.9072. Note that the confidence interval contains zero, which is expected since difference was not significant.

All-pairwise Comparison

It's important you understand the difference between (1) the hypotheses being tested by the overall F test for a main treatment effect in analysis of

variance and (2) the hypotheses being tested by pairwise comparisons procedures. A contrast is any linear combination of treatment means such that the linear coefficients sum to zero (see **Contrasts** on page 269). A pairwise comparison of two means is a special case of a contrast where the contrast coefficients are 1 and -1 for the means being compared, and 0 for all other means.

The overall F test for a treatment effect in AOV is testing the hypothesis that all of the means are equal. You can think of it as a test of whether all possible contrasts are zero. If the overall F is significant, it means that there is **some** contrast that's significant, but it doesn't guarantee that any pairwise comparison is particularly important. The set of pairwise comparisons is a small subset of the entire set of all possible contrasts. The F test has to be a conservative test because it must guard against type I errors (rejecting a null hypothesis when it's true) over the entire set of all possible contrasts, not just the smaller subset of pairwise comparisons. If you're interested only in the set of pairwise comparisons, you can construct a more powerful test than the overall F test over this restricted space.

If you're interested in a single comparison of two means, the most powerful procedure is the T test. For a single such comparison, the probability of falsely rejecting a true null hypothesis (type I error) is whatever the significance level of the T test is. Suppose that there are two comparisons of interest to you. Suppose you test each one at the level α with a T test. The probability of making a type I error in each comparison is α, so the probability of making at least one type I error over both comparisons is greater than α. As the number of comparisons grows, the probability of making at least one type I error grows toward 1. This probability of making at least one type I error for all comparisons is called the experimentwise error rate, in contrast to the comparisonwise error rate. The T test controls the comparisonwise error rate at α but allows the experimentwise error rate to increase as the number of comparisons increases. Experimentwise error rate refers to the maximum experimentwise error rate under a complete or partial null hypothesis. Under a complete null hypothesis, all the population means are equal; under a partial null hypothesis, only some of the population means are equal.

If there are P means, there are $m = P(P-1)/2$ pairwise comparisons, so the number of comparisons grows rapidly as the number of means increases. Some control over the experimentwise error rate is desirable. Numerous methods have been proposed for this, and there is some disagreement as to the best procedures. The following discussion describes the procedures

available in *Statistix*.

First, some terminology: Suppose M_i and M_j are two means. The comparison between means I and j is $L_{ij} = M_i - M_j$. For a complete, balanced AOV, the standard error of L_{ij} is $SE(L_{ij}) = (2 * MSE/n)^{½}$, where MSE is the mean square for error and n is the number of samples present at a level of the factor of interest. For the comparison L_{ij} to be significant, its absolute value must exceed some critical value C, where C depends on the method of comparison being used. Confidence intervals for a comparison are computed as $L_{ij} \pm C$.

The most powerful (least conservative) comparison procedure is the **LSD**, or Least Significant Difference method. The critical value for a comparison is $SE(L_{ij})$ T, where T is Student's t-statistic for the degrees of freedom associated with MSE. This method is also called the T method. LSD controls the comparisonwise error rate at α but allows the experimentwise error rate to increase as the number of comparisons increases. Some advocate using this method only if the overall F test is significant, leading to what has sometimes been called the PSD, or Protected Significant Difference. Contrary to what has sometimes been claimed, the PSD method does not control the experimentwise error rate if there are more than three levels for the factor of interest.

As we noted earlier, the overall F test is testing a much broader range of hypotheses than a multiple comparison test and so it must be more conservative. If the set of pairwise comparisons are of primary interest, then the so-called protected approach (proceeding only if the overall F is significant) can be refuted to some extent regardless of the comparison method because the F test sacrifices power to test hypotheses that are not of direct interest. However, such cases are probably exceptions rather than the rule.

You should use the LSD method if there are a **few** preplanned comparisons that are of primary interest. However, inspecting the means for large differences before deciding which comparisons to make invalidates its use. The LSD procedure is the most powerful pairwise comparison procedure, but it will generally have the highest experimentwise error rate. We mentioned earlier that the LSD approach controls the comparisonwise error rate at α. If you use the LSD method and report significant comparisons, you should be prepared to justify why you didn't find it necessary to control the experimentwise error rate.

The LSD procedure can be modified to prevent the experimentwise error

from growing as the number of comparisons increases. The general idea is to make it more difficult to reject as the number of comparisons increases, which can be done by increasing the critical value of T as the number of comparisons increases. Suppose T(p) is the T value corresponding to a two-tailed significance level of p for Student's t. For the LSD procedure, p is the constant α. To control the experimentwise error rate, p should be some decreasing function of m, where m is the number of comparisons. Two common methods for this are **Bonferroni's** and **Sidak's**. Bonferroni's—probably the more popular of the two—uses the function $p = \alpha/m$, and Sidak's uses the function $p = 1 - (1 - \alpha)^{1/m}$. Using either of these methods results in an experimentwise error rate of less than α. The problem with these procedures is they rapidly grow conservative as m increases; in effect, the experimentwise error rate is reduced too much and real differences do not get detected (test power is lost). Bonferroni's is generally more conservative than Sidak's. Because of rapidly decreasing power, these procedures are not recommended for general use although they can be useful when the number of means, and hence the number of comparisons, is small.

Tukey's method is the most useful pairwise comparison procedure *Statistix* performs. It controls the experimentwise error rate, yet still retains good power. It's based on the Studentized range statistic. Suppose there are P means for the factor of interest, with $X_{(1)}$ being the smallest and $X_{(P)}$ being the largest. The standard error of a mean is $(MSE/n)^{½}$, where MSE is the mean square for error and n is the number of samples within each level. (In terms of $SE(L_{ij})$, which is displayed by *Statistix*, the standard error of a mean is $SE(Lij)/(2^{1/2})$.) Under the usual assumptions, the statistic $(X_{(1)}-X_{(P)}) / (MSE/n)^{½}$ then has a Studentized range distribution if there are no differences between the population means. The critical value for a comparison L_{ij} is $C = (MSE/n)^{½} Q(P,DF)$, where Q(P,DF) is the Studentized range value for P means and DF degrees of freedom (degrees of freedom associated with MSE) at the desired rejection level α. Tukey's procedure may find significant comparisons even if the overall F test is not significant because Tukey's test restricts itself to the pairwise comparison subset of contrast space. This is mentioned because it helps in deciding whether to use Tukey's procedure or Scheffe's procedure.

Basically, **Scheffe's** procedure treats pairwise comparisons as "just another contrast". Suppose you've just observed a significant overall F. Clearly you'd be interested in investigating the pattern(s) among the means that produced this result. In this context, pairwise comparisons are just one of any number of contrasts that may interest you; you are interested in general

"data-snooping". Scheffe's procedure controls the experimentwise error rate, but here the "experiment" is not just the m = P(P-1)/2 comparisons but all possible contrasts. The price you pay for such general protection is that Scheffe's procedure is more conservative than Tukey's; it will not detect some differences between means that Tukey's will. If the overall F test was not significant, Scheffe's comparisons will never be significant either. The critical value for Scheffe's is $C = SE(L_{ij}) [(P-1) F(P-1, DF)]^{½}$, where F(P-1, DF) is the appropriate F value.

Two comparison procedures that are very popular in the natural sciences and other areas are **Duncan's New** method and the **Student-Newman-Keuls**, or SNK, method. These procedures are not recommended and *Statistix* doesn't compute them. Duncan's New method controls the comparisonwise error rate at α and generally gives results similar to the LSD procedure. The SNK procedure doesn't control the experimentwise error rate under a partial null hypothesis and cannot be recommended (Einot and Gabriel 1975). There are a number of procedures more powerful than Tukey's that still control the experimentwise error rate (Ryan 1960, Einot and Gabriel 1975, Welsch 1977, Begun and Gabriel 1981). Like Duncan's New and SNK, these procedures are multiple-stage tests, which means the critical value doesn't remain constant for all comparisons but rather varies as the homogenous subsets are constructed. The disadvantages of such multiple-stage procedures are that they're more complex to explain and compute and, in particular, do not permit the construction of confidence intervals, which is often useful when you present your results.

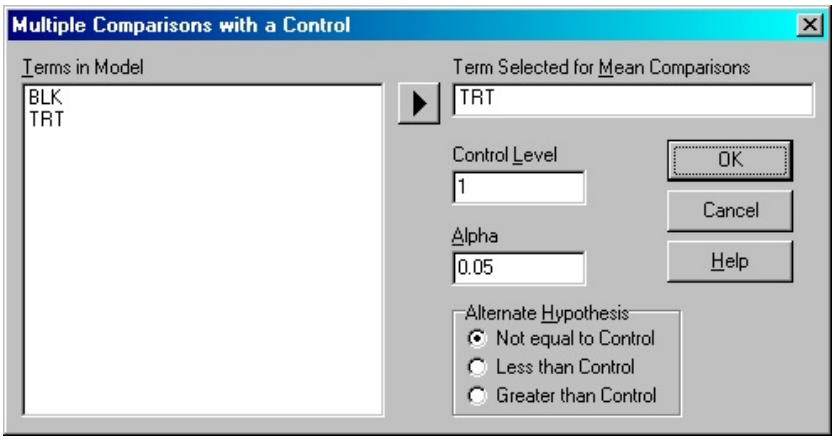
The basis for deciding which procedure to use is somewhat subjective and philosophical. Hsu (1996) recommends Tukey's method (also called the Tukey-Krammer method when used for unbalanced data) for preplanned all-pairwise comparisons.

Comparisons of Means - Computational Notes

Statistix computes quantiles for Student's t distribution using a procedure patterned after Hill (1970). Quantiles for the F distribution are found by finding the inverse of the corresponding beta distribution using Newton's method. The algorithm used to perform this is similar to Majumder and Bhattacharjee's (1973), although a different procedure, described in Probability Functions (Chapter 12), is used to compute the cumulative distribution function of the beta distribution. The quantiles for the Studentized range distribution are computed with a procedure patterned after Lund and Lund (1983).

Multiple Comparisons with a Control

This procedure makes use of Dunnett's test for comparing all treatment means with the mean of a control. This test is more powerful than using Tukey's all-pairwise test because there are fewer comparisons.



We'll use the randomized complete block example from page 230, where the treatment factor TRT was the type of fungicide applied to batches of 100 soybeans. TRT has five levels; the first is a no-fungicide control, and the remaining four are different types of fungicide. The dependent variable FAILURES is the number of beans out of 100 that failed to sprout.

First select the main effect or interaction for which you want comparisons and move it to the ***Term Selected for Mean Comparisons*** box. Next enter the value of the level that identifies the control treatment. If you're testing an interaction term, enter one value for each factor in the term, separated by commas. Enter a value for the rejection level in the ***Alpha*** edit control. Select an ***Alternate Hypothesis***. Select "not equal to the control" to perform the two-sided test.

The results for the analysis specified in the example dialog box above are shown on the next page.

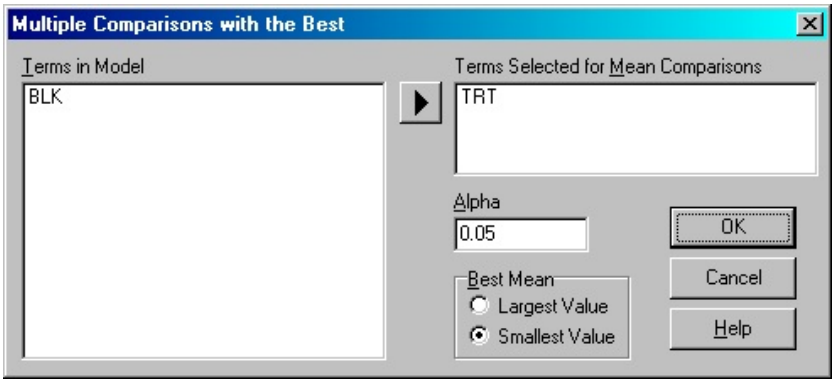
The table displays the difference between the control mean and the remaining means. Means significantly different from the control mean (differences significantly different from zero) are flagged with an asterisk. 95% simultaneous confidence interval of the mean differences are also displayed.

```
Two-sided Dunnett's Multiple Comparisons with a Control of FAILURES
Control: TRT=Control

Simultaneous 95% confidence intervals of treatment mean - control mean

                        Lower                    Upper
TRT          Mean       Bound  Difference        Bound
Control    10.800
Fung #1     6.200      -8.584      -4.600*      -0.616
Fung #2     8.200      -6.584      -2.600        1.384
Fung #3     6.600      -8.184      -4.200*      -0.216
Fung #4     5.800      -8.984      -5.000*      -1.016


Alpha              0.05     Standard Error for Comparison  1.4711
Critical D Value  2.708     Critical Value for Comparison  3.9839
Error term used: BLK*TRT, 16 DF
```

See Hsu (1996) for computational details.

Multiple Comparisons with the Best

This procedure makes use of Hsu's test for multiple comparisons with the best (Hsu, 1996). It's useful when you're most interested in identifying those treatments that may provide the best result. This test is more powerful than using Tukey's all-pairwise test because there are fewer comparisons.



First select the main effect or interaction for which you want comparisons and move it to the ***Terms Selected for Mean Comparisons*** box. Enter a value for the rejection level in the ***Alpha*** edit control. Select a value for what constitutes the ***Best Mean***: the one with the largest value, or the one

with the smallest value.

We'll use the randomized complete block example from page 230, where the treatment factor TRT was the type of fungicide applied to batches of 100 soybeans. The dependent variable FAILURES is the number of beans out of 100 that failed to sprout. The results for the analysis specified in the dialog box one the preceding page are shown below.

Hsu's Multiple Comparisons with the Best of FAILURES for TRT

Simultaneous 95% confidence intervals of mean - smallest of other means

TRT	Mean	Lower Bound	Difference	Upper Bound
Control	10.800	0.000	5.000*	8.447
Fung #1	6.200	-3.047	0.400	3.847
Fung #2	8.200	-1.047	2.400	5.847
Fung #3	6.600	-2.647	0.800	4.247
Fung #4	5.800	-3.847	-0.400	3.047

Alpha 0.05 Standard Error for Comparison 1.4711
Critical D Value 2.343 Critical Value for Comparison 3.4468
Error term used: BLK*TRT, 16 DF

The table displays the difference between each mean and the best (lowest in this example) of the remaining means. Means significantly different from the best mean are flagged with an asterisk. 95% simultaneous confidence interval of the differences from the best are also displayed.

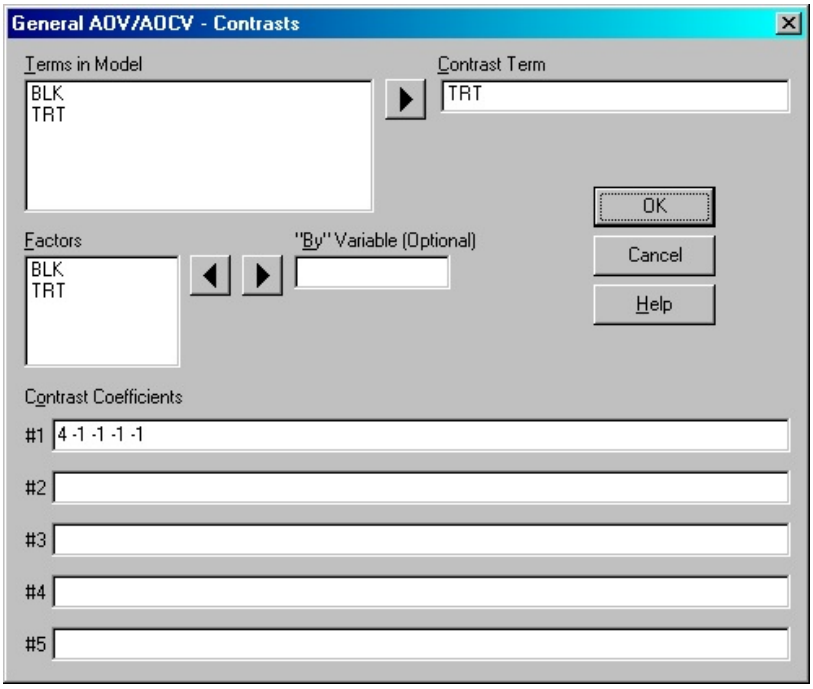
See Hsu (1996) for computational details.

Contrasts

This powerful option computes any linear contrast for any effect or interaction. Linear contrasts are linear combinations of the means for any effect or interaction, and they're valuable for examining the "fine structure" of the data after the overall F test indicates that the effect or interaction is significant.

Suppose, in the randomized block example on page 230, you're interested in whether the mean for the control (no fungicide) is different from the mean of the four treatments (fungicides applied). To make this comparison, you

enter “4 -1 -1 -1 -1" for the contrast coefficients.



To specify the general contrast, first select the main effect or interaction for which you want to construct contrasts and move it to the ***Contrast Term*** box—TRT in our example.

You can specify up to five contrasts at once. The coefficients entered must sum to 0, but their absolute values don’t matter. The coefficients can be entered as integer or real values. The ordering of the coefficients is determined by the values used to represent the levels of the factors. In our example, the no-fungicide control is represented in the variable TRT as 1, the four fungicides are represented as 2, 3, 4, and 5. The list of coefficients “4 -1 -1 -1 -1" can be abbreviated using a repeat factor: 4 4(-1). When entering coefficients for an interaction term, you enter them in the order with rightmost subscripts changing fastest (in the same order that means are listed using the **Mean and Standard Errors** procedure).

The results are for the example specified above are presented on the next page.

AOV Contrasts of FAILURES by TRT Fungicide treatments

Contrast Coefficients: 4 -1 -1 -1 -1

Contrast	16.400	SS (Contrast)	67.240
Scheffe's F	3.11	P (Scheffe's F)	0.0453
T-Statistic	3.53	P (T-Statistic)	0.0028
SE (Contrast)	4.6519		

Error term used: BLK*TRT, 16 DF

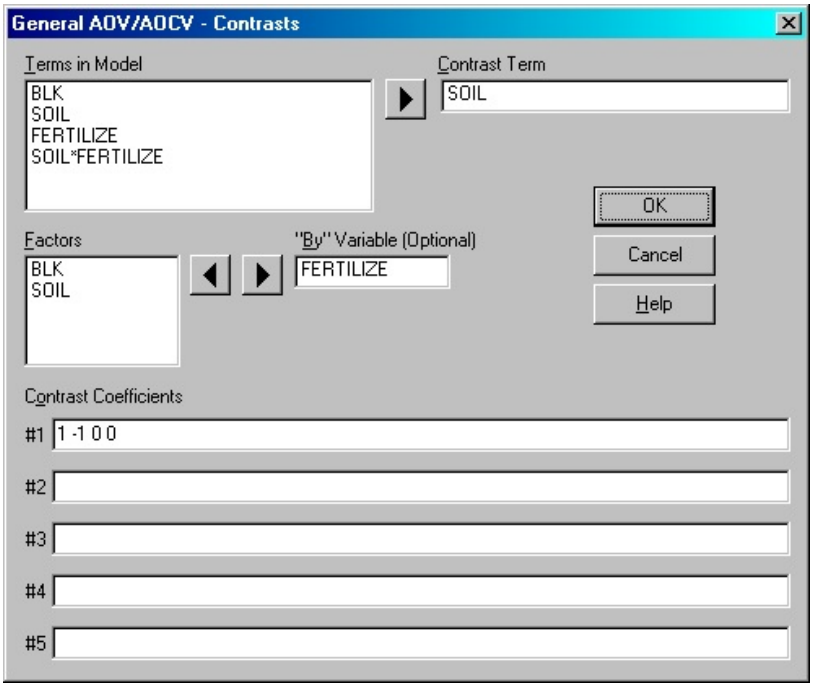
Scheffe's F method of significance testing for arbitrary simultaneous contrasts is used to test the hypothesis that the contrast is zero. The contrast in the above example is seen to be significant at the 5% level (p = 0.0453). Scheffe's procedure is appropriate for any number of a posteriori contrasts, which means it can be used to test hypotheses that arise after the data are collected and inspected. It protects you from making too many type I errors (rejecting a correct null hypothesis) during such "data-snooping". In the example output on the preceding page, the sum of squares due to contrast is computed in the usual way, as illustrated in Sections 12.7 and 12.8 of Snedecor and Cochran (1980). The computational methods used are discussed in Section 6.4 of Scheffe (1959).

The statistic Scheffe's F is computed as SSC/(DF∗MSE), which is equivalent to $L^2/(DF*SE(L)^2)$, where SSC is the sum of squares due to the contrast, MSE is the mean square for error, L is the value of the contrast, SE(L) is the standard error of the contrast, and DF is the degrees of freedom associated with the contrast. The same error term is used as would be used for the F test in the original AOV table. Scheffe's F will not be computed for contrasts of interaction terms in models that have multiple error terms; neither will it be computed for terms used as error terms in the model.

In addition to Scheffe's F method, Student's t test is performed. Student's t test is appropriate for a priori tests (contrasts that had been planned before the data were inspected). Student's t test doesn't control the experiment-wise error rate, as we discussed in the comparisons of means section (page 263). Student's t-statistic is computed as L/SE(L).

By Variable

It's possible to compute contrasts for a term in the model, but at each level of another factor in the model. Using the guinea pig forage example from page 256, we'll compute a contrast for comparing soil types 1 and 2, at each level for the factor FERTILIZE (1 = not fertilized, 2 = fertilized). This is specified in the dialog box on the next page.



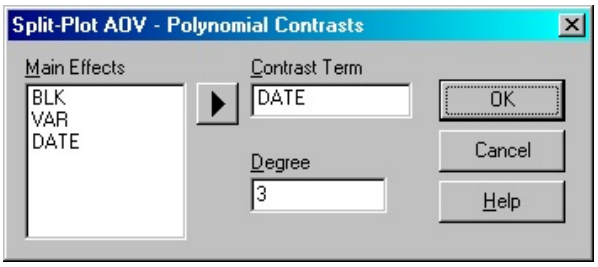
The results are presented below.

```
AOV Contrasts of WTGAIN by SOIL for FERTILIZE

Contrast Coefficients: 1 -1 0 0

FOR FERTILIZE = 1
Contrast           121.17         SS (Contrast)        22024
Scheffe's F         17.15         P (Scheffe's F)     0.0001
T-Statistic          7.17         P (T-Statistic)     0.0000
SE (Contrast)      16.894

FOR FERTILIZE = 2
Contrast           144.93         SS (Contrast)        31508
Scheffe's F         24.53         P (Scheffe's F)     0.0000
T-Statistic          8.58         P (T-Statistic)     0.0000
SE (Contrast)      16.896

Error term used: BLK*SOIL*FERTILIZE, 12 DF
```

The contrast for the FERTILIZE=1 test is computed using the means for the SOIL* FERTILIZE interaction, but only using the means for the not fertilized cells. The contrast for the FERTILIZE=2 test is computed using only the means for the fertilized cells.

Polynomial Contrasts

This option computes the polynomial decomposition of the sums of squares for any main effect. This is very useful for determining the existence and nature of trends in the treatment level means.



In the split-plot example on page 242, we're interested in examining the trends of yield TONS as a function of cutting date DATE. In the dialog box above, we've entered DATE for the ***Contrast Term*** and entered a value of 3 for ***Degree***.

By specifying polynomials up to degree 3 be computed, we will get sums of squares for linear (degree 1), quadratic (degree 2), and cubic (degree 3) trends due to DATE. The results are shown below.

```
Polynomial Contrasts of TONS by DATE  Date of harvest

Degree = 1, Linear Trend

Contrast           0.3073         SS (Contrast)       1.7003
Scheffe's F         20.27         P (Scheffe's F)     0.0000
T-Statistic          7.80         P (T-Statistic)     0.0000
SE (Contrast)      0.0394

Degree = 2, Quadradic Trend

Contrast          -0.1200         SS (Contrast)       0.2594
Scheffe's F          3.09         P (Scheffe's F)     0.0363
T-Statistic         -3.05         P (T-Statistic)     0.0039
SE (Contrast)      0.0394

Degree = 3, Cubic Trend

Contrast           0.0123         SS (Contrast)     2.72E-03
Scheffe's F          0.03         P (Scheffe's F)     0.9920
T-Statistic          0.31         P (T-Statistic)     0.7565
SE (Contrast)      0.0394

Error term used: BLK*VAR*DATE, 45 DF
```

Strong support exists for a linear trend with DATE ($p = 0.0000$) in the

example. The positive value for the contrast (0.3073) indicates that TONS increase with increasing DATE. There also appears to be some evidence of a quadratic trend in addition to the linear trend. Perhaps this indicates that a date before November 10 should have been used as the end of the effective growing season. For example, if you use October 15 instead (DATE = 80 instead of 106), the linear trend is stronger and the quadratic trend disappears.

Remember that the actual spacings of the treatment levels are used to compute polynomial contrasts. Because calculating unequally spaced polynomials can be quite tedious, researchers commonly ignore the unequal spacing of levels and treat them as equally spaced, even though this can result in substantial errors. As the example shows, the choice of level spacings can have considerable influence on the results. Because of the ease with which this option can handle unequal spacings, there is little excuse for not using them.

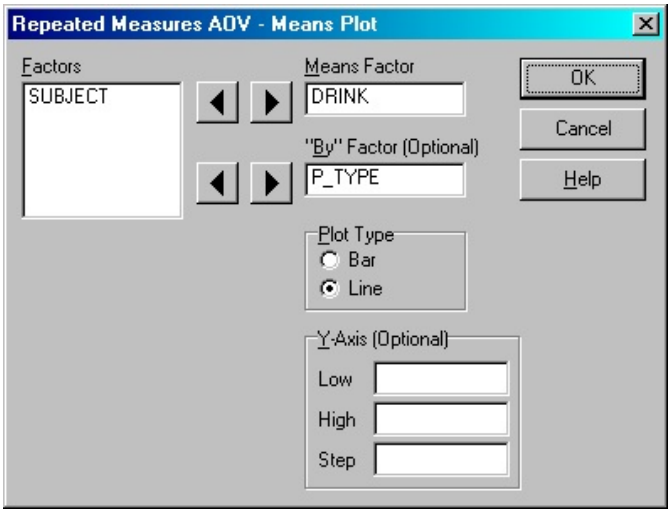
Plots

The Plots submenu offers an option to plot means for main effects and two-factor interactions, and two plots for examining the residuals.

The **Normal Probability Plot** plots the residuals against the rankits. Plots for normal data form a straight line. The Shapiro-Wilk statistic for normality is also reported on the plot. See Chapter 9 for details.

The **Resids By Fitted Values** plot is useful for examining whether the variances are equal among the groups. If the order of the groups is meaningful, then systematic departures from equality can be seen in the plot.

We'll illustrate the **Means Plot** using the two-factor repeated measures example discussed on page 250. We're interested in looking at how alcohol affects the aggressiveness of people with and without antisocial personality disorder (ASP).



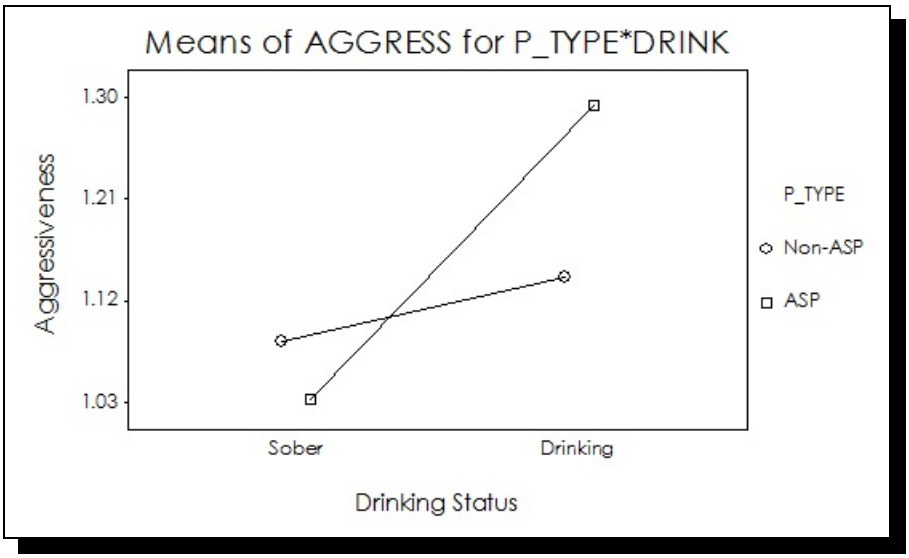
A mean plot plots the dependent variable along the Y axis. The factor you specify in the ***Means Factor*** box is plotted along the X axis. If you specify a ***By Factor***, one line-plot of means are plotted for each level of the factor specified. The Means Factor is limited to 30 levels. The By Factor is limited to 6 levels.

Next select the ***Plot Type***, either bar chart or line chart. The bar chart uses vertical bars to represent the means. The line chart uses circles to mark the means, and the circles are connected sequentially with lines.

You can enter ***Low***, ***High***, and ***Step*** values to control the Y axis scale. You can use this feature to create a meaningful interval width and interval boundaries.

The means plot for the alcohol example specified above is shown on the next page.

Recall that the F test for the P_TYPE*DRINK interaction was significant. The means plot shows that while aggressiveness increases for both personality types while drinking alcohol, the increase is more dramatic for the subjects diagnosed with ASP.



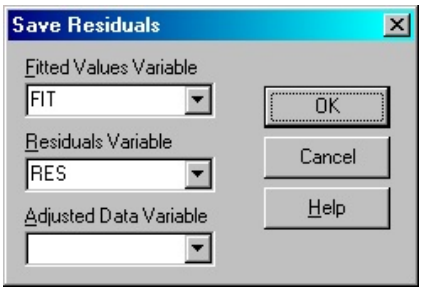
If you specified more than one dependent variable for your analysis of variance, you view a plot for one dependent variable at a time. Arrows appear on the toolbar, as shown below.

Press the right-arrow button on the toolbar to display the plot for the next dependent variable. Press the left-arrow button to display the plot for the previous dependent variable.

The **Titles** procedure on the Plots menu is used to changes the titles of the plot displayed. The **Graph Preferences** procedure is used to change details of the plot, such as font and symbol type. See Chapter 1 for details.

Save Residuals

Select Save Residuals to compute the residuals, fitted values, or adjusted data for a particular model and store them for later examination. An example dialog box is shown below.



You can type in the names of new variables in the spaces provided, or you can click on the down arrow to select the name of an existing variable from a drop down list.

The residuals are used to evaluate how well a model fits the data. A residual is defined as the difference between the actual observed response and that predicted by the fitted model. Residuals can help detect bad values (outliers) and can also help suggest more appropriate models or transformations to apply. Consult Daniel (1976) for more detail.

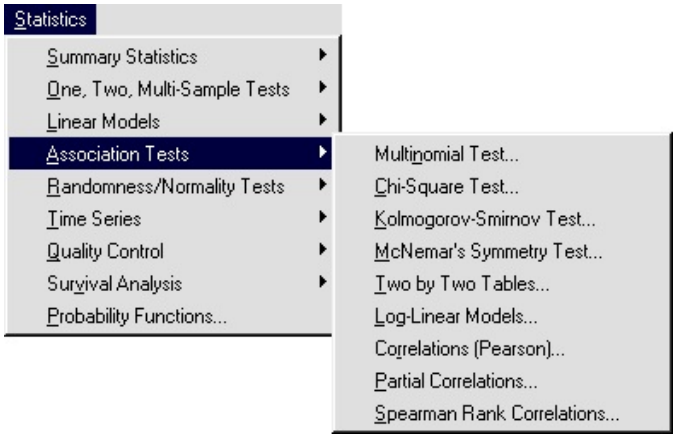
The ***Adjusted Data Variable*** option is only available when you included covariates in the model using the General AOV/AOCV procedure. The adjusted data are the dependent variable data adjusted for the values for the covariates on a case by case basis.

If you specified more than one dependent variable for your analysis of variance, you can only save the residuals for the first dependent variable listed.

C H A P T E R

8

Association Tests



Statistix offers many association tests that can be used to examine the similarity or association among two or more variables.

The **Multinomial Test** is a goodness-of-test that tests how well frequencies of mutually exclusive categories fit a hypothesized distribution.

The **Chi-Square Test** computes the traditional chi-square goodness-of-fit test for two-way tables. Two hypotheses can be examined with this test: the hypothesis of independence, and the hypothesis of homogeneity.

The **Kolmogorov-Smirnov Test** is useful for comparing the similarity of the distributions of samples from two populations. If there is an intrinsic ordering to the categories, the Kolmogorov-Smirnov test is usually better than the chi-square test because it can exploit the information in the ordering while the chi-square analysis cannot.

The **McNemar's Symmetry Test** is a goodness-of-fit test that's often useful for measuring change. It's used to analyze square contingency tables; often the rows represent classifications before some event, while the columns represent the same classes after some event. Individuals may be in one class before the event but in another class after the event. However, if the table is symmetric about the diagonal from the upper left to the lower right, there will be no net shift in the row and column proportions before and after. McNemar's test examines whether the table is in fact symmetric.

The **Two by Two Tables** procedure computes a variety of tests of association for two by two contingency tables. A typical example of a two by two table is where a number of individuals are cross-classified by two dichotomous variables, such as treated-not treated and survived-died. The tests include Fisher's exact test, Pearson chi-square, log odds ratio, and others, along with standard errors.

The **Log-Linear Models** procedure is a powerful tool for analyzing discrete multidimensional categorical data. Log-linear models are the discrete data analogs to analysis of variance. If a set of discrete data has more than two classifying variables, you may be tempted to analyze such data as a series of two-way tables with traditional chi-square tests. However, the danger of such an approach is that collapsing the data over some categorical variables results in these variables becoming confounded with the remaining two categorical variables. Log-linear models allows all dimensions of multidimensional contingency tables to be treated simultaneously and so avoids such potential confounding.

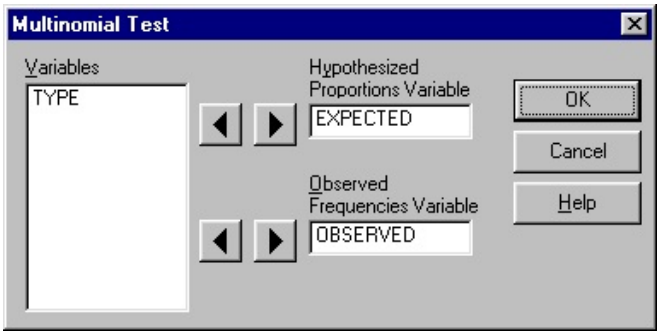
The **Correlations** procedure measures the degree of linear association between two variables. The **Partial Correlations** procedure allows you to examine the degree of linear association between two variables after the effect of other variables have been "adjusted out". These procedures also appear on the Linear Models menu and are discussed in Chapter 6.

The **Spearman Rank Correlations** procedure produces nonparametric correlation coefficients that are suitable for examining the degree of association when the samples violate the assumption of bivariate normality.

Multinomial Test

The Multinomial Test is a goodness-of-test that tests how well frequencies of mutually exclusive categories fit a hypothesized distribution. For example, it can be used to test whether or not a sample of 100 rolls of a die support the hypothesis that each number is equally likely to be rolled. The large-sample chi-square approximation is used for this test.

Specification



The test requires two variables. The ***Hypothesized Proportions Variable*** contains the list of hypothesized proportions. The values can be entered as proportions that sum to 1, or on any arbitrary scale such that the relative values represent the hypothesized proportions. The ***Observed Frequencies Variable*** contains the corresponding list of observed frequencies.

Example

In crosses between two types of maize, four distinct types of plants were found in the second generation (Snedecor and Cochran, 1980): green, golden, green-striped, and golden-green-striped. According to a simple type of Mendelian inheritance, the probabilities of obtaining these four types of plants are 9/16, 3/16, 3/16, and 1/16. The frequencies tabulated for a sample of 1301 plants, and the expected ratios, are listed below.

TYPE	OBSERVED	EXPECTED
green	773	9
golden	231	3
green-striped	238	3
golden-green-striped	59	1

The analysis is specified using the dialog box above. The results appear on the next page.

Multinomial Test

Hypothesized Proportions Variable: EXPECTED
Observed Frequencies Variable: OBSERVED

Category	Hypothesized Proportion	Observed Frequency	Expected Frequency	Chi-Square Contribution
1	0.56250	773	731.81	2.32
2	0.18750	231	243.94	0.69
3	0.18750	238	243.94	0.14
4	0.06250	59	81.31	6.12

Overall Chi-Square 9.27
P-value 0.0259
Degrees of Freedom 3

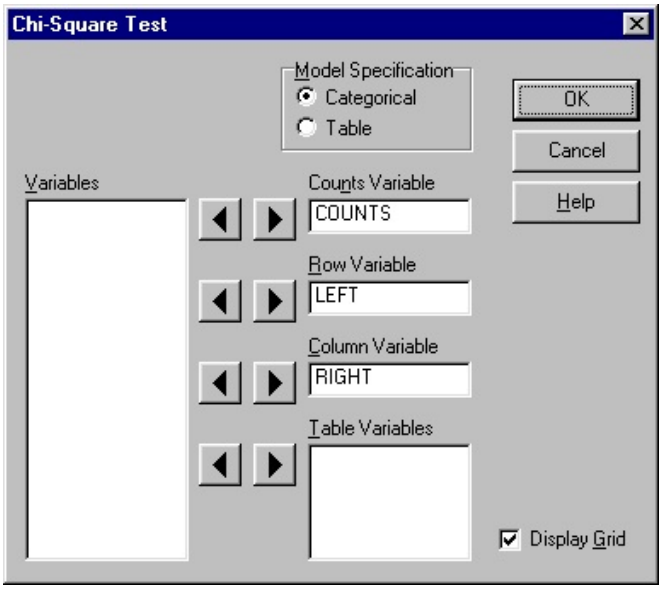
The p-value for the overall chi-square test is 0.0259, so we reject the null hypothesis that the 9:3:3:1 ratio is correct. We see from the chi-square contribution column that category 4 gives the largest contribution to chi-square. The original researcher noted that the golden-green-striped plants were not vigorous due to their chlorophyll abnormality.

Chi-Square Test

The **Chi-Square Test** procedure is used to analyze two-dimensional tables of discrete data. Two hypotheses can be examined; the hypothesis of independence examines whether the row-classifying variable acts independently of the column-classifying variable, and the hypothesis of homogeneity tests whether the relative frequency distributions for each of the rows or columns are the same. The appropriate hypothesis choice depends on how the sampling was performed. The calculations involved are identical for the two tests.

Specification

The analysis can be specified in two ways, depending on how you choose to enter the data. If you enter the data in columns using two categorical variables to identify the rows and columns, select the ***Categorical*** method and identify the two classifying variables and, optionally, a dependent variable containing counts. You can also enter the data in as a two-dimensional table, with variables identifying the columns and cases identifying the rows. This is called the ***Table*** method.



Select either the Categorical or the Table method that fits the way your data have been entered. Using the Categorical method, select the classifying variable that identifies the rows of the contingency table and move it to the ***Row Variable*** box. Select the variable that identifies the columns and move it to the ***Column Variable*** box. If each case represents one observation, don't select a ***Counts Variable***. If the data are summarized, select the variable that contains the counts for each case.

Using the Table method, select the variables that will represent the columns of the contingency table and move them to the ***Table Variables*** box.

Data Restrictions

No more than 500 row categories and 500 column categories are allowed. Missing values are not permitted. The count data must be nonnegative integer values.

Example

We use the data from Table 8.2-1 in Bishop, Fienberg, and Holland (1975) for our example. The data are the results of vision tests for 7,477 women. The variables LEFT and RIGHT are the scores for the left and right eyes, respectively. Each eye was assigned a score 1, 2, 3, or 4. The counts that fall in each of the cells of the contingency table are in the variable COUNTS. The data are presented on the next page, an are stored in the file Sample Data\vision.sx

CASE	COUNTS	LEFT	RIGHT
1	1520	1	1
2	234	1	2
3	117	1	3
4	36	1	4
5	266	2	1
6	1512	2	2
7	362	2	3
8	82	2	4
9	124	3	1
10	432	3	2
11	1772	3	3
12	179	3	4
13	66	4	1
14	78	4	2
15	205	4	3
16	492	4	4

The analysis is specified on the preceding page. The results are displayed below.

Chi-Square Test for Heterogeneity or Independence
for COUNTS = LEFT RIGHT

		RIGHT				
LEFT		1	2	3	4	
1	Observed	1520	234	117	36	1907
	Expected	503.98	575.39	626.40	201.23	
	Cell Chi-Sq	2048.32	202.55	414.25	135.67	
2	Observed	266	1512	362	82	2222
	Expected	587.22	670.43	729.87	234.47	
	Cell Chi-Sq	175.72	1056.38	185.41	99.15	
3	Observed	124	432	1772	179	2507
	Expected	662.54	756.43	823.48	264.55	
	Cell Chi-Sq	437.75	139.14	1092.53	27.66	
4	Observed	66	78	205	492	841
	Expected	222.26	253.75	276.25	88.75	
	Cell Chi-Sq	109.86	121.73	18.38	1832.37	
		1976	2256	2456	789	7477

Overall Chi-Square 8096.88
P-Value 0.0000
Degrees of Freedom 9

Cases Included 16 Missing Cases 0

In each cell, the original observed value, the value expected under the hypothesis of independence or homogeneity, and the cell contribution to the overall chi-square are displayed. In this example, the hypothesis of independence is appropriate. If either the row or the column totals had been predetermined before the sample was taken, the hypothesis of homogeneity would have been appropriate. The calculations are the same for either hypothesis.

3In our example, the large differences between the expected values and the

observed values indicate that the model of independence is not an acceptable model for this data set (p-value = 0.0000). This seems quite reasonable; one might reasonably expect the vision score for one eye to have a positive association with that for the other eye.

It's more convenient in some situations to store the column categories as separate variables. For example, the data listed on the preceding page can rearranged using the ***Table*** format.

CASE	RIGHT1	RIGHT2	RIGHT3	RIGHT4
1	1520	234	117	36
2	266	1512	362	82
3	124	432	1772	179
4	66	78	205	492

The variables RIGHT1, RIGHT2, RIGHT3, and RIGHT4 contain the data for columns 1 through 4 of the table (right eye scores 1 through 4). Each variable has four cases, and the order of the cases represent rows 1 through 4 of the table (left eye scores 1 through 4). The analysis for these data is specified by moving the four variables to the ***Table Variables*** box.

The most common problem in applying the chi-square test is that it becomes unreliable when numerous expected cell values are near zero. Snedecor and Cochran (1980, p. 77) give the following general rules:

1) No expected values should be less than one.
2) Two expected values may be close to one if most of the other expected values exceed five.
3) Classes with expectations less than one should be combined to meet 1) and 2).

If your contingency table is a two by two table (two categories in both the rows and columns), use the **Two by Two Tables** procedure (page 291).

Computational Notes

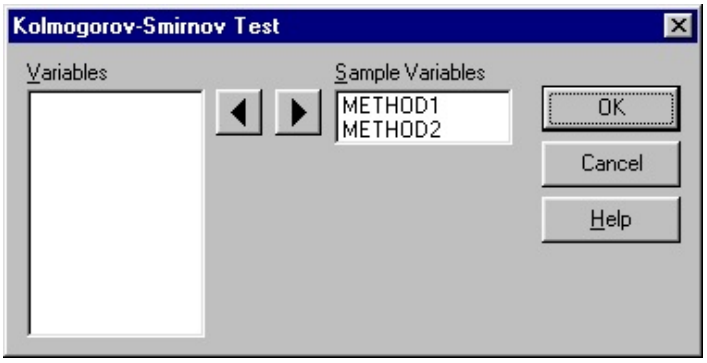
You should consult Snedecor and Cochran (sec. 10.11) for more detail on this test. We don't use the correction for continuity; we believe that the arguments given by Fienberg (1980) for not using the correction are compelling.

Kolmogorov-Smirnov Test

The **Kolmogorov-Smirnov Test** procedure examines whether two samples have the same distribution. You must order the categories within the samples. The test, which is also known as the Smirnov test, is sensitive to any differences between the distributions, including differences in means and variances. It's generally preferable to a chi-square test because it exploits the information in the ordering of the categories.

Specification

The sample counts for one distribution are in one variable, and the counts for the other distribution are in a second variable. It's assumed that the ordering of the cases in the two variables reflects the ordering of the categories.



Select the names of the two variables containing the samples. The order of the variables isn't important, except that the signs of the resulting one-tailed statistics are reversed.

Data Restrictions

The data must be nonnegative whole numbers and can't exceed 99,999. There must be at least five cases.

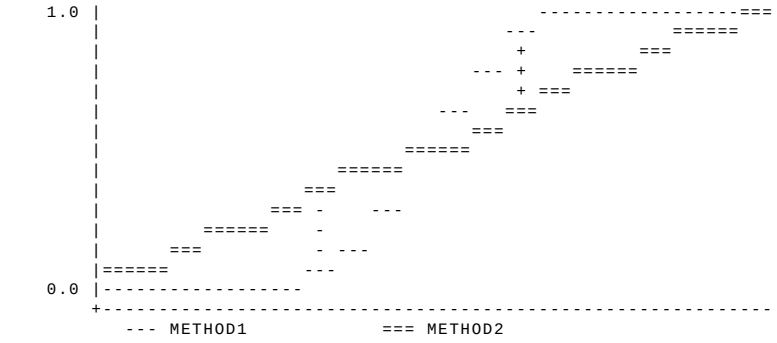
Example

The data for our example are fabricated. Suppose you want to examine whether students' test scores are the same under two teaching methods. The variables METHOD1 and METHOD2 represent the number of students receiving a particular grade. There are 20 questions on the test, so there are 20 categories corresponding to the total number of possible scores. The data are listed in the table on the next page.

CASE	METHOD1	METHOD2
1	1	7
2	0	5
3	0	7
4	1	5
5	1	6
6	0	8
7	6	6
8	12	7
9	18	7
10	30	4
11	24	5
12	20	7
13	21	8
14	8	9
15	2	5
16	1	5
17	0	6
18	0	7
19	0	5
20	1	7

The analysis is specified on the preceding page. The results are as follows:

Two Sample Kolmogorov-Smirnov (Smirnov) Test




```
Sample size for METHOD1                      146
Sample size for METHOD2                      126

Hypothesis : METHOD1 <> METHOD2
Two-tailed Kolmogorov-Smirnov Statistic     0.29
P-Value (Smirnov's Chi-Square Approx.)    0.0000

Hypothesis : METHOD1 < METHOD2
One-tailed Kolmogorov-Smirnov Statistic    -0.29
P-Value (Smirnov's Chi-Square Approx.)    0.0000

Hypothesis : METHOD1 > METHOD2
One-tailed Kolmogorov-Smirnov Statistic     0.27
P-Value (Smirnov's Chi-Square Approx.)    0.0001
```

The graph at the top of the replort displays the sample cumulative distribution functions. The Kolmogorov-Smirnov two-sample test is based on the maximum difference between the cumulative distribution functions. The location of the greatest negative difference between the two curves is

shown by the column of "-"s. The greatest positive difference is located at the column of "+"s. The two-tailed test, which tests the hypothesis that the two curves are different, is based on the difference with the greatest absolute value, 0.29 in the example. The very small p-value indicates that the distributions are indeed different. One-tailed tests are also computed to test (1) whether the largest positive difference is greater than expected by chance if the two distributions are identical, and (2) whether the smallest negative difference is smaller than expected. Note that in this example both one-tailed tests are significant. The p-values are based on Smirnov's (1939) approximation using the chi-square distribution. P-values are computed only if the sum of the observations in the variables both exceed 15.

See Lehmann (1975) and Hollander and Wolfe (1973) for more detail.

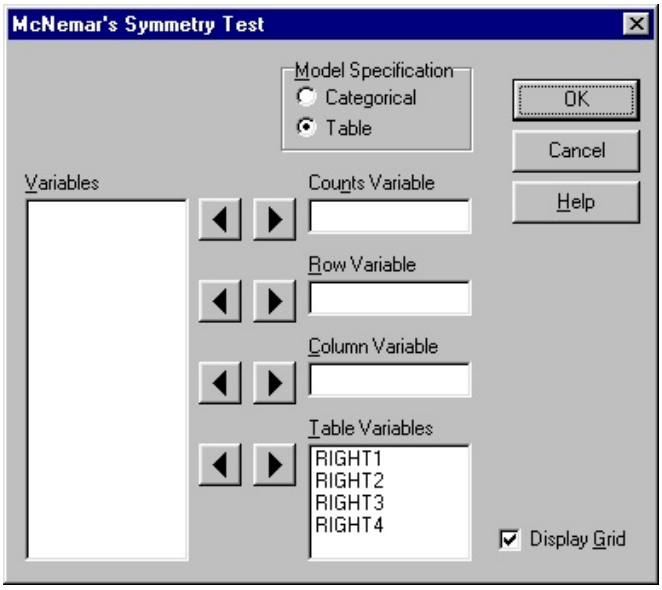
McNemar's Symmetry Test

The **McNemar's Symmetry Test** procedure tests whether a square contingency table is symmetric about the diagonal running from the upper left to the lower right. It is actually a generalized version of McNemar's test developed by Bowker (1948).

Specification

The analysis can be specified in two ways, depending on how you choose to enter the data. If you enter the data in columns using two categorical variables to identify the rows and columns, select the ***Categorical*** method and identify the two classifying variables and, optionally, a dependent variable containing counts. You can also enter the data in as a two-dimensional table, with variables identifying the columns and cases identifying the rows. This is called the ***Table*** method.

Select either the Categorical or the Table method that fits the way your data have been entered. Using the Categorical method, select the classifying variable that identifies the rows of the contingency table and move it to the ***Row Variable*** box. Select the variable that identifies the columns and move it to the ***Column Variable*** box. If each case represents one observation, don't select a ***Counts Variable***. If the data are summarized, select the variable that contains the counts for each case.



Using the Table method, select the variables that will represent the columns of the contingency table and move them to the ***Table Variables*** box.

Data Restrictions

The contingency table must be square, i.e., the number of rows and columns must be equal. No more than 500 row and column categories are allowed. Missing values aren't permitted. The count data should be nonnegative integer values.

Example

We use the data from Table 8.2-1 in Bishop, Fienberg, and Holland (1975) for our example. The data are the results of vision tests for 7,477 women. The variables LEFT and RIGHT are the scores for the left and right eyes, respectively. Each eye was assigned a score 1, 2, 3, or 4. The data were entered in the Table format (see the **Chi-Square Test** on page 282 for an example of the Categorical method).

CASE	RIGHT1	RIGHT2	RIGHT3	RIGHT4
1	1520	234	117	36
2	266	1512	362	82
3	124	432	1772	179
4	66	78	205	492

The analysis of the vision test data is specified on the preceding page. The results are presented below.

McNemar's Symmetry Test

Case		Variable RIGHT1	RIGHT2	RIGHT3	RIGHT4	
1	Observed	1520	234	117	36	1907
	Expected	1520.00	250.00	120.50	51.00	
	Cell Chi-Sq	0.00	1.02	0.10	4.41	
2	Observed	266	1512	362	82	2222
	Expected	250.00	1512.00	397.00	80.00	
	Cell Chi-Sq	1.02	0.00	3.09	0.05	
3	Observed	124	432	1772	179	2507
	Expected	120.50	397.00	1772.00	192.00	
	Cell Chi-Sq	0.10	3.09	0.00	0.88	
4	Observed	66	78	205	492	841
	Expected	51.00	80.00	192.00	492.00	
	Cell Chi-Sq	4.41	0.05	0.88	0.00	
		1976	2256	2456	789	7477

Overall Chi-Square	19.11
P-Value	0.0040
Degrees of Freedom	6

In each cell, the original observation, the value expected under the hypothesis of symmetry, and the cell contribution to the overall chi-square are displayed. Clearly, this model fits the data better than the model of independence examined in **Chi-Square Test**, but there's still a significant lack of fit (p-value = 0.0040). By examining the cell chi-squares, you'll see that the left and right pairs on the diagonal running from the lower left to the upper right are primarily responsible for the lack of symmetry.

The most common problem in the application of McNemar's test is that it becomes unreliable when numerous expected cell values are near zero. The guidelines usually given for the chi-square test for independence and homogeneity are also applicable to McNemar's test. Snedecor and Cochran (1980, p. 77) give the following general rules:

1) No expected values should be less than one.
2) Two expected values may be close to one if most of the other expected values exceed five.
3) Classes with expectations less than one should be combined to meet 1) and 2).

Computational Notes

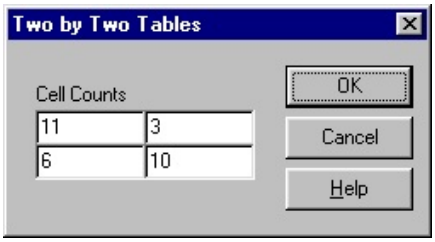
Consult Bishop, Fienberg, and Holland (1975) for more detail on this test and example.

Two By Two Tables

The **Two By Two Tables** procedure computes several measures and tests of association for two by two contingency tables.

Specification

When the two by two tables procedure is selected, an empty two by two table is displayed, with the cursor positioned in the upper left cell:



Simply enter the number you want in each of the four cells, pressing Tab to move forward to the next cell. Press the ***OK*** button to compute the analysis.

Data Restrictions

The cell values must be nonnegative integers. All row and column totals must be greater than zero. The individual cell values can't exceed 99,999.

Example

Suppose 30 people are selected at random and asked two questions. The first question is whether they favor increasing the budget for space research, and the second is whether they favor increasing the defense budget. The responses are as follows:

		Increase the space research budget? Yes	No
Increase the defense budget?	Yes	11	3
	No	6	10

The goal of the analysis is to examine whether the responses to the two questions are related. That is, if a person favors increasing space research, would he or she also favor increased defense spending? The results of the analysis are shown on the next page.

Two by Two Tables

11	3	14
6	10	16
17	13	30

Fisher Exact Tests: Lower Tail 0.0279 Upper Tail 0.0051 Two Tailed 0.0329

Pearson's Chi-Square	5.13	Yule's Q	0.72
P (Pearson's)	0.0235	SE (Q)	0.2009
Yates' Corrected Chi-Sq	3.59	SE (H0: Q = 0)	0.3693
P (Yates)	0.0580	Yule's Y	0.42
Log Odds Ratio	1.8101	SE (Y)	0.1704
SE (LOR)	0.8312	SE (H0: Y = 0)	0.1846
SE (H0: LOR = 0)	0.7385	C Max	0.63
Odds Ratio	6.1111	Phi	0.41
Lower 95% CI for OR	1.1983	Phi Max	0.82
Upper 95% CI for OR	31.164	Contingency Coeff	0.38

Descriptions of most of these measures can be found in Bishop et al. (1975) and the BMDP-83 manual. Notice that where standard errors are reported for a measure, two types of standard errors are given—unrestricted and restricted. The restricted standard error is appropriate for constructing hypothesis tests that the measure equals zero. The unrestricted measure is appropriate for constructing confidence intervals around the measure if the hypothesis that the measure equals zero is rejected. Brown and Benedetti (1977) should be consulted for more detail. Large sample theory can be used to test whether a measure differs from zero. If the samples are "large enough", the measures should be nearly normally distributed. Under the null hypothesis, a measure divided by its restricted standard error should have a Z, or standard normal distribution. For example, to test whether the log odds ratio is different from zero, we compute Z = 1.8101 / 0.7385 = 2.451. When we calculate the p-value for Z = 2.451 with the Z2TAIL procedure (see Chapter 13 **Probability Functions**), we find that p = 0.014, suggesting that the log likelihood ratio is different from zero. Approximate 95% confidence intervals around the log odds ratio would be constructed as $1.8101 \pm 1.96 * 0.8312$ using the unrestricted standard error.

The results of this analysis suggest that there is, in fact, a relationship between the responses to the two questions.

Fisher's exact test is displayed for total sample sizes of 500 or less. The other statistics are only displayed for total sample sizes of 26 or greater.

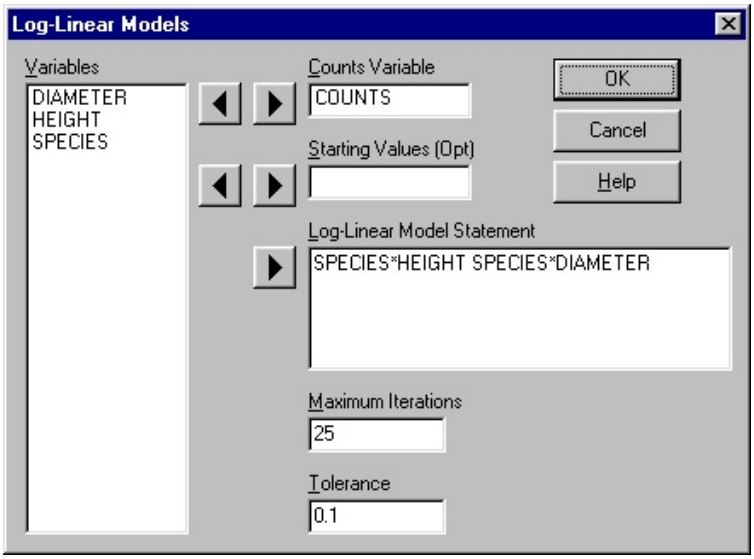
Computational Notes

The procedure for determining the upper- and lower-tail probabilities for Fisher's exact test is described in Bradley (1968). Woolf's method is used to compute the confidence intervals for the odds ratio and is described in Lee (1992, p. 286).

Log-Linear Models

The **Log-Linear Models** procedure fits a hierarchical log-linear model to the data in a multidimensional contingency table and computes several goodness-of-fit statistics. You can also save the expected values and residuals.

Specification



The analysis requires a variable which contains the discrete count data. Move the variable name containing this data to the ***Dependent Variable*** box. You can specify starting values for the estimated cell counts by moving the name of the variable that contains the starting values to the ***Starting Values*** box. This isn't necessary normally. Starting values are discussed in detail on page 297.

Log-linear model configurations are specified in a manner quite similar to the traditional approach of Bishop, Fienberg, and Holland (1975) and Fienberg (1980). Some examples are given below. The variables X1, X2, …, and XN represent the categorical variables by which the data in the dependent variable are cross-classified. The Xi's should be integers; decimal values are truncated. (Hint: the CAT function in **Transformations** is often very useful for generating the classifying variables.)

Example 1: Two-way table. Model for complete independence.

```
X1 X2
```

Example 2: Three-way table. Conditional independence of X1 and X2 given X3.

```
X1*X3 X2*X3
```

Example 3: Four-way table. Model includes a three-way interaction.

```
X1 X2*X3*X4
```

Variables within the same interaction term are separated by "*". Interaction terms are set off from other interaction terms by spaces. The order in which the terms are specified in the model doesn't matter, and neither does the order of variables within the interaction terms.

Log-Linear Models is based on a procedure called iterative proportional fitting (IPF), which is described in the references mentioned above. Three aspects of the IPF procedure can be controlled by the user: (1) starting values for the estimated expected values, (2) maximum number of iterations, and (3) estimated expected value convergence criterion, or tolerance. You should specify these only if you're certain the defaults are unacceptable. The defaults are appropriate for the majority of applications. The details of changing the defaults follow.

The starting values for the estimated expected values default to 1.0 unless they're specified otherwise. This is appropriate for most log-linear analyses (but see Specifying Starting Values on page 297). To override the default, a variable containing the starting values must be supplied. These initial values don't need to be integers.

The IPF algorithm continues until one of two termination criteria is satisfied. These criteria are (1) the number of iterations and (2) the maximum absolute difference in the cell estimates from one iteration to the next, which is called the tolerance. The default value for the tolerance is

0.01; if none of the absolute differences in the estimated expected cell counts from one iteration to the next exceeds the tolerance value, satisfactory convergence is assumed to have occurred. If the maximum allowable number of iterations is reached, it's assumed that the procedure didn't converge satisfactorily and no results will be given. The default for the maximum allowable number of iterations is 25.

Decreasing the size of the tolerance often increases the required number of iterations. The default tolerance value will usually be satisfactory.

Data Restrictions

The maximum number of classifying variables that are allowed in one model is seven. Within each classifying variable, the maximum number of classes permitted is 500. The combinations of cross-classifications must be unique and exhaustive. For example, suppose the model is Y = X1 X2 and that X1 has four classes (1, 2, 3, 4) associated with it and X2 has three classes (1, 2, 3) associated with it. This requires a total of 12 cases; the only pairs of classes for X1, X2 that may appear are 1,1; 1,2; 1,3; 2,1; 2,2; 2,3; 3,1; 3,2; 3,3; 4,1; 4,2; and 4,3. If all possible cross-classifications are not present or if some are duplicated, an error message is given. No missing values are allowed in the cell data.

Example

The data in the multidimensional contingency table may have been generated by either a Poisson, multinomial, or product multinomial sampling scheme. Consult Bishop et al. (1975) for more detail. It's assumed that the model to be fitted is hierarchical; all lower-order interactions that are subsets within the specified configurations are always included in the model.

The example is from Table 3-2 of Fienberg (1980). The variable COUNTS contains the counts for two species of tree-dwelling lizards. The object of the analysis is to examine how the height of a perch, the diameter of a perch, and the species of lizard influence perch selection. The categorical variable SPECIES indicates which species a count is for. (SPECIES = 1 and SPECIES = 2 for the two species, respectively.) There are two perch-diameter classes recorded in the variable DIAMETER. Likewise, there are two perch-height classes in HEIGHT. A variety of models could be fitted to this data. Refer to Bishop et al. (1975) and Fienberg (1980) for strategies for finding the "best" models.

The results for a model of conditional independence are shown. That is,

suppose it's hypothesized that, given the species of a lizard, the diameter of the perch it selects is independent of the height of the perch. The model is specified in the dialog box on page 293. The results are displayed below.

Log-linear Model Analysis on COUNTS

Configuration 1 SPECIES*HEIGHT
Configuration 2 SPECIES*DIAMETER

Goodness-of-fit Summary Statistics

Statistic	Chi-Sq	DF	P
Pearson	6.11	2	0.0471
Likelihood	4.88	2	0.0871
Freeman-Tukey	3.99	2	0.1360

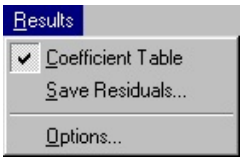
Number of Near Zero Expected Cells 0
Number of Iterations Performed 2
Termination Criterion Difference 0.10

The Pearson, likelihood ratio, and Freeman-Tukey goodness-of-fit statistics are discussed in Bishop et al. (1975). In this example, it's difficult to decide whether the model of conditional independence fits the data; the Pearson chi-square would lead to the rejection of this model at the 5% significance level, while the other two goodness-of-fit statistics would not. In an actual analysis, it's often desirable to look at the results for all feasible models.

The degrees of freedom are computed as the total number of cells in the multidimensional table minus the number of independent parameters that were estimated. This is appropriate for a "complete" table that has no zero marginals. (See Bishop et al. for the definition of complete tables.) Incomplete tables or zero marginals will require special treatment. The number of near zero cell estimates are displayed to indicate when the degrees of freedom need to be adjusted.

Log-Linear Results Menu

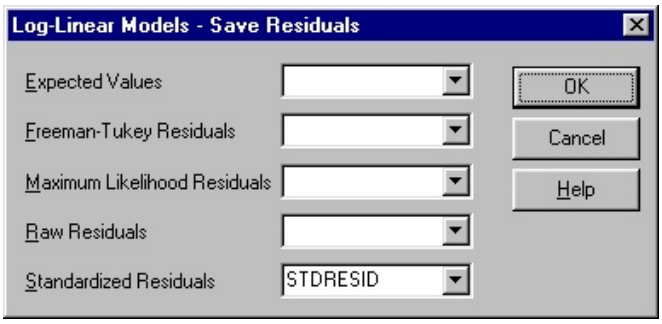
Once you've specified the analysis and obtained the results, a ***Results*** menu appears on the menu at the top of the main *Statistix* window. After you've viewed the log-linear results, click on the Results menu to access the log-linear results menu.



Select Coefficient Table to redisplay the log-linear results table. Select Options to return the log-linear dialog box used to specify the model. The Save Residuals option is discussed below.

Save Residuals

The **Save Residuals** command is used to save residuals and expected values as new variables. You can save standardized, maximum likelihood, Freeman-Tukey, or raw residuals. You save residuals by entering variable names in the dialog box below.



Enter a new or existing variable name in the field next to the residual or expected value you want to save. You can press the down-arrow button next to a box to display the list of existing variables. You can select as many options as you like. When you've entered all the names you want, press the ***OK*** button.

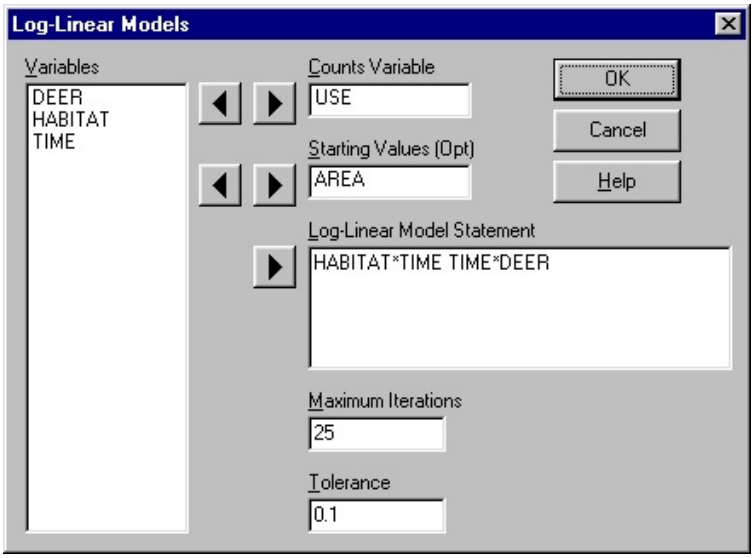
Residuals are useful for diagnosing why a model doesn't fit the data well, which in turn may suggest models that will fit well. When the model is correct, the Freeman-Tukey and standardized residuals are approximately normally distributed with a mean of zero and a variance slightly less than one.

Specifying Starting Values

Occasionally, it's useful to specify starting values for the estimated cell counts. One such circumstance is when the cell counts follow a Poisson distribution and the counts have been taken from different reference populations. Haberman (1978) and Heisey (1985) give some examples of such analyses; the starting values will often be non-integer values.

Heisey (1985) gives an example where the goal is to examine factors affecting the preferences shown by white-tailed deer for different habitat types. Frequent use of a habitat may result from a preference for that

habitat or from that habitat just being common, so it's desirable to adjust for the areas of habitats available to the animals. In Heisey's study, the categorical variable indicating the habitat type was called HABITAT and the areas of the different habitats in the study area were called AREA. The categorical variables ("factors") were DEER and TIME—the animal under observation and the time of day, respectively. The number of times a deer was found in a particular habitat was USE. The best model was found to be:



The main point of interest is that AREA was adjusted for by using it as the starting values for the estimated expected cell frequencies. Consult Heisey (1985) for more detail.

Another situation where the user will want to specify initial values is for models of quasi-independence; the starting values will be either zero or one (Fienberg 1980). Quasi-independence is often of interest for tables that by definition must include empty cells, or so-called structural zeros. Incomplete tables are handled in a similar way; missing cells are given a starting value of zero. A note is given in the results when initial values are specified.

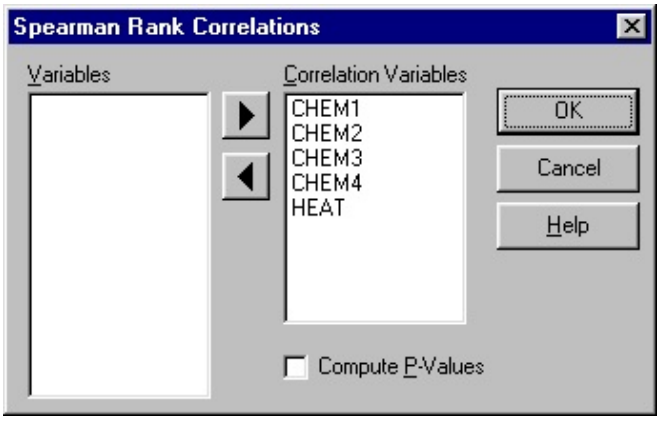
Computational Notes

Iterative proportional fitting is used to estimate the expected cell frequencies. The program was generally patterned after Haberman (1972).

Spearman Rank Correlations

The **Spearman Rank Correlations** procedure computes the Spearman rank correlation coefficient matrix for a list of variables. Typical significance tests for simple correlations usually require that the samples follow a bivariate normal distribution. This is often a difficult assumption to justify, especially for ordinal data, i.e., ranks. Spearman rank correlations are suitable for examining the degree of association when the samples violate the assumption of bivariate normality.

Specification



Select the variables for which you want to compute rank correlations. Highlight the variables you want to select in the ***Variables*** list box, then press the right-arrow button to move them to the ***Correlation Variables*** list box. To highlight all variables, click on the first variable in the list, and while holding the mouse button down, drag the cursor to the last variable in the list. Check the ***Compute P-Values*** check box to have p-values for the correlation coefficients computed and reported.

Data Restrictions

Up to 50 variables can be specified. If a case in your data has missing values for any variable, the entire case is deleted (listwise deletion).

Example

We use the data of Hald (1952) used in Draper and Smith (1981) for our example. This data set is also used for the example for the **Correlations** procedure in Chapter 6. The variable HEAT is the cumulative heat of hardening for cement after 180 days. The variables CHEM1, CHEM2,

CHEM3, and CHEM4 are the percentages of four chemical compounds measured in batches of cement. The data are listed below, and are stored in the file Sample Data\Hald.sx.

CASE	HEAT	CHEM1	CHEM2	CHEM3	CHEM4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

The analysis is specified on the preceding page. The results are as follows:

```
Spearman Rank Correlations, Corrected for Ties

            CHEM1     CHEM2     CHEM3     CHEM4
CHEM2      0.3301
CHEM3     -0.7186    0.0527
CHEM4     -0.3320   -0.9903   -0.0806
HEAT       0.7912    0.7373   -0.4488   -0.7521

Maximum Difference Allowed Between Ties  0.00001

Cases Included 13    Missing Cases 0
```

The Spearman correlation coefficient is the usual (Pearson product moment) correlation coefficient computed from the rank scores of the data rather than the original data. If ties are found when the data are ranked, the average rank is assigned to the tied values, as suggested by Hollander and Wolfe (1973). Values are considered to be tied if they are within 0.00001 of one another. A message “corrected for ties” is displayed in the first line of the report when ties are found.

A similar nonparametric correlation coefficient is Kendall’s tau. In most cases, inference based on Kendall’s tau will produce results nearly identical to that based on Spearman’s rho (Conover 1980).

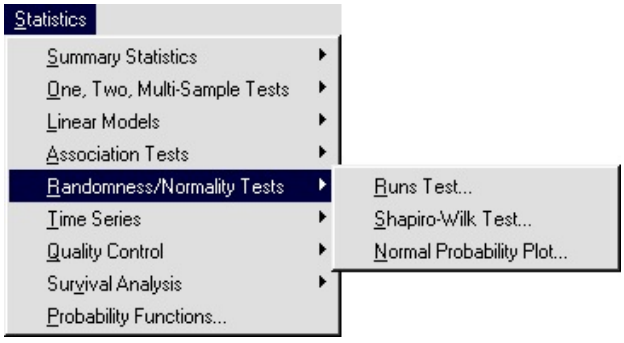
Computational Notes

The ranks are first computed for the data. The correlations are then computed with the same procedures used to produce the Pearson correlations. See **Correlations (Pearson)** in Chapter 6 for more detail.

C H A P T E R

9

Randomness/Normality Tests



Statistix offers three procedures for testing data for randomness and normality.

The **Runs Test** is useful for examining whether samples have been drawn at random from a single population. It can detect patterns that often result from autocorrelation.

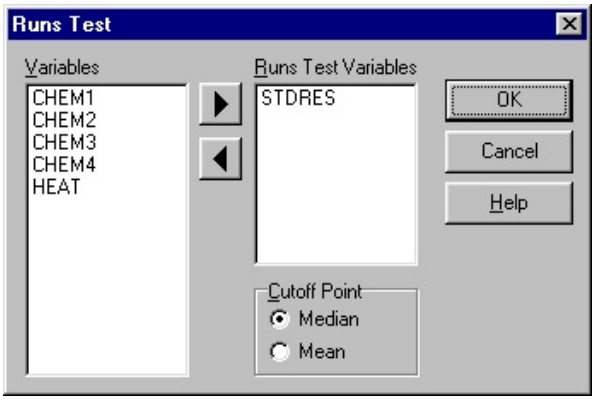
The **Shapiro-Wilk Test** calculates the Shapiro-Wilk statistic with p-value used to test whether data conform to a normal distribution.

The **Normal Probability Plot** produces a rankit plot, also useful for examining whether data conform to a normal distribution.

Runs Test

The **Runs Test** procedure examines whether the number of runs found in a variable are consistent with the hypothesis that the samples are order-independent. A run is defined as consecutive samples that are either consistently above or below the sample median (mean). Autocorrelation often results in more or fewer runs than if the samples are independent.

Specification



Select the names of the variables that you want to test for runs. A separate runs test is performed for each variable you move to the ***Runs Test Variables*** list box.

Choose either the median or the mean for the ***Cutoff Point***. Using the mean can work better than the median when the data contain only a few discrete values that would result with a large number of ties with the median.

Example

We'll apply the runs test to residuals resulting from a linear regression analysis. The data—the Hald data from Draper and Smith (1981)—are used for the example data set in **Linear Regression**. Standardized residuals were computed for the regression model HEAT = CHEM1 CHEM4 and stored in the variable STDRES. The runs test is specified by selecting the variable name as shown on the preceding page. The results are presented on the next page.

Runs Test for STDRES

Median	0.0553	
Values Above the Median	6	
Values below the Median	6	
Values Tied with the Median	1	
Runs Above the Median	4	
Runs Below the Median	4	
Total Number of Runs	8	
Expected Number of Runs	7.0	
P-Value, Two-Tailed Test		0.7835
Probability of getting 8 or fewer runs		0.8247
Probability of getting 8 or more runs		0.3918

A value was counted as a tie with the Median if it was within 0.00001

Cases Included 13 Missing Cases 0

In residual analysis, you're more likely to see too few runs rather than too many. Too few runs result from runs generally being too long, which indicates positive autocorrelation. When observed in residuals, this could be because you didn't include important explanatory variables in the model. The one-tailed p-value of 0.8247 indicates that there is no evidence for too few runs in these residuals.

Too many short runs is less common, and results from negative autocorrelation. An example where negative autocorrelation can be expected is in a situation where a process is being constantly monitored and adjusted, and there's a tendency toward overcompensation when adjustments are made. The probability of getting eight or more runs in our example is 0.3918, which again is no cause for alarm.

The **Durbin-Watson** statistic in Linear Regression is also very important for diagnosing autocorrelation in regression residuals.

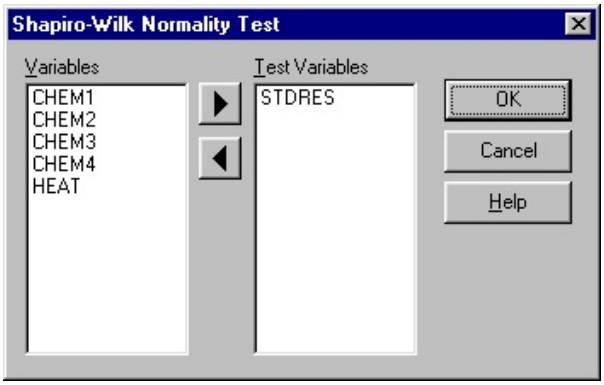
Computational Notes

The equations used to calculate the runs probabilities can be found in Bradley (1968, p. 254). These exact equations are used unless the number of values above or below the median exceeds 20, in which case normal approximations are used (p. 262).

Shapiro-Wilk Normality Test

The **Shapiro-Wilk Test** examines whether data conforms to a normal distribution. The W statistic and corresponding p-values are calculated.

Specification



Select the variables you want to test and move them to the ***Test Variables*** box.

Example

We'll apply the Shapiro-Wilk test to residuals resulting from a linear regression analysis. The data—the Hald data from Draper and Smith (1981)—are used for the example data in **Linear Regression**. Standardized residuals were computed for the regression model HEAT = CHEM1 CHEM4 and stored in the variable STDRES. If the assumptions of linear regression are met, the standardized residuals should be approximately normally distributed with mean 0 and variance 1. The variable STDRES is selected for the Shapiro-Wilk test in the dialog box above. The results are presented below.

Shapiro-Wilk Normality Test

Variable	N	W	P
STDRES	13	0.9668	0.8541

The W statistic approaches one for normally distributed data. We reject the null hypothesis that the data are normally distributed when the p-value is small (e.g., smaller than 0.05). We conclude from the p-value of 0.8541 for

the variable STDRES that the residuals are normal and that this assumption for linear regression has been satisfied.

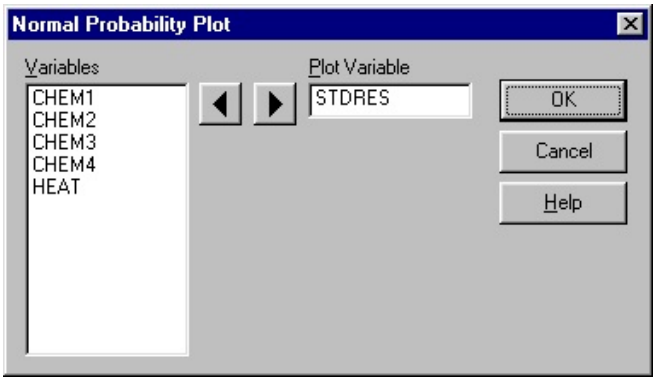
Computational Notes

Algorithms used for the W statistic and the p-value are given in Royston (1995).

Normal Probability Plot

The normal probability plot, also called a rankit plot, plots the ordered data points against the corresponding rankits. When the data plotted are drawn from a normal population, the points appear to fall on a straight line. The Shapiro-Wilk W statistic is also reported on the plot.

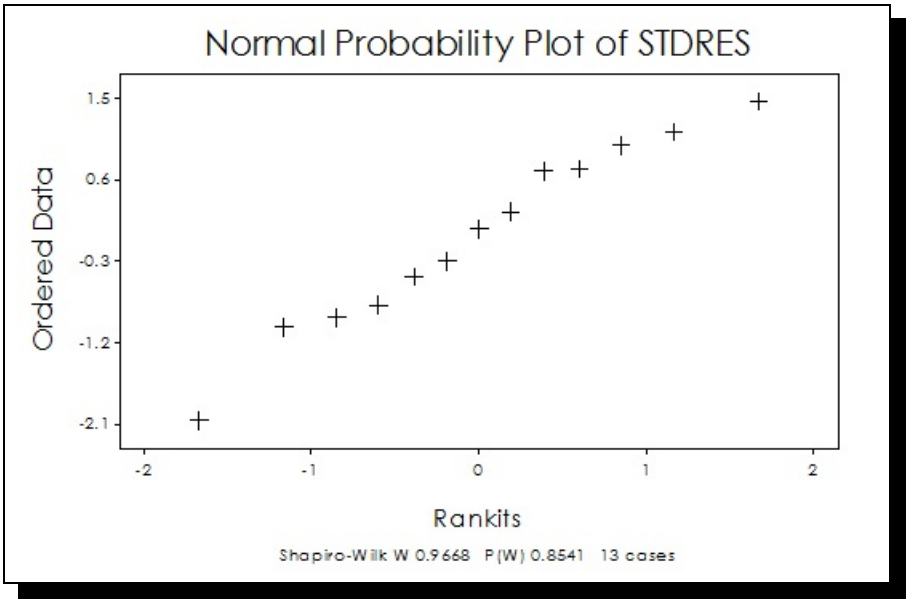
Specification



Select the name of the variable you want to plot and move it to the ***Plot Variable*** box.

Example

We'll apply the Normal Probability Plot procedure to residuals resulting from a linear regression analysis. The data—the Hald data from Draper and Smith (1981)—are used for the example data in **Linear Regression**. Standardized residuals were computed for the regression model HEAT = CHEM1 CHEM4 and stored in the variable STDRES. The variable STDRES is selected as the plot variable in the dialog box above. The results are presented no the next page.



If the assumptions of linear regression are met, the standardized residuals should be approximately normally distributed with mean 0 and variance 1. The i-th rankit is defined as the expected value of the i-th order statistic for the sample, assuming the sample was from a normal distribution. The order statistics of a sample are the sample values reordered by their rank. If the sample conforms to a normal distribution, a plot of the rankits against the order statistics should result in a straight line, except for random variation.

Systematic departure of the normal probability plot from a linear trend indicates non-normality, as does a small value for the Shapiro-Wilk W statistic (see page 304). The example plot above shows no evidence of non-normality. One or a few points departing from the linear trend near the extremes of the plot are indicative of outliers. Consult Daniel and Wood (1971), Daniel (1976), and Weisberg (1985) for more detail.

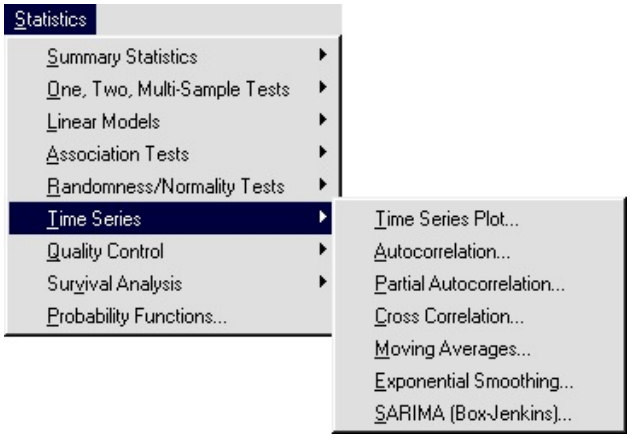
Computational Notes

Rankits are computed with an algorithm similar to Royston's (1982) NSCOR2. The procedure for calculating the required percentage points of the normal distribution and the Shapiro-Wilk W statistic uses the algorithms provided by Royston (1995).

C H A P T E R

10

Time Series



A time series is a list of observations collected sequentially, usually over time. Common time series subjects are stock prices, population levels, product sales, rainfall, and temperature. It's assumed that the observations are taken at uniform time intervals, such as every day, month, or year. Not all time series occur over time. For example, a list of diameters taken at every meter along a telephone cable is a legitimate time series.

Observations in a time series are often sequentially dependent. For example, population levels in the future often depend on levels at present and in the past. The goal of time series analysis is to model the nature of these dependencies, which in turn allows you to predict, or forecast,

observations that have not yet been made.

The **Time Series Plot** procedure is used to create a time series plot for one or more variables.

The **Autocorrelation Plot** procedure is used to create an autocorrelation plot for a specified variable.

The **Partial Autocorrelation Plot** procedure is used to create a partial autocorrelation plot for a specified variable.

The **Cross Correlation** procedure is used to create a cross correlation plot for two variables.

The **Moving Averages** procedure is used to compute forecasts for time series data based on moving averages.

The **Exponential Smoothing** procedure computes forecasts for time series data using exponentially weighted averages.

The **SARIMA** procedure allows you to fit a variety of models to data, including both nonseasonal and multiplicative, and nonmultiplicative seasonal models.

Model Building

The methods described by Box and Jenkins (1976) is a popular tool for modeling time series. The Box-Jenkins approach assumes that the time series can be represented as an ARIMA process, which stands for AutoRegressive Integrated Moving Average.

Box and Jenkins advocate an iterative three-step approach to model building:

1) Identification of terms to be included in the model
2) Parameter estimation
3) Model evaluation

Model term identification relies on the use of Time Series Plots, Autocorrelation Plots, and Partial Autocorrelation Plots. These plots are examined to suggest what transformations, differencing operators, AR terms, and MA terms should be included in the model.

Once a tentative model has been identified, parameter estimation is accom-

plished with the SARIMA procedure, which uses the unconditional least squares method (sometimes called the backcasting method) to fit the identified model to the series.

Model evaluation is accomplished by examining the results of the SARIMA fitting. If the model isn't adequate, a new tentative model is identified and the process repeated until a good model is found. A good model can then be used to forecast future observations. Box and Jenkins (1976) should be consulted for more details.

The forecasting procedures Exponential Smoothing and Moving Averages are easier to use and understand than ARIMA models.

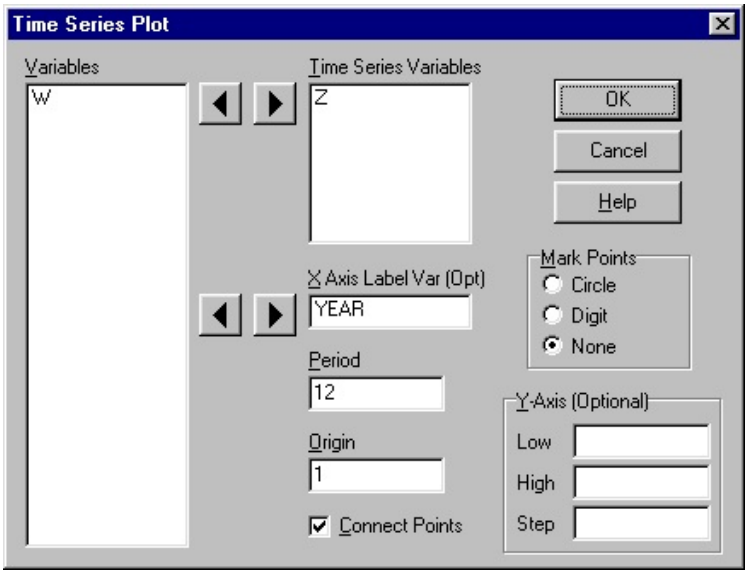
Treatment of Missing or Omitted Cases

Time series data sets can't have embedded missing values or omitted cases. There can be blocks of missing or omitted cases at the beginning and end of the data set. *Statistix* time series procedures use the first continuous block of data that doesn't contain missing values or omitted cases for the series.

Time Series Plot

The **Time Series Plot** procedure is used to create a time series plot for one or more variables. The values of the variables are plotted in case order.

Specification



Select the names of one or more time series variables. If you select more than one name, all the variables will be plotted on a single time series plot using different colors, line patterns, and point symbols.

Normally, the points along the X axis are labeled as case numbers starting with 1. You can customize the X axis labels specifying an ***X Axis Label Var*** containing the labels. The label variable can be a string, date, integer, or real variable. Strings are truncated to ten characters.

Additional options are available to control the appearance and labeling of the plot. Points on a time series plot are usually connected by line segments. You control this using the ***Connect Points*** check box. Use the ***Mark Points*** radio buttons to have the individual points marked with a circle, a digit, or no mark at all.

Use the ***Origin*** option to change the starting case number for the X axis labels. For example, if you have annual data starting with 1955, enter 1955 in the Origin edit control to have the axis labeled 1955, 1956, and 1957

instead of 1, 2, and 3.

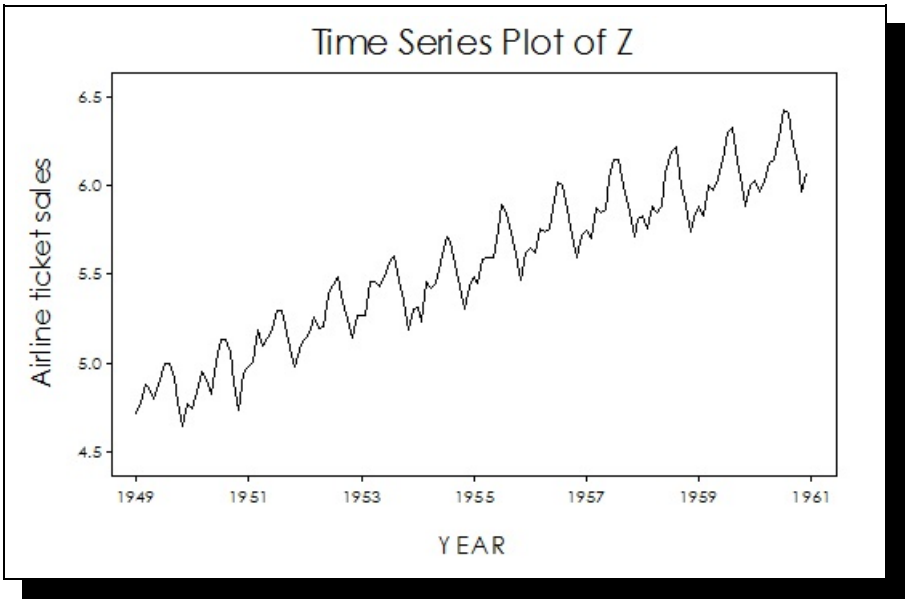
Numbering the points using digits is useful for identifying seasonal trends. You should enter a value for ***Period*** that reflects the frequency of data collection. For example, enter 4 for quarterly data or a 12 for monthly data. A seasonal trend stands out clearly when the same digit appears in either peaks or valleys.

The Period also affects how often the X axis is labeled. The example dialog box on the preceding page specifies a period of 12, which is suitable for labeling monthly data. The X axis will be labeled every 12 cases.

You can enter values for low, high, and step to control the ***Y-Axis*** scale.

Example

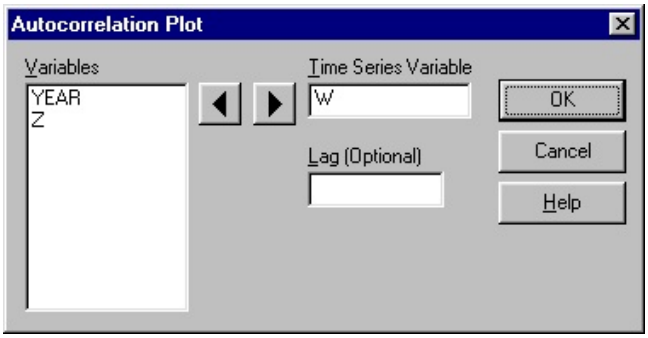
The example data are the natural logs of monthly passenger totals (in thousands) in international air travel for 12 years from January 1949 to December 1960 (Box and Jenkins 1976, p. 304). You can view the data by opening Sample Data\airline.sx. A total of 144 cases are in the variable named Z. The variable YEAR was created to annotate the X axis of our example plot—the year was entered for every 12th case. The dialog box on the preceding page is used to obtain the time series plot below.



Autocorrelation

The **Autocorrelation Plot** procedure is used to create an autocorrelation plot for the specified variable. Approximate 95% confidence intervals are also shown.

Specification



Select the name of a time series variable and move it to the ***Times Series Variable*** box. You can specify a lag in the ***Lag*** edit control. If you don't specify a lag, the maximum lag is used, which is calculated as the square root of the sample size plus five.

Example

We use the data from Box and Jenkins (1976, p. 304) for our example (see the sample file Sample Data\airline.sx). The data are the natural logs of monthly passenger totals (in thousands) in international air travel for 12 years from January 1949 to December 1960. A total of 144 cases are in a variable named Z. The variable W is created by first seasonally differencing Z and then nonseasonally differencing the seasonal difference. Using **Transformations** from the ***Data*** menu, W is created in two steps by:

```
W = Z - LAG (Z, 12)
W = W - LAG (W)
```

That is, $W = DD^{12}Z$, where D is the differencing operator.

The dialog box above shows the entries for the differenced airline data. A specific lag is not specified, so the maximum lag is used. The results are presented in the autocorrelation plot on the next page.

```
Autocorrelation Plot for W

            -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
 Lag    Corr  +----+----+----+----+----+----+----+----+----+----+----+
   1  -0.341                  *****<****    <
   2   0.105                      >    **** <
   3  -0.201                     <*****    <
   4   0.020                      >    **   <
   5   0.056                      >    **   <
   6   0.031                      >    **   <
   7  -0.056                      >   **    <
   8   0.001                      >    *    <
   9   0.174                      >    *****<
  10  -0.076                      >  ***    <
  11   0.064                      >    ***  <
  12  -0.387                  *****<*****    <
  13   0.154                     >     ***** <
  14  -0.060                     >    **     <
  15   0.151                     >     ***** <
  16  -0.139                     >  ****     <

Mean of the Series   2.824E-04
Std Dev of Series      0.04565
Number of Cases            131
```

The first column indicates the lag for which the autocorrelation is computed. The next column displays the value of the autocorrelation. The autocorrelation is displayed graphically as a horizontal bar. Approximate 95% confidence bounds are indicated with angled brackets (< >). The direction in which the confidence bounds point indicates where the observation lies relative to the confidence bound. For example, with a lag of 3, the autocorrelation is -0.201, which lies outside of the confidence bound to the left.

The confidence intervals for the lag p are based on the assumption that autocorrelations for lags p and greater are zero.

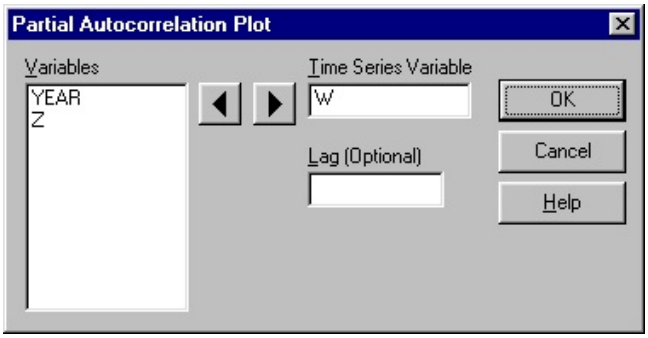
Computational Notes

Computations follow those outlined in Box and Jenkins (1976).

Partial Autocorrelation

The **Partial Autocorrelation Plot** procedure is used to create a partial autocorrelation plot for the specified variable. Approximate 95% confidence intervals are also displayed.

Specification



Select the name of a ***Time Series Variable***. You may specify a ***Lag***. If you don't, the maximum lag is used, calculated as the square root of the sample size plus five.

Example

We use the data from Box and Jenkins (1976, p. 304) for our example. You can view the data by opening Sample Data\airline.sx. The data are the natural logs of monthly passenger totals (in thousands) in international air travel for 12 years from January 1949 to December 1960. The dialog box above shows the entries for the differenced airline data in the variable W (page 312). A specific lag isn't specified, so the maximum lag is used. The resulting partial autocorrelation plot is shown on the next page.

The first column indicates the lag for which the partial autocorrelation is computed. The next column displays the value of the partial autocorrelation. The partial autocorrelation is displayed graphically as a horizontal bar. Approximate 95% confidence bounds are indicated with angled brackets (< >). The direction in which the confidence bounds point indicates where the observation lies relative to the confidence bound.

The confidence intervals for lag p are based on the assumption that the series results from an autoregressive process of order p - 1.

```
Partial Autocorrelation Plot for W

            -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
 Lag    Corr  +----+----+----+----+----+----+----+----+----+----+----+
   1  -0.341                  *****<****   <
   2  -0.013                       >   *   <
   3  -0.191                      *<****   <
   4  -0.125                       >****   <
   5   0.034                       >   **  <
   6   0.036                       >   **  <
   7  -0.061                       > ***   <
   8  -0.019                       >   *   <
   9   0.224                       >   ****>**
  10   0.041                       >   **  <
  11   0.046                       >   **  <
  12  -0.341                  *****<****   <
  13  -0.107                       >****   <
  14  -0.077                       > ***   <
  15  -0.021                       >  **   <
  16  -0.138                       >****   <

Mean of the Series  2.824E-04
Std Dev of Series     0.04565
Number of Cases           131
```

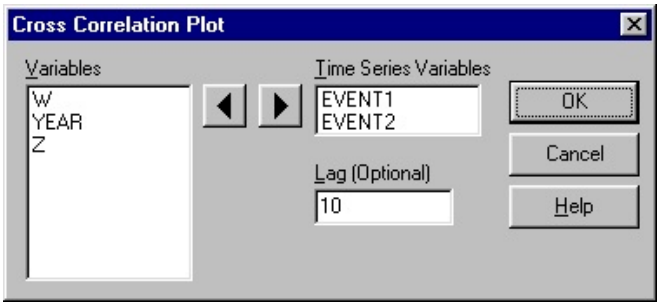
Computational Notes

Computations follow those outlined in Box and Jenkins (1976).

Cross Correlation

The **Cross Correlation** procedure is used to create a cross correlation plot for two variables.

Specification



Select the names of the two ***Time Series Variables***. Unless a ***Lag*** is

explicitly specified, the maximum absolute value of the lag is used, computed as the square root of the sample size plus five.

Example

To demonstrate cross correlation, we'll fabricate some data. The variable EVENT1 is created simply as a list of uniform random numbers. We then create the variable EVENT2 as a moving average process of EVENT1. These two variables are generated in **Transformations** as:

```
EVENT1 = RANDOM

EVENT2 = EVENT1 + LAG (EVENT1) / 2 + LAG (EVENT1, 2) / 3 +
LAG (EVENT1, 6) / 4
```

The dialog box displayed on the preceding page is used to find the cross correlations for EVENT1 and EVENT2 for 10 lags. The results are presented below in the cross correlation plot.

```
Cross Correlation Plot for EVENT1 and EVENT2

            -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
 Lag    Corr  +----+----+----+----+----+----+----+----+----+----+----+
 -10  -0.100                          ***
  -9  -0.014                            *
  -8  -0.008                            *
  -7  -0.032                           **
  -6  -0.073                          ***
  -5   0.013                            *
  -4   0.022                            **
  -3  -0.037                           **
  -2  -0.117                         ****
  -1  -0.119                         ****
   0   0.823                            ************************
   1   0.375                            **********
   2   0.177                            *****
   3  -0.157                        *****
   4  -0.060                           **
   5   0.032                           **
   6   0.201                           ******
   7  -0.053                           **
   8  -0.070                          ***
   9   0.003                            *
  10  -0.033                           **

Mean of Series 1          0.55584
Std. Dev. of Series 1     0.27460
Mean of Series 2          1.16506
Std Dev of Series 2       0.31157
Number of Cases               139
```

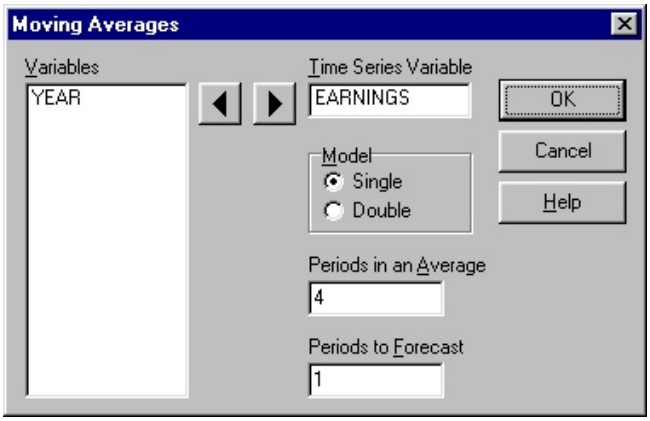
Computational Notes

Computations follow those outlined in Box and Jenkins (1976).

Moving Averages

The **Moving Averages** procedure is used to compute forecasts for time series data. Both single and double moving averages techniques are available. Use single moving averages when there is no trend in the time series data. Use double moving averages when there is a trend in the data,

Specification



Select the name of a ***Time Series Variable***. Then select single or double moving averages using the ***Model*** radio buttons.

Enter a value for the number of periods in the moving average in the ***Periods in an Average*** edit control. In general you can select a value for the period that minimizes the maximum absolute deviation (MAD) or the mean squares of the forecast errors (MSE). The number of periods in the moving average is sometimes selected to remove seasonal effect; 4 is used for quarterly data, 12 for monthly data, and so on.

When using double moving averages, you can also specify the number of future periods for which you want to compute forecasts in the ***Periods to Forecast*** box.

Data Restrictions

There must be enough cases in the time series to compute the moving averages. The minimum number depends on the value you enter for the length of the moving averages. If the number of periods you specify for the moving averages is n, there must be at least n + 1 cases for single moving averages and 2 x n cases for double moving averages.

Example

We illustrate single moving averages using earnings per share for Exxon for

the years 1962 to 1976. The data are stored in Sample Data\Exxon.sx.

YEAR	EARNINGS	YEAR	EARNINGS
1962	1.94	1970	2.96
1963	2.37	1971	3.39
1964	2.44	1972	3.42
1965	2.41	1973	5.45
1966	2.53	1974	7.02
1967	2.77	1975	5.60
1968	2.97	1976	5.90
1969	2.89		

The dialog box on the preceding page shows the moving averages options selected for the Exxon data. The results are:

Single Moving Averages for EARNINGS

Moving Average Length 4

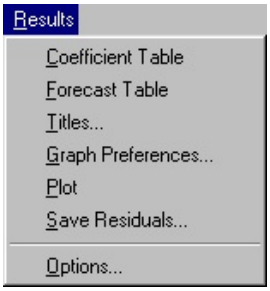
Sum of Squared Errors (SSE)	17.2542
Mean Squared Error (MSE)	1.56857
Standard Error (SE)	1.25242
Mean Absolute Deviation (MAD)	0.82386
Mean Abs Percentage Error (MAPE)	16.84
Mean Percentage Error (MPE)	16.84
Number of Cases	15

Lead	95% C.I. Lower Bound	Forecast	95% C.I. Upper Bound
1	3.53775	5.99250	8.44725

The results list a number of summary statistics that are useful for checking the adequacy of the model. The forecast for the year 1977 is displayed with a 95% confidence interval.

Moving Averages Results Menu

Once you've specified and computed a moving average analysis, a Results menu appears on the menu at the top of the *Statistix* window. Click on Results to access the pull-down menu displayed below.



Selecting Coefficient Table from the menu will redisplay the results

presented on the preceding page. Selecting Options from the menu will return you to the Moving Averages dialog box used to generate these results. Use the Plot option to produce a time series plot of the actual and fitted data. The remaining menu selections are discussed below.

Forecast Table

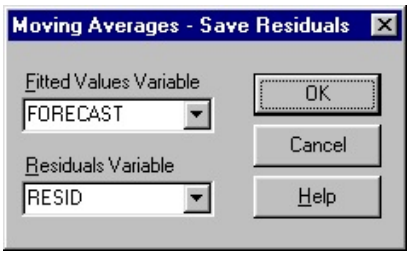
The Forecast Table lists the actual value, moving average, forecast, and forecast error for each time period in the data.

Single Moving Averages Forecast Table for EARNINGS

Time	Actual Value	Moving Average	Forecast	Forecast Error
1	1.94000			
2	2.37000			
3	2.44000			
4	2.41000	2.29000		
5	2.53000	2.43750	2.29000	0.24000
6	2.77000	2.53750	2.43750	0.33250
7	2.97000	2.67000	2.53750	0.43250
8	2.89000	2.79000	2.67000	0.22000
9	2.96000	2.89750	2.79000	0.17000
10	3.39000	3.05250	2.89750	0.49250
11	3.42000	3.16500	3.05250	0.36750
12	5.45000	3.80500	3.16500	2.28500
13	7.02000	4.82000	3.80500	3.21500
14	5.60000	5.37250	4.82000	0.78000
15	5.90000	5.99250	5.37250	0.52750

Save Residuals

Use the Save Residuals procedure to save the fitted values (forecasts) and/or residuals (forecast errors) in new or existing variables for later analysis. Enter a variable name in the ***Fitted Values Variable*** box and/or the ***Residuals Variable*** box.



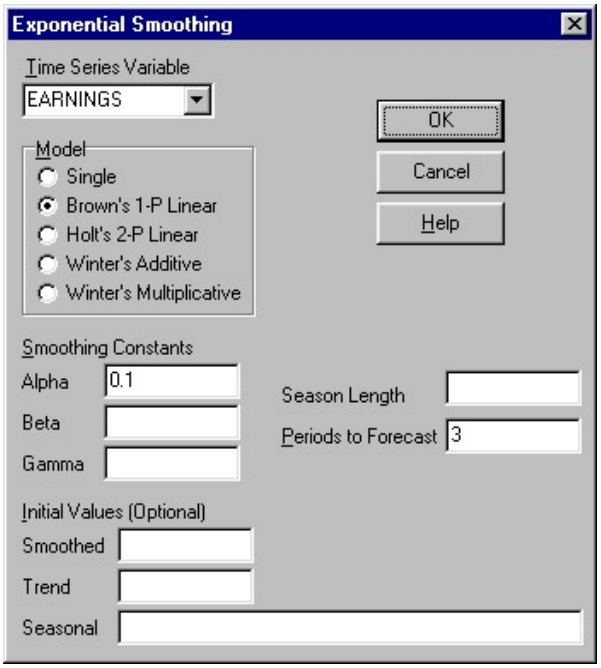
Computational Notes

Computations follow those described in Hanke and Reitsch (1989) and Mercier (1987).

Exponential Smoothing

The **Exponential Smoothing** procedure computes forecasts for time series data using exponentially weighted averages. Five different models are available to fit no trend data, trend data, and trend with seasonal data.

Specification



First select the name of a ***Time Series Variable***. Press the down-arrow button to display the list of variables in your data set.

Next select the model you want to use from the ***Model*** radio buttons. Single exponential smoothing is used when the time series data doesn't exhibit any trend. The smoothing constant ***Alpha*** determines how much past observations influence the forecast. A small smoothing constant results in a slow response to new values; a large constant results in a fast response to new values. Values for the smoothing constant are normally selected in the range 0.05 to 0.60. For the best model, select a smoothing constant that minimizes the mean squares of the forecast errors (MSE).

Brown's 1-P linear model and Holt's 2-P linear model are used when the data exhibit trend. Brown's model uses one smoothing constant (Alpha) to

smooth both the local average and trend estimates. Holt's model uses two smoothing constants, one for the smoothed local average (Alpha) and one for the trend estimate (Beta).

Winter's models are used when the data exhibit trend and seasonality. The additive model is used when the seasonal influences are additive, that is, the seasonal influences have the same magnitude from year to year. The multiplicative model is appropriate when the seasonal swings are wider for years with higher levels. Both models use three smoothing constants, for the local average (Alpha), trend (Beta), and season (Gamma). You must also specify the ***Season Length***.

All five models require that you enter values for the applicable ***Smoothing Constants***. Some models give you the option of entering ***Initial Values*** and the number of future ***Periods to Forecast***.

Example

Brown's 1-P linear exponential smoothing is illustrated using earnings per share data for Exxon for the years 1962 to 1976. The data are listed on page 318, and are stored in the file Sample Data\Exxon.sx.

The dialog box on the preceding page shows the exponential smoothing options selected for the Exxon data. The results are:

Brown's 1-P Linear Exponential Smoothing for EARNINGS

Smoothing Constant 0.10

Sum of Squared Errors (SSE)	9.47382
Mean Squared Error (MSE)	0.63159
Standard Error (SE)	0.79472
Mean Absolute Deviation (MAD)	0.59165
Mean Abs Percentage Error (MAPE)	16.57
Mean Percentage Error (MPE)	-1.81
Number of Cases	15

Forecast (T) = 5.84858 + 0.31442 * T

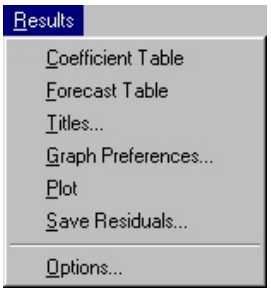
	95% C.I.		95% C.I.
Lead	Lower Bound	Forecast	Upper Bound
1	4.60534	6.16300	7.72066
2	4.91179	6.47742	8.04305
3	5.21789	6.79184	8.36579

The coefficient table above lists a number of summary statistics that are useful for checking the adequacy of the model (Hanke and Reitsch 1989). You can also use the residuals (see ***Save Residuals*** on page 323) to examine the model fit.

The forecasts for the years 1977 to 1979 are displayed with their 95% confidence intervals.

Exponential Smoothing Results Menu

Once you've specified and computed an exponential smoothing analysis, a Results menu appears on the menu at the top of the *Statistix* window. Select the Results menu to access the pull-down menu displayed below.



Select Coefficient Table from the menu to redisplay the results presented above. Select Options from the menu to return to the Exponential Smoothing dialog box used to generate these results. The remaining options are discussed below.

Forecast Table

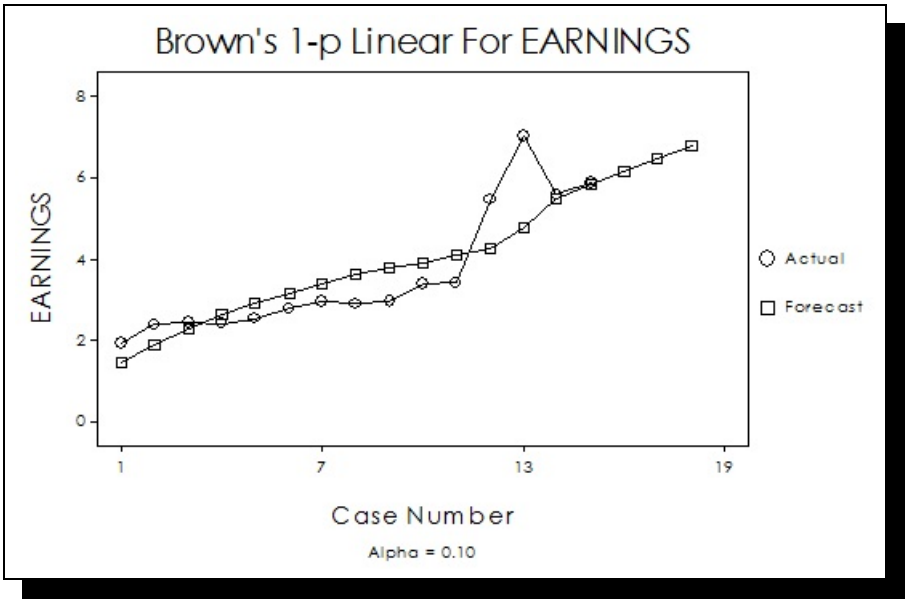
The forecast table lists the actual value, first and second exponential averages, slope, forecast, and forecast error for each time period in the data. The forecast table for the model specified on page 320 is displayed below.

Brown's 1-P Linear Exponential Forecast Table for EARNINGS

Time	Actual Value	1st Exp Average	2nd Exp Average	Trend	Forecast	Forecast Error
0		-1.62107	-4.38729	0.307		
1	1.94000	-1.26496	-4.07505	0.312	1.45250	0.48750
2	2.37000	-0.90147	-3.75769	0.317	1.85736	0.51264
3	2.44000	-0.56732	-3.43866	0.319	2.27212	0.16788
4	2.41000	-0.26959	-3.12175	0.317	2.62305	-0.21305
5	2.53000	0.01037	-2.80854	0.313	2.89948	-0.36948
6	2.77000	0.28633	-2.49905	0.309	3.14249	-0.37249
7	2.97000	0.55470	-2.19368	0.305	3.38120	-0.41120
8	2.89000	0.78823	-1.89549	0.298	3.60845	-0.71845
9	2.96000	1.00541	-1.60540	0.290	3.77014	-0.81014
10	3.39000	1.24387	-1.32047	0.285	3.90630	-0.51630
11	3.42000	1.46148	-1.04228	0.278	4.09313	-0.67313
12	5.45000	1.86033	-0.75201	0.290	4.24343	1.20657
13	7.02000	2.37630	-0.43918	0.313	4.76294	2.25706
14	5.60000	2.69867	-0.12540	0.314	5.50461	0.09539
15	5.90000	3.01880	0.18902	0.314	5.83652	0.06348

Plot

Selecting Plot from the menu produces a time series plot that shows both the actual time series data and the fitted data. The forecasts are also plotted. The plot for the Exxon data is shown below.



Select Titles from the Results menu to change the plot's titles. Select Graph Preferences to change other details of the plot such as colors and point symbols (see Chapter 1 for details).

Save Residuals

Use the Save Residuals procedure to save the fitted values (forecasts) and/or residuals (forecast errors) in new or existing variables for later analysis. Enter a variable name in the ***Fitted Values Variable*** box and/or the ***Residuals Variable*** box.

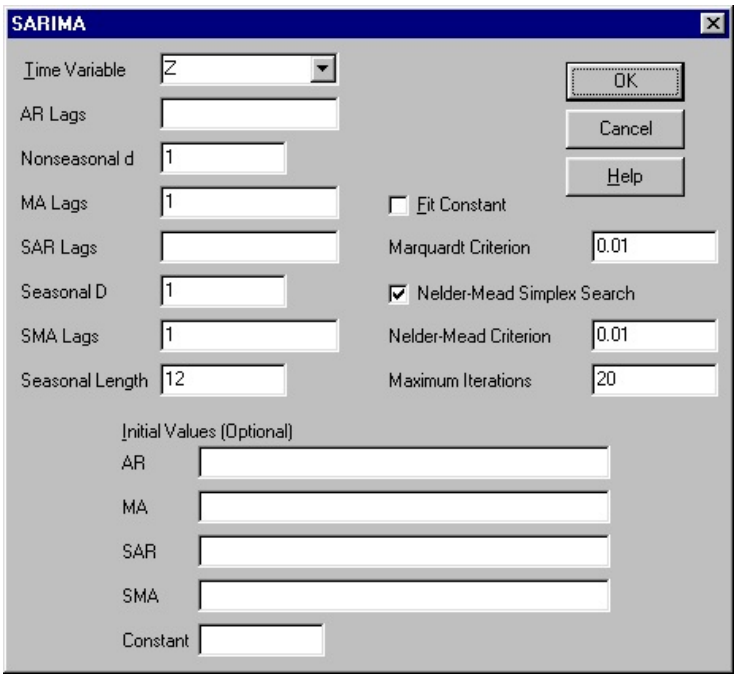
Computational Notes

Computations for the single and linear trend models follow those described in Abraham and Ledolter (1983). Computations for Winter's models follow Thomopoulos (1980).

The **SARIMA** procedure allows you to fit a variety of models to data, including both nonseasonal and multiplicative, and nonmultiplicative seasonal models. SARIMA stands for Seasonal AutoRegressive Integrated Moving Average. It's a procedure for modeling time series popularized by Box and Jenkins (1976). MA and AR terms need not be sequential, so non-significant terms don't need to be included in the model.

Output includes parameter estimates, approximate significance levels, and several statistics useful for diagnosing model fit. You can also obtain forecasts with confidence intervals. You can save fitted values and residuals to evaluate model adequacy. Estimation is based on unconditional least squares, also known as the backcasting method.

Specification



First select the name of your ***Time Series Variable***. You can press the down-arrow button next to the variable box to display the list of your variables. Next you specify the ARIMA model by filling in values in the relevant controls. You can also specify the Marquardt and Nelder-Mead

criteria, the maximum number of iterations, and initial values. Press the ***OK*** button to begin the analysis.

For example, the dialog box on the preceding page fits the ARIMA model $(0, 1, 1)X(0, 1, 1)_{12}$ to the airline data described on the next page.

A description of each of the dialog box fields is listed below:

Time Variable: Variable name for the time series variable
AR Lags: Lags of nonseasonal autoregressive terms in model
Nonseasonal d: Order of nonseasonal differencing
MA Lags: Lags of nonseasonal moving average terms in model
SAR Lags: Seasonal lags of seasonal autoregressive terms in model
Seasonal D: Order of seasonal differencing
SMA Lags: Seasonal lags of seasonal moving average terms in model
Season Length: Period of seasonality
Fit Constant: Model should include a constant term
Marquardt Criterion: Criterion to stop nonlinear least squares
Nelder-Mead Search: Perform simplex search after nonlinear least squares
Nelder-Mead Criterion: Criterion to stop Nelder-Mead simplex search
Maximum iterations: Number of iterations allowed
Initial Values: List of starting values for any of the model parameters

Most of these fields are self-explanatory. Estimation is initially performed with Marquardt's nonlinear least squares procedure. After each iteration, the changes in the parameter estimates are checked. If all parameters have changed less in absolute value than the ***Marquardt Criterion***, the procedure terminates and is assumed to have converged successfully. If you checked the ***Nelder-Mead Simplex Search*** box, a Nelder-Mead simplex search is performed after the Marquardt procedure in an attempt to further reduce the unconditional sums of squares. Unlike the Marquardt criterion, the Nelder-Mead criterion is based on the reduction of the unconditional sums of squares. After each iteration, the reduction in the unconditional sums of squares is checked. If the reduction is less than the number specified for the Nelder-Mead Criterion, the procedure terminates and convergence is assumed. Both Marquardt and Nelder-Mead stop when the number of iterations equals ***Maximum Iterations*** if convergence hasn't occurred.

Note: When specifying AR, MA, SAR, or SMA terms, **all** lags must be specified. For example, to specify an AR(3) model, you would enter "1 2 3", not "3" The latter specification is appropriate if you want the AR coefficients for lags 1 and 2 constrained to zero.

Example

We use the data from Box and Jenkins (1976, p. 304) for our example. You can view the data by opening Sample Data\airline.sx. The data in the variable Z are the natural logs of monthly passenger totals (in thousands) in international air travel for 12 years from January 1949 to December 1960. The variable W is created by first seasonally differencing Z and then nonseasonally differencing the seasonal difference. Using **Transformations** in Data Management, W is created in two steps by:

```
W = Z - LAG (Z, 12)
W = W - LAG (W)
```

That is, $W = DD^{12}Z$, where D is the differencing operator. The dialog box on page 324 specifies the ARIMA model $(0, 1, 1)X(0, 1, 1)_{12}$.

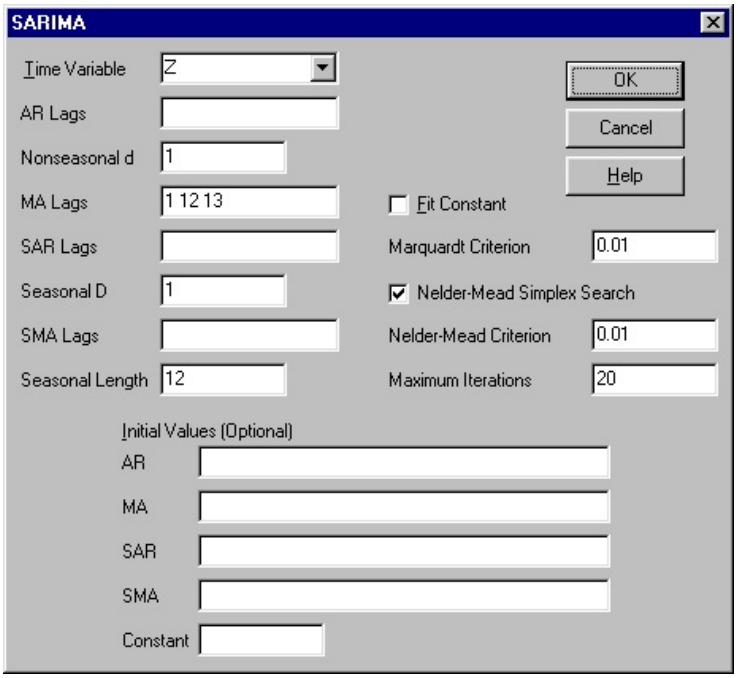
The results are summarized in the coefficient table below.

```
Unconditional Least Squares SARIMA Model for Z

Nonseasonal Differencing of Order 1
Seasonal Differencing of Order 1, Period 12
NOTE: No Constant Term in Model

  Term    Coefficient     Std Error    Coef/SE          P
  MA  1       0.39308       0.08033       4.89     0.0000
  SMA 1       0.61263       0.06941       8.83     0.0000

MS (Backcasts Excluded)     0.00133
DF                             129
SS (Backcasts Excluded)     0.17213   SS Due to Backcasts     0.00387
N Before Differencing          144
N After Differencing           131
Marquardt Criterion of 0.010 was met.
Simplex Criterion of   0.010 was met.
Ljung-Box Portmanteau Lack-of-fit Diagnostics
Lag (DF)   =    12(    10)      24(    22)      36(    34)      48(    46)
Chi-Sq (P) =  9.33(0.5015)   25.43(0.2769)   35.38(0.4028)   44.03(0.5553)
```

Parameter significance can be judged with the t-like statistic Coef/SE. The p-value for this statistic assumes a standard normal distribution. Overall model fit can be judged with the Ljung-Box statistic (Ljung and Box, 1978), which is calculated for lags at multiples of 12. Small p-values indicate that the model fit is poor. In the example above, the p-values are large enough to suggest lack of fit isn't a problem.

As noted, model terms don't need to be sequential. As an example, we'll fit a nonmultiplicative seasonal model to the airline data set. The model we've already used is $DD^{12}z_t = (1 - \Phi B)(1 - \Theta B^{12})a_t$ (D is the differencing operator; other terms are explained in Box and Jenkins 1976). When expanded, this model is $DD^{12}z_t = a_t - \Phi a_{t-1} - \Theta a_{t-12} + \Phi\Theta a_{t-13}$. The more general nonmultiplicative model is $DD^{12}z_t = a_t - \Phi_1 a_{t-1} - \Phi_{12} a_{t-12} - \Phi_{13} a_{t-13}$.

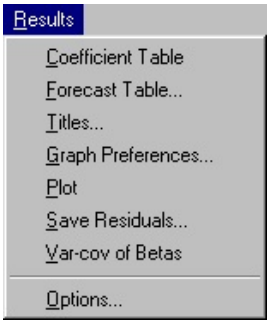
This model is specified in the dialog box below.



While this model has a smaller MS (0.00126) than the multiplicative model, it also requires an additional parameter. Box and Jenkins (1976, p. 324) briefly consider how to examine whether such models are improvements over their multiplicative counterparts.

SARIMA Results Menu

Once you've specified and computed an ARIMA model, a Results menu appears on the menu at the top of the *Statistix* window. Click on Results to access the pull-down menu displayed below.

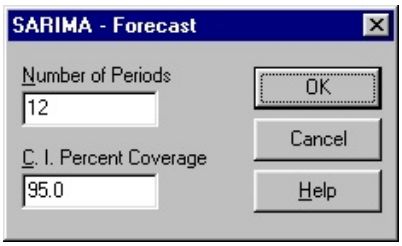


Select Coefficient Table from the Results menu to redisplay the results

presented on page 326. Select Options from the menu to return to the SARIMA dialog box used to generate these results. The remaining options are discussed below.

Forecasts

The Forecast procedure lets you forecast future observations. It also gives confidence intervals for the forecasts.



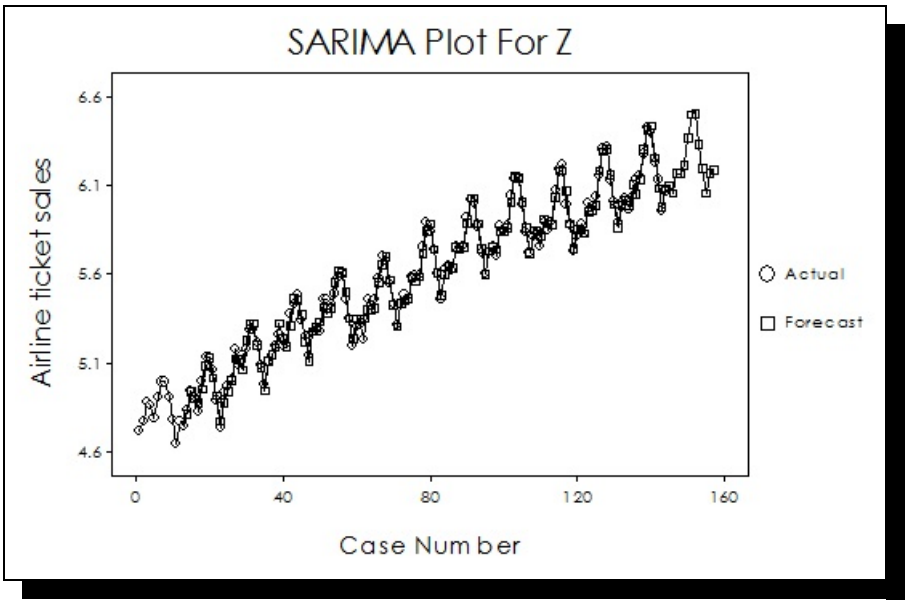
First enter a value for the ***Number of Periods***, which is the number of future time intervals you want to forecast. You can also specify the ***C. I. Percentage Coverage*** for computing confidence levels.

The first 12 forecasts for our airline ticket sales example are presented below.

Lead	95% C.I. Lower Bound	Forecast	95% C.I. Upper Bound
1	6.03779	6.10938	6.18098
2	5.97163	6.05537	6.13912
3	6.08325	6.17760	6.27195
4	6.09434	6.19821	6.30209
5	6.11796	6.23056	6.34315
6	6.24756	6.36825	6.48894
7	6.37619	6.50446	6.63274
8	6.36544	6.50088	6.63631
9	6.18287	6.32510	6.46734
10	6.05870	6.20742	6.35614
11	5.90898	6.06392	6.21886
12	6.00841	6.16933	6.33025

Plot

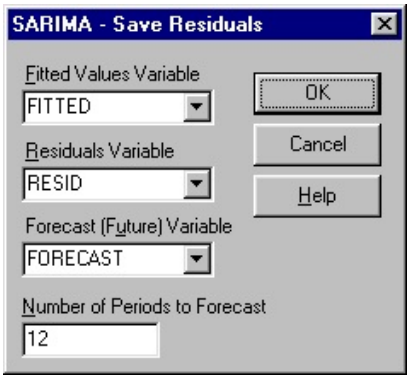
Selecting Plot from the results menu produces a time series plot that shows both the actual time series data and the fitted data. The forecasts are also plotted. The plot for the airline ticket sales data is shown on the next page. The number of future time periods forecast can be changed by using the Forecasts procedure described above.



Select Titles from the Results menu to change the plot's titles. Select Graph Preferences to change other details of the plot such as colors and point symbols (see Chapter 1 for details).

Save Residuals

Use the Save Residuals procedure to save the fitted values, residuals (forecast errors), or forecasts for later analysis.



Because of differencing and lagging, fitted values and residuals may have fewer cases than the original series. The ***Fitted Values Variable*** stores fitted values corresponding to cases used in the analysis. The ***Forecast (Future) Variable*** is used to save forecasts corresponding to time periods

after the cases used for the analysis. Enter the ***Number of Periods to Forecast*** when using the Forecast (Future) Variable.

Variance-Covariance Matrix

The Variance-covariance matrix selection displays the variance-covariance matrix of the estimated model coefficients.

```
Variance - Covariance Matrix for Coefficients

                MA  1       SMA 1
  MA  1       0.00645
  SMA 1    -3.165E-04     0.00482
```

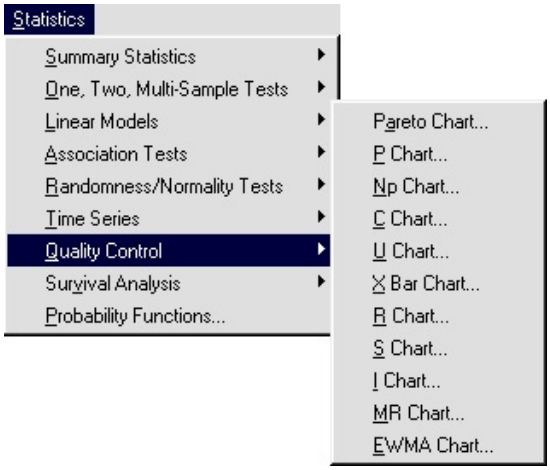
Computational Notes

Computations generally follow those outlined in Box and Jenkins (1976). Initial values for the nonseasonal AR parameters are computed as described on page 499. These, along with the mean of the series, are used to construct an initial estimate of the constant (Box and Jenkins 1976, p. 500) if the model contains one. All other parameters are initially set to 0.1. The Marquardt procedure follows Box and Jenkins (1976) with modifications suggested by Nash (1979). Numerical derivatives use "Nash's compromise" (Nash, 1979, eq. 18.5). The Nelder-Mead procedure is patterned after Nash's outline.

C H A P T E R

11

Quality Control



Statistix offers a number of quality control or statistical process control (SPC) procedures. SPC methods are used to improve the quality of a product or service by examining the process employed to create the product or service.

The **Pareto Chart** procedure produces a Pareto chart, which is used in SPC to identify the most common problems or defects in a product or service. It is a histogram with the bars sorted by decreasing frequency.

A control chart plots a measurement sampled from a process by sample

number over time. A center line is drawn to represent the average value of the quality characteristic when the process is stable, that is, "in control". Two lines are drawn on the control chart to represent the upper and lower control limits (UCL and LCL). A point that falls outside the control limits is evidence that the process is "out of control". *Statistix* uses "3-sigma" control limits that are computed as the center line value plus or minus three times the standard deviation of the process statistic being plotted.

A quality characteristic that can't be measured on a quantitative scale but can be classified as conforming or nonconforming, is called an *attribute*. Consider a cardboard juice container as an example: The seams are either conforming (will not leak) or nonconforming (will leak). *Statistix* computes four attributes control charts—the p chart, np chart, c chart, and u chart.

The **P Chart** procedure plots the fraction nonconforming. The **Np Chart** procedure plots the number nonconforming.

The **C Chart** procedure plots the number of nonconformities per inspection unit (e.g., flaws on the finish of a television set). The **U Chart** procedure plots the average number of nonconformities per unit.

A quantitative quality characteristic, such as the diameter of piston rings, is called a *variable*. *Statistix* computes six control charts for variables—the X bar chart, R chart, S chart, I chart, MR chart, and EWMA chart.

The **X Bar Chart** procedure plots the average of samples; it's used to control the process average of a variable.

The **R Chart** procedure plots the sample range. The **S Chart** procedure plots the sample standard deviation. These plots are used to control the variability of a process.

The **I Chart** procedure plots individuals—variables with a sample size of one. The **MR Chart** procedure plots the moving range of individuals.

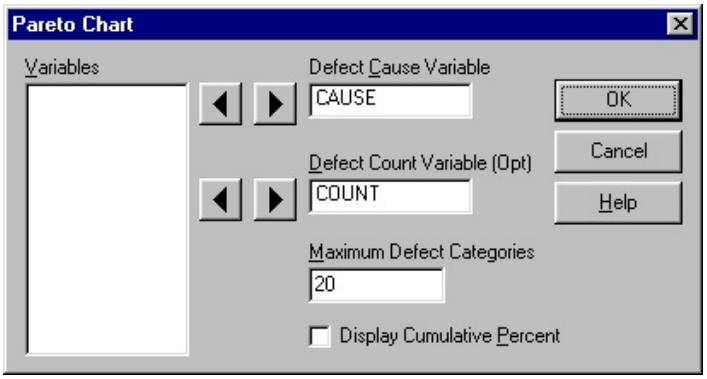
The **EWMA Chart** procedure plots an exponentially weighted moving average. It can be used to control the process mean using individuals or sample averages.

See Montgomery (1991) for computational details for all of the procedures discussed in this chapter.

Pareto Chart

The **Pareto Chart** is used to identify the most frequent causes of defects. It displays a histogram with the bars ordered by frequency.

Specification



First move the name of the variable that contains the defect classifications to the ***Defect Cause Variable*** box. This variable can be of any type—real, integer, date, or string. Strings are truncated to ten characters.

You can enter your data one defect at a time, so that each case in your data set represents one defect. These types of data are often tabulated as they are collected, in which case it's more convenient to enter the data in two columns, one for the defect cause and one for the count of defects. In this case you must move the name of the variable containing the counts of defects for each cause to the ***Defect Count Variable*** box.

You can limit the number of bars by entering a value in the ***Maximum Defect Categories*** edit control. If there are more categories than this value, the least frequently recorded categories are excluded from the chart.

Check the ***Display Cumulative Percent*** check box to have a cumulative distribution curve drawn on the Pareto chart.

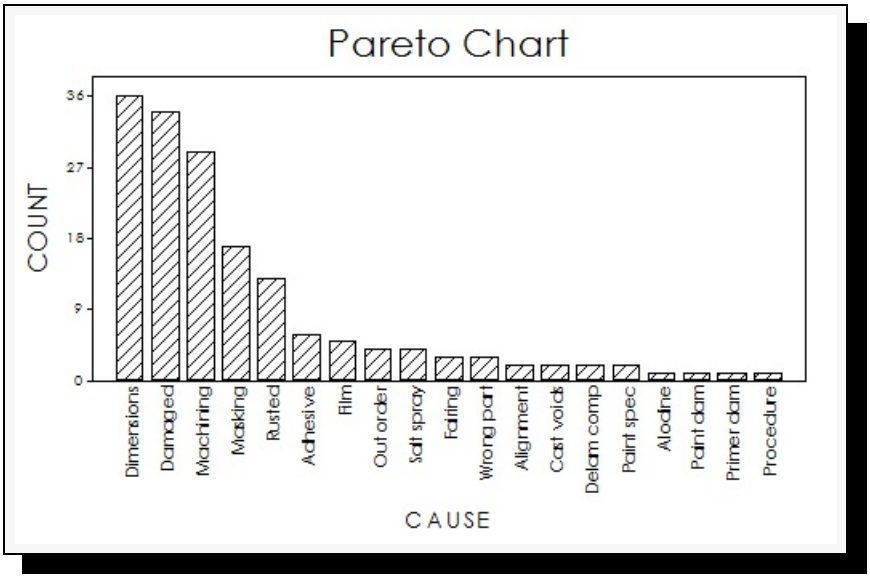
Example

We use data from Montgomery (1991, p. 119) for our example. The various reasons that aircraft tanks were classified defective were collected over several months. The data were entered into *Statistix* using two variables. The variable CAUSE is used to identify the defect cause, and the variable

COUNT records the count for each cause.

CASE	CAUSE	COUNT
1	Adhesive	6
2	Alignment	2
3	Alodine	1
4	Cast voids	2
5	Damaged	34
6	Delam comp	2
7	Dimensions	36
8	Fairing	3
9	Film	5
10	Machining	29
11	Masking	17
12	Out order	4
13	Paint dam	1
14	Paint spec	2
15	Primer dam	1
16	Procedure	1
17	Rusted	13
18	Salt spray	4
19	Wrong part	3

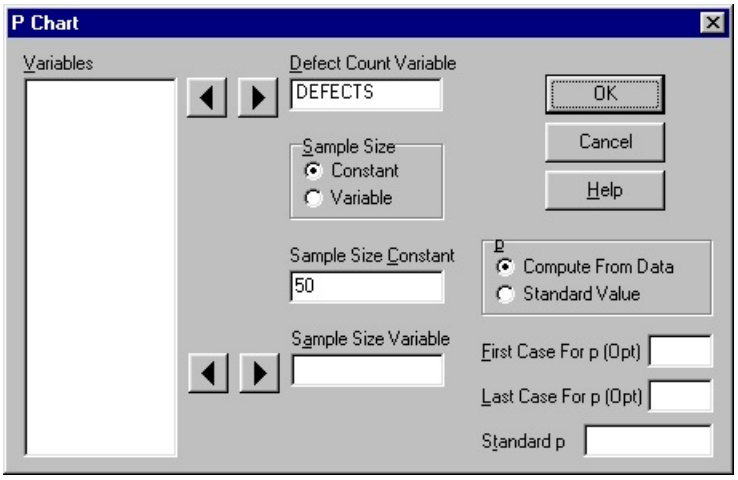
The analysis is specified on the preceding page. The results are as follows:



P Chart

The **P Chart** is the control chart for the fraction of a sample that is nonconforming. P charts are used for attributes—quality characteristics that can be classified as conforming or nonconforming.

Specification



The p chart is computed both from the number of defects per sample and the sample size. First move the name of the variable that contains the counts of defects for each sample to the ***Defect Count Variable*** box. If the sample size is constant, select the ***Constant*** sample size radio button and enter the sample size in the ***Sample Size Constant*** edit control. If the sample size is not always the same, select the ***Variable*** sample size radio button and move the name of the variable that contains the sample size for each case to the ***Sample Size Variable*** box.

The center line and the control limits are computed from p, the fraction nonconforming. You can choose to enter a standard or historical value for p for this purpose, or have p computed from the data. Make your choice by selecting one of the two ***p*** radio buttons. When you select ***Compute From Data***, you may specify the first and last case number used to compute p. If these case numbers are left blank, all cases are used. When you select ***Standard Value***, you must enter a value in the ***Standard p*** edit control.

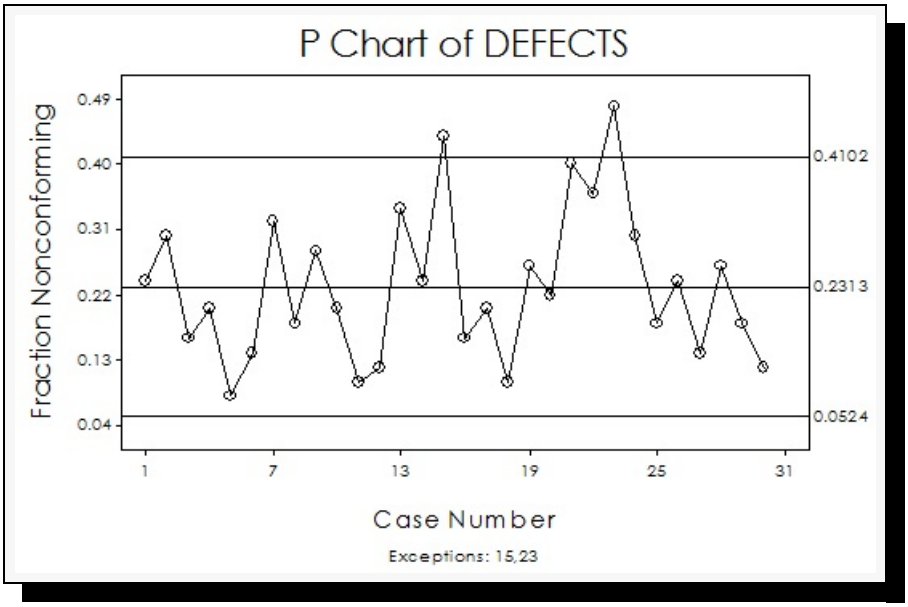
Example

We use data from Montgomery (1991, p. 151) for our example. Cardboard cans being manufactured for orange juice concentrate were sampled from

the machine at half-hour intervals. Each sample contained 50 cans. The number of cans with defective seams were recorded for each sample.

CASE	DEFECTS	CASE	DEFECTS
1	12	16	8
2	15	17	10
3	8	18	5
4	10	19	13
5	4	20	11
6	7	21	20
7	16	22	18
8	9	23	24
9	14	24	15
10	10	25	9
11	5	26	12
12	6	27	7
13	17	28	13
14	12	29	9
15	22	30	6

The p chart model is specified in the dialog box on the preceding page. The results are shown in the figure below.

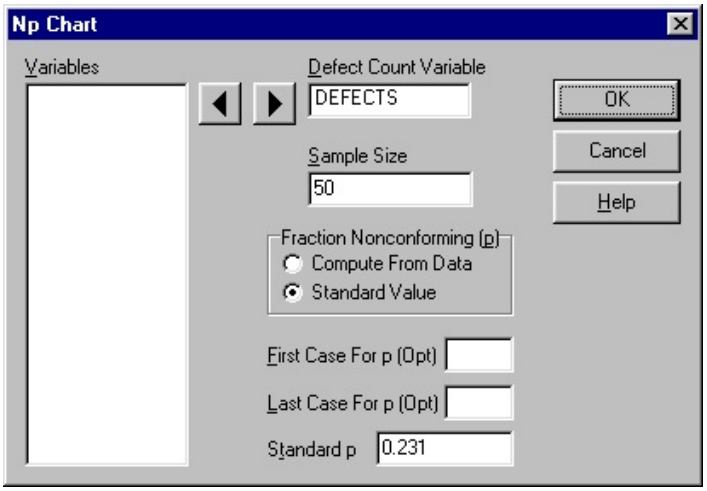


The fraction of defective cans is plotted for each sample. The center line is plotted at the average value for p which is 0.2313. The 3-sigma control limits are labeled on the right side: UCL = 0.4102, LCL = 0.0524. Two values exceed the UCL, indicating that the process is out of control. These case numbers (15 and 23) are noted at the bottom of the chart.

Np Chart

The **Np Chart** is the control chart for the number of nonconforming units. Np charts are used for attributes—quality characteristics that can be classified as conforming or nonconforming. The np chart gives the same results as the p chart discussed earlier (page 335), the only difference is the units of the vertical axis.

Specification



The np chart is computed from the number of defects per sample and the sample size. First select the name of the variable that contains the counts of defects for each sample and move it to the ***Defect Count Variable*** box. The sample size must be a constant. Enter the number in the ***Sample Size*** box.

The center line and the control limits are computed from p, the fraction nonconforming. You can choose to enter a standard or historical value for p for this purpose, or have p computed from the data. Make your choice by selecting one of the two ***p*** radio buttons. When you select ***Compute From Data***, you may specify the first and last case number used to compute p. If these case numbers are left blank, all cases are used. When you select ***Standard Value***, you must enter a value in the ***Standard p*** edit control.

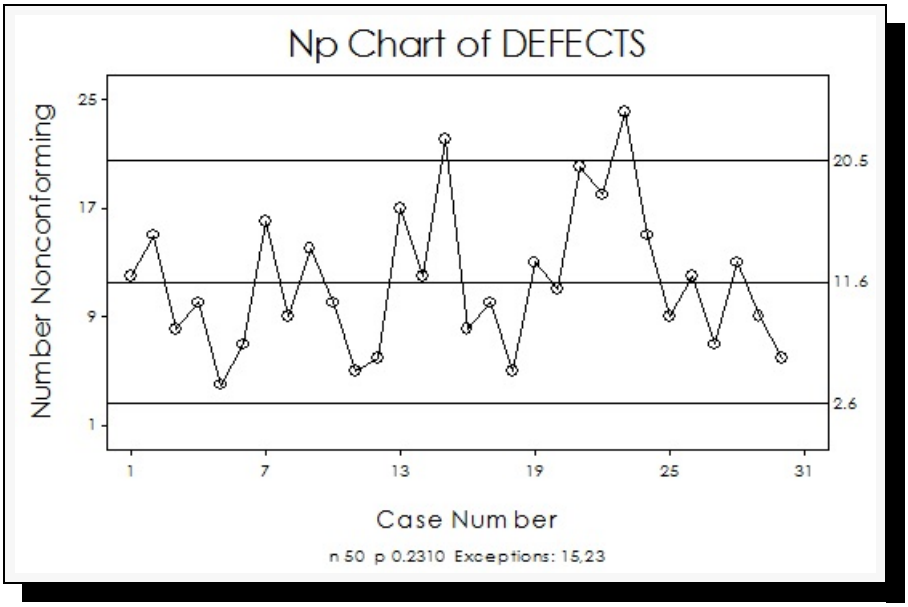
Example

We use data from Montgomery (1991, p. 151) for our example. Cardboard cans being manufactured for orange juice concentrate were sampled from the machine at half-hour intervals. Each sample contained 50 cans. The

number of cans with defective seams were recorded for each sample. The data are listed below, and stored in the file Sample Data\juice.sx.

CASE	DEFECTS	CASE	DEFECTS
1	12	16	8
2	15	17	10
3	8	18	5
4	10	19	13
5	4	20	11
6	7	21	20
7	16	22	18
8	9	23	24
9	14	24	15
10	10	25	9
11	5	26	12
12	6	27	7
13	17	28	13
14	12	29	9
15	22	30	6

The Np chart model is specified in the dialog box on the preceding page. The results are shown in the figure below.

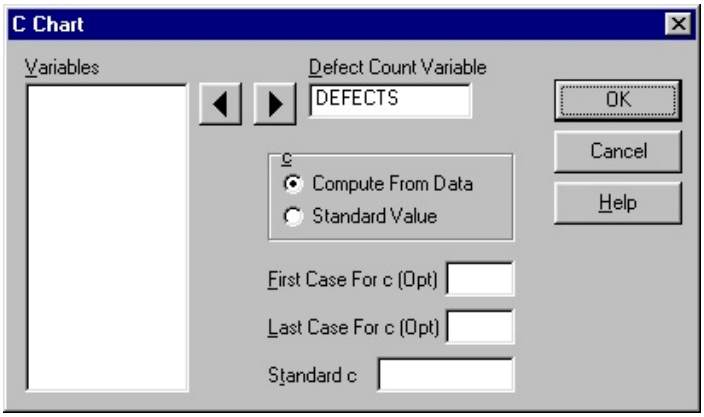


The number of defective cans is plotted for each sample. The center line is plotted at the historical value for np, 50 × 0.231 = 11.6. The 3-sigma upper and lower control limits are labeled on the right side at 20.5 and 2.6. Two values exceed the UCL at cases 15 and 23, which are noted at the bottom of the chart.

C Chart

The **C Chart** is the control chart for nonconformities (defects). A product may contain one or more defects and not be considered defective. If, for example, a television cabinet has a flaw in the finish, we wouldn't necessarily want to reject the television. In these situations we're more interested in the number of defects per inspection unit.

Specification



The c chart is computed from the number of defects per inspection unit. Enter the name of the variable that contains the counts of defects for each inspection unit.

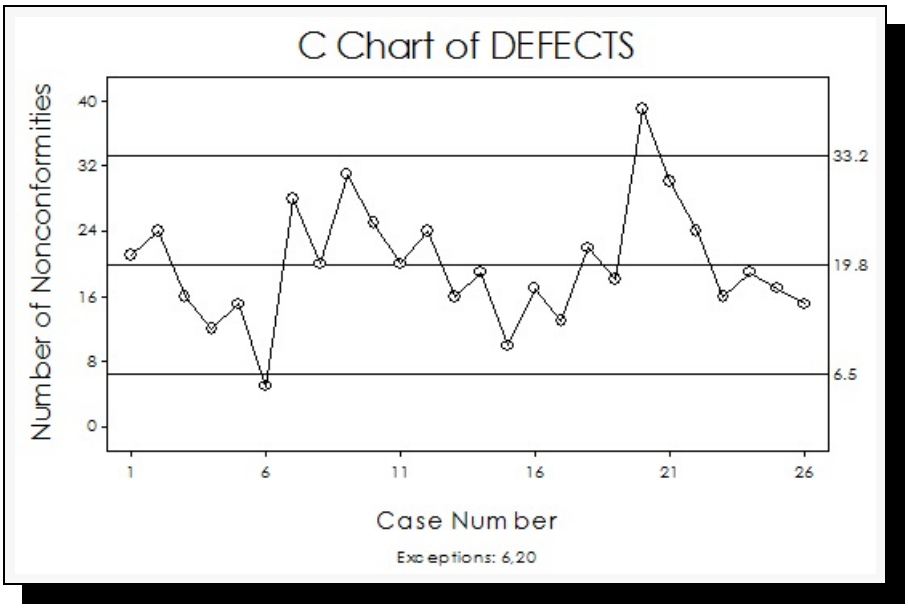
The center line and the control limits are computed from c, the number of nonconformities per inspection unit. You can enter a standard or historical value for c, or you can use the average value of c computed from the data. When c is estimated from the data, you can specify the first and last case number used to compute the average value. If these case numbers are left blank, as in the dialog box above, all cases are used. When you select ***Standard Value***, you must enter a value in the ***Standard c*** edit control.

Example

We use data from Montgomery (1991, p. 173) for our example. Printed circuit boards were inspected for defects. The inspection unit was defined as 100 circuit boards. The number of defects per 100 circuit boards for 26 samples are listed in the table on the next page. The data are stored in the file Sample Data\circuit.sx.

CASE	DEFECTS	CASE	DEFECTS
1	21	14	19
2	24	15	10
3	16	16	17
4	12	17	13
5	15	18	22
6	5	19	18
7	28	20	39
8	20	21	30
9	31	22	24
10	25	23	16
11	20	24	19
12	24	25	17
13	16	26	15

The resulting c chart for the variable DEFECTS is illustrated below.

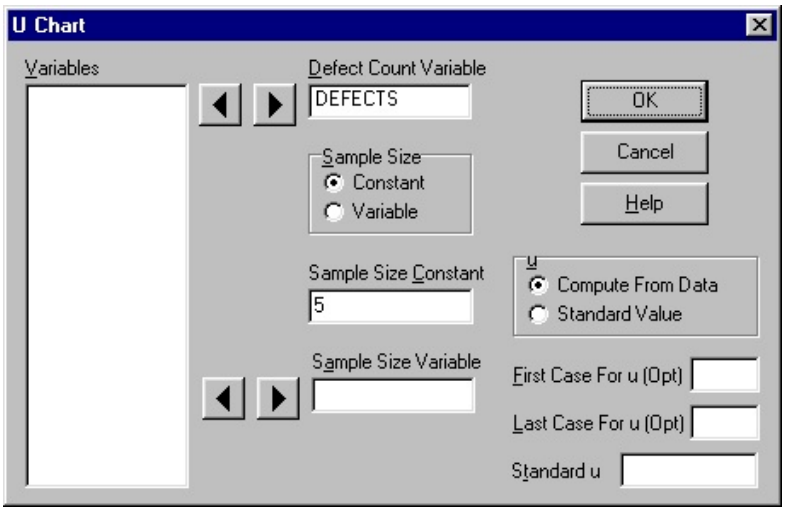


The number of nonconformities is plotted for each inspection unit (100 printed circuit boards). The center line is plotted at the average value for c 19.8. The 3-sigma upper and lower control limits are labeled on the right side at 33.2 and 6.5. The value at case 6 is below the LCL, and the value at case 20 is above the UCL. The process is not in control.

U Chart

The **U Chart** is the attribute control chart for nonconformities per unit. It is used in controlling the nonconformities per unit when the sample size is not one inspection unit. The c chart discussed earlier (page 339) is used when the sample size is one.

Specification



First move the name of the variable containing the number of defects for each sample to the ***Defect Count Variable*** box. If the sample size is constant, select the ***Constant*** sample size radio button and enter the sample size in the ***Sample Size Constant*** edit control. If the sample size is not always the same, select the ***Variable*** sample size radio button and move the name of a second variable that contains the sample size for each case to the ***Sample Size Variable*** box.

The center line and control limits are computed from u, the number of nonconformities per unit. You can enter a standard or historical value for u, or you can use the average value of u computed from the data. When u is estimated from the data, you can specify the first and last case number used to compute the average value. If these case numbers are left blank, as in the dialog box above, all cases are used. When you select ***Standard Value***, you must enter a value in the ***Standard u*** edit control.

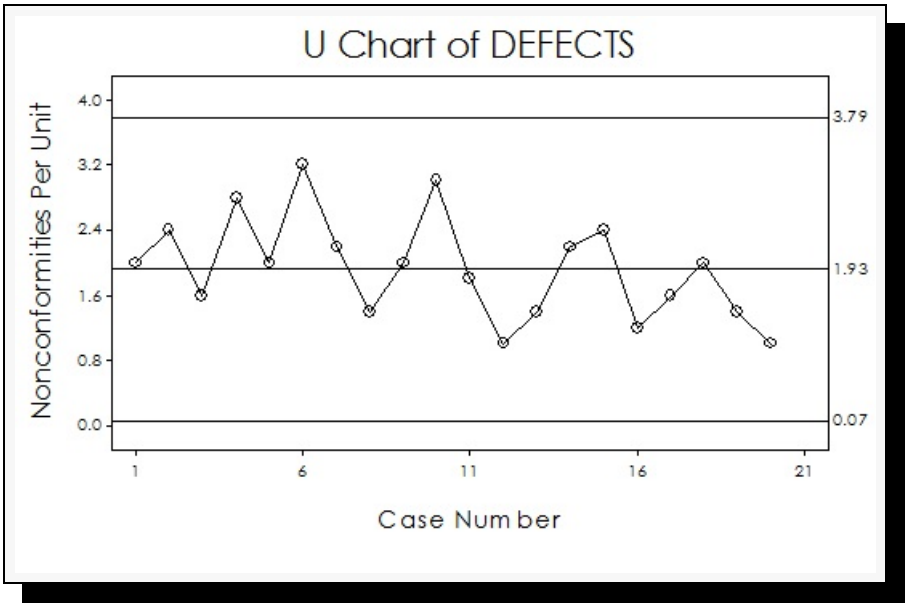
Example

We use data from Montgomery (1991, p. 181) for our example. Five personal computers were periodically sampled from the final assembly line and inspected for defects. The inspection unit was one computer. The sample size was five.

The total number of defects for each sample of five computers are listed below, and are stored in the file Sample Data\computers.sx.

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

The variable for the defect counts and the sample size are entered in the u chart dialog box on the preceding page. The results are presented in the figure below.



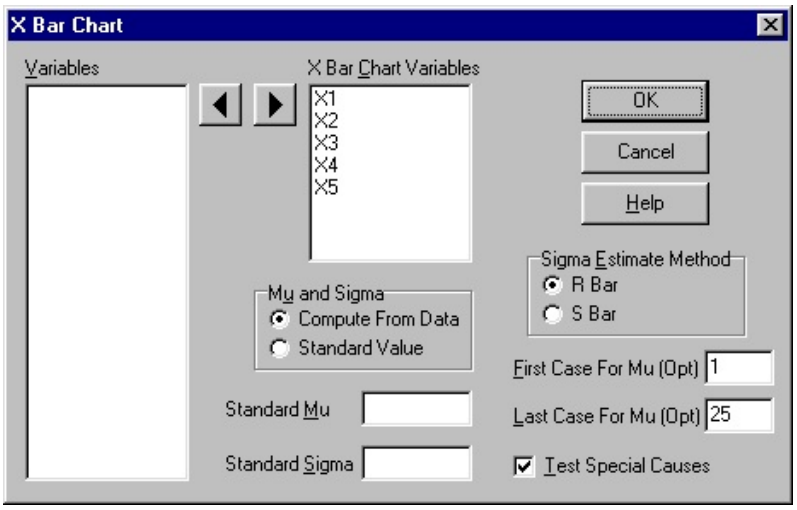
The process appears to be in control.

X Bar Chart

The **X Bar Chart**, which plots sample averages, is used to control the process average of a variable. A quantitative quality characteristic, such as the diameter of a piston ring, is called a variable. This chart is normally used in conjunction with either the R chart or the S chart to control the variability in the process.

Specification

The X bar chart requires that the quality characteristic is sampled with a sample size of at least two (use the I chart when the sample size is one). The data must be arranged in a *Statistix* data set so each case represents one sample and the individual measurements of a sample are recorded in separate variables (see the example data on the next page). If your data are presented in a single column, you can use the **Unstack** command to rearrange the data (see Chapter 2).



Move the names of the variables containing the individual measurements of the quality characteristic to the ***X Bar Chart Variables*** list box. You must select at least two variables, but no more than 20.

The center line and control limits are computed from the mean (mu) and standard deviation (sigma) of the process. You can choose to enter standard or historical values for mu and sigma, or you can have estimates computed from the data. If you select the ***Compute From Data*** radio button, you must also select a method for estimating sigma—the ***R-Bar*** or ***S-Ba***r method.

You should choose the method that corresponds to the control chart you're using to control the process variability, either the R chart or the S chart. You can also specify the ***First Case*** and ***Last Case*** to use to compute mu and sigma. If these case numbers are left blank, all cases are used.

If you select the ***Standard Value*** method, you must enter values in the ***Standard Mu*** and ***Standard Sigma*** edit controls.

Check the ***Test Special Causes*** check box to have eight tests for special causes performed. These tests are described on the next page.

Example

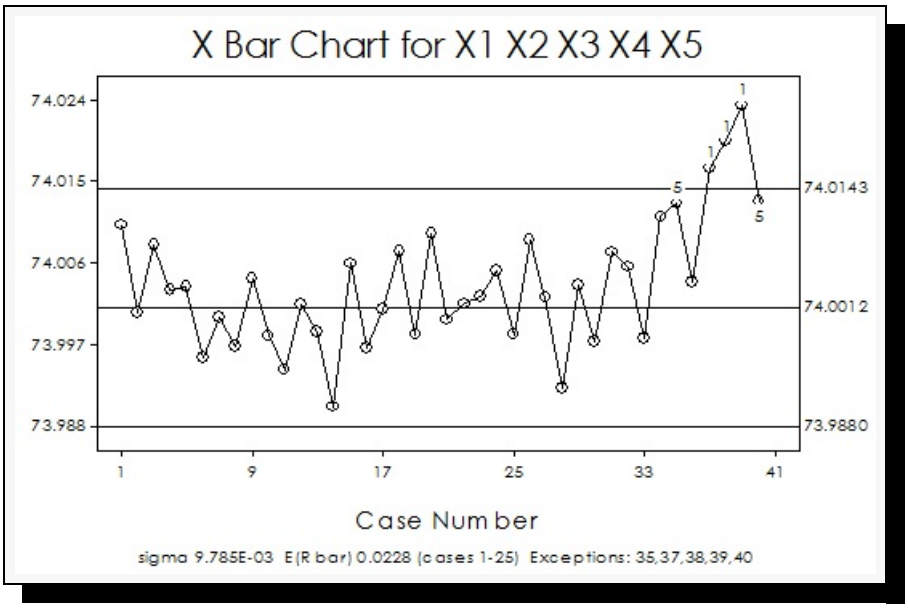
We use data from Montgomery (1991, p. 206) for our example. The data are listed below, and are stored in the file Sample Data\pistons.sx.

CASE	X1	X2	X3	X4	X5
1	74.030	74.002	74.019	73.992	74.008
2	73.995	73.992	74.001	74.011	74.004
3	73.988	74.024	74.021	74.005	74.002
4	74.002	73.996	73.993	74.015	74.009
5	73.992	74.007	74.015	73.989	74.014
6	74.009	73.994	73.997	73.985	73.993
7	73.995	74.006	73.994	74.000	74.005
8	73.985	74.003	73.993	74.015	73.988
9	74.008	73.995	74.009	74.005	74.004
10	73.998	74.000	73.990	74.007	73.995
11	73.994	73.998	73.994	73.995	73.990
12	74.004	74.000	74.007	74.000	73.996
13	73.983	74.002	73.998	73.997	74.012
14	74.006	73.967	73.994	74.000	73.984
15	74.012	74.014	73.998	73.999	74.007
16	74.000	73.984	74.005	73.998	73.996
17	73.994	74.012	73.986	74.005	74.007
18	74.006	74.010	74.018	74.003	74.000
19	73.984	74.002	74.003	74.005	73.997
20	74.000	74.010	74.013	74.020	74.003
21	73.988	74.001	74.009	74.005	73.996
22	74.004	73.999	73.990	74.006	74.009
23	74.010	73.989	73.990	74.009	74.014
24	74.015	74.008	73.993	74.000	74.010
25	73.982	73.984	73.995	74.017	74.013
26	74.012	74.015	74.030	73.986	74.000
27	73.995	74.010	73.990	74.015	74.001
28	73.987	73.999	73.985	74.000	73.990
29	74.008	74.010	74.003	73.991	74.006
30	74.003	74.000	74.001	73.986	73.997
31	73.994	74.003	74.015	74.020	74.004
32	74.008	74.002	74.018	73.995	74.005
33	74.001	74.004	73.990	73.996	73.998
34	74.015	74.000	74.016	74.025	74.000
35	74.030	74.005	74.000	74.016	74.012
36	74.001	73.990	73.995	74.010	74.024
37	74.015	74.020	74.024	74.005	74.019
38	74.035	74.010	74.012	74.015	74.026
39	74.017	74.013	74.036	74.025	74.026
40	74.010	74.005	74.029	74.000	74.020

Automotive piston rings were sampled from a forge. Five rings were taken

per sample. The inside diameter of the piston ring—listed on the preceding page—was recorded.

The parameters for the X bar chart are specified in the dialog box on page 343. Note that cases 1 through 25 have been entered as the cases (or samples) to use to estimate mu and sigma. The results are presented below.



The sample averages are plotted for each case. The values for the center line and the 3-sigma control limits are labeled on the right side. The estimates for sigma and expected R bar computed from the first 25 cases are given in the footnote.

Note the upward drift in the process mean, starting around case 34. The points at cases 37, 38, and 39 exceed the UCL. The points at cases 36 and 40 are marked with a 5. This means that these points failed test number 5: two out of three consecutive points in zone A (beyond the 2-sigma limits).

Tests for Special Causes

A "special cause", or "assignable cause", is signaled on control charts when a point is plotted outside the 3-sigma control limits. In addition to a point outside the UCL and LCL, there are other tests based on patterns of more than one point that can be applied to the X bar and I charts. These tests are summarized by Nelson (1984).

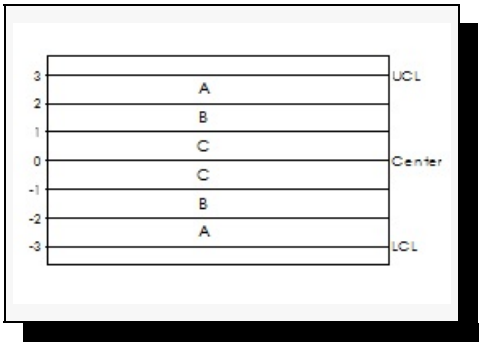
When these tests are selected, *Statistix* will indicate an out of control point using a digit indicating the test that caused the signal. If a point fails more than one test, the lowest digit is used to mark the point.

Test #1: A point outside the 3-sigma control limits.

Test #2: Nine points in a row on one side of the center line.

Test #3: Six points in a row, either all increasing or all decreasing.

Test #4: Fourteen points in a row, alternating up and down.



Tests #5 through #8 refer to zones A, B, and C. Zone A is the area of the chart between the 2- and 3-sigma lines. Zone B is the area between the 1- and 2-sigma lines. Zone C is the area between the center line and the 1-sigma line.

Test #5: Two out of three points in a row in zone A or beyond on one side of the center line.

Test #6: Four out of five points in a row in zone B or beyond on one side of the center line.

Test #7: Fifteen points in a row in zone C on either side on the center line.

Test #8: Eight points in a row on either side of the center line, but none of them in zone C.

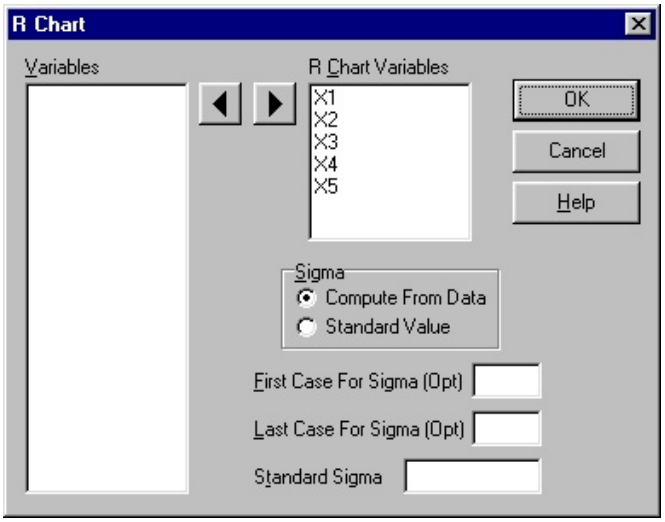
By using a number of these tests at once, you increase the sensitivity of the control chart. But you also increase the chance of a false alarm.

R Chart

The **R Chart**, which plots sample ranges, is used to control the process variability of a variable. A quantitative quality characteristic, such as the diameter of a piston ring, is called a variable. The S chart is an alternative control chart for process variability.

Specification

The R chart requires that the quality characteristic is sampled with a sample size of at least two (use the MR chart when the sample size is one). The data must be arranged in a *Statistix* data set such that each case represents one sample and the individual measurements of a sample are recorded in separate variables. If your data are presented in a single column, you can use the **Unstack** command to rearrange the data (see Chapter 2).



Select the names of the variables containing the individual measurements of the quality characteristic and move them to the ***R Chart Variables*** list box. You must select at least two variables, but no more than 20.

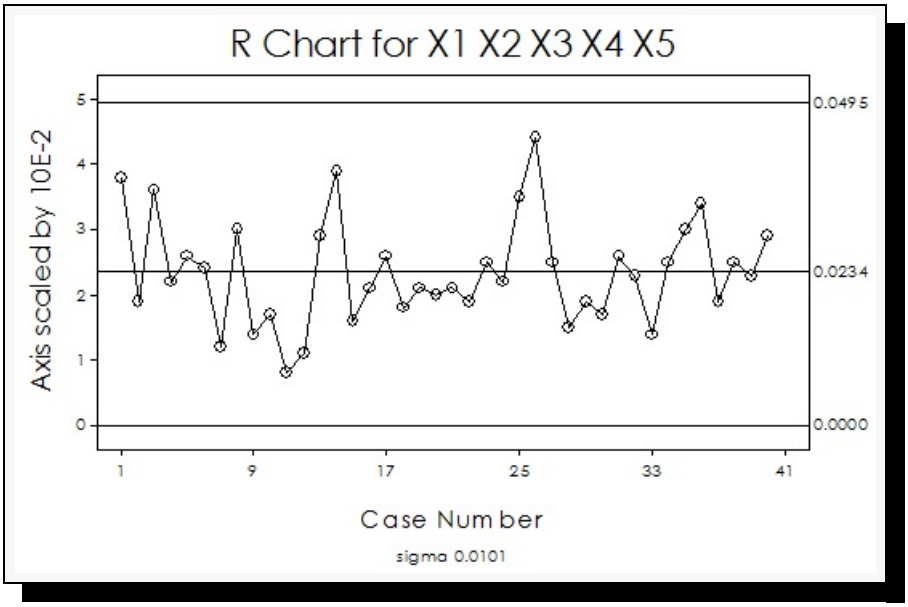
The center line and control limits are computed from the standard deviation (sigma) of the process. You can choose to enter a standard or historical value for sigma, or you can have an estimate computed from the data. If you select the ***Compute From Data*** radio button, you can specify the ***First Case*** and ***Last Case*** to use to compute sigma. If these case numbers are left blank, all cases are used.

If you select the ***Standard Value*** method, you must enter a value in the ***Standard Sigma*** edit control.

Example

We use example data from Montgomery (1991, p. 206) for our example. Automotive piston rings were sampled from a forge. Five rings were selected per sample. The inside diameter of each piston ring was measured. This data set is also used as the X Bar Chart example and is listed on page 344. The data are available from the file Sample Data\pistons.sx.

The R chart model is specified in the dialog box on the preceding page. The results are presented below.



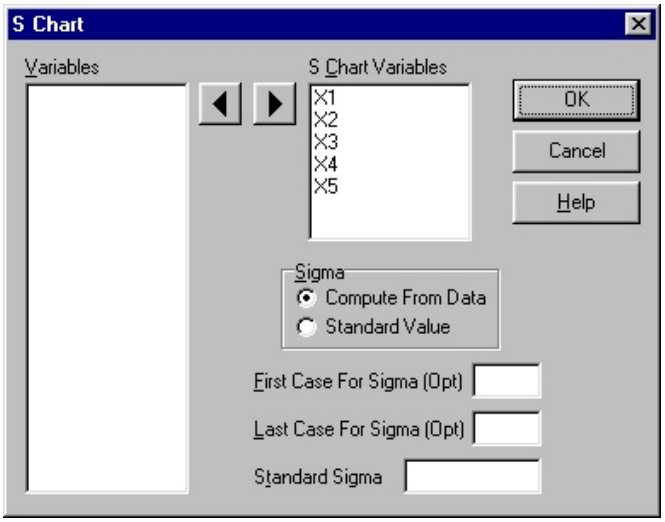
The sample ranges are plotted for each case. The values for the center line and the 3-sigma control limits are labeled on the right side. The value used for sigma, in this case estimated from the data, is given at the bottom of the R chart.

S Chart

The **S Chart**, which plots sample standard deviations, is used to control the process variability of a variable. A quantitative quality characteristic, such as the diameter of a piston ring, is called a variable. The R chart discussed earlier is another control chart for process variability.

Specification

The S chart requires that the quality characteristic is sampled with a sample size of at least two (use the MR chart when the sample size is one). The data must be arranged in a *Statistix* data set so each case represents one sample and the individual measurements of a sample are recorded in separate variables. If your data are presented in a single column, you can use the **Unstack** command to rearrange the data (see Chapter 2).



Select the names of the variables containing the individual measurements of the quality characteristic and move them to the ***S Chart Variables*** list box. You must list at least two variables, but no more than 20.

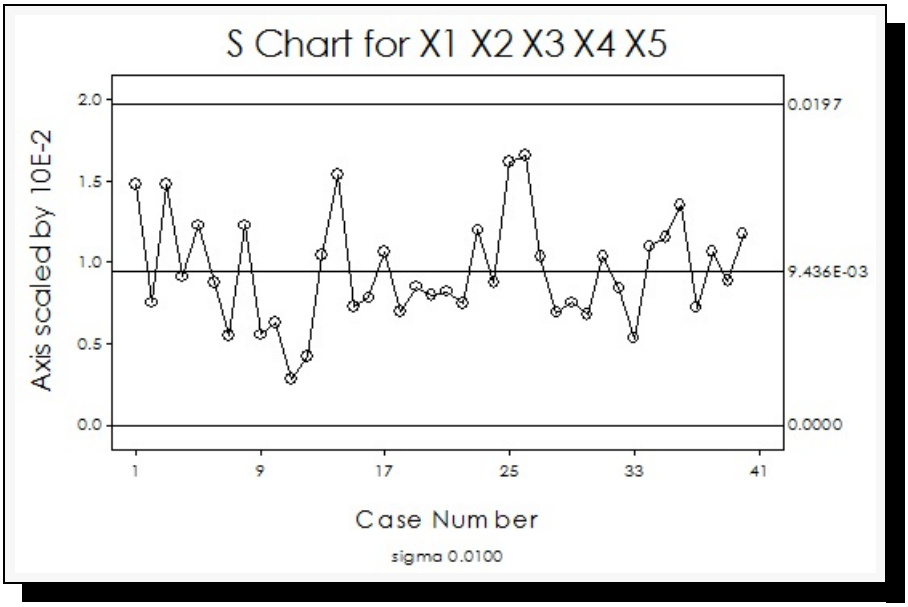
The center line and control limits are computed from the standard deviation (sigma) of the process. You can choose to enter a standard or historical value for sigma, or you can have an estimate computed from the data. If you select the ***Compute From Data*** radio button, you can specify the ***First Case*** and ***Last Case*** to use to compute sigma. If these case numbers are left blank, all cases are used.

If you select the ***Standard Value*** method, you must enter a value in the ***Standard Sigma*** edit control.

Example

We use data from Montgomery (1991, p. 206) for our example. Automotive piston rings were sampled from a forge, five rings per sample. The inside diameter of the piston ring was the quality characteristic of interest. This data set was also used as the X Bar Chart example and is listed on page 344. The data are available from the file Sample Data\pistons.sx.

The S chart model is specified in the dialog box on the preceding page. The results are displayed below.

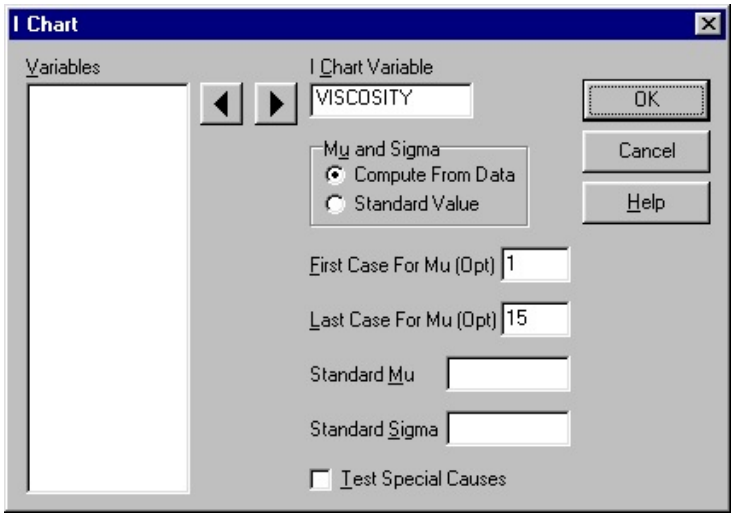


The sample standard deviations are plotted for each sample. The values for the center line and the 3-sigma control limits are labeled on the right side. The value used for sigma, in this case estimated from the data, is given at the bottom of the S chart.

I Chart

The **I Chart** is the control chart for individuals used to control the process mean. This chart is sometimes called the X chart. The process variability is estimated from the moving range, which is the difference between two successive observations. The MR chart is the companion chart used to control the process variability.

Specification



Select the name of the variable that contains the individual observations and move it to the ***I Chart Variable*** box.

The center line and control limits are computed from the mean (mu) and standard deviation (sigma) of the process. You can choose to enter standard or historical values for mu and sigma, or you can have estimates computed from the data. If you select the ***Compute From Data*** radio button, you can specify the ***First Case*** and ***Last Case*** to use to compute mu and sigma. If these case numbers are left blank, all cases are used.

If you select the ***Standard Value*** method, you must enter values in the ***Standard Mu*** and ***Standard Sigma*** edit controls.

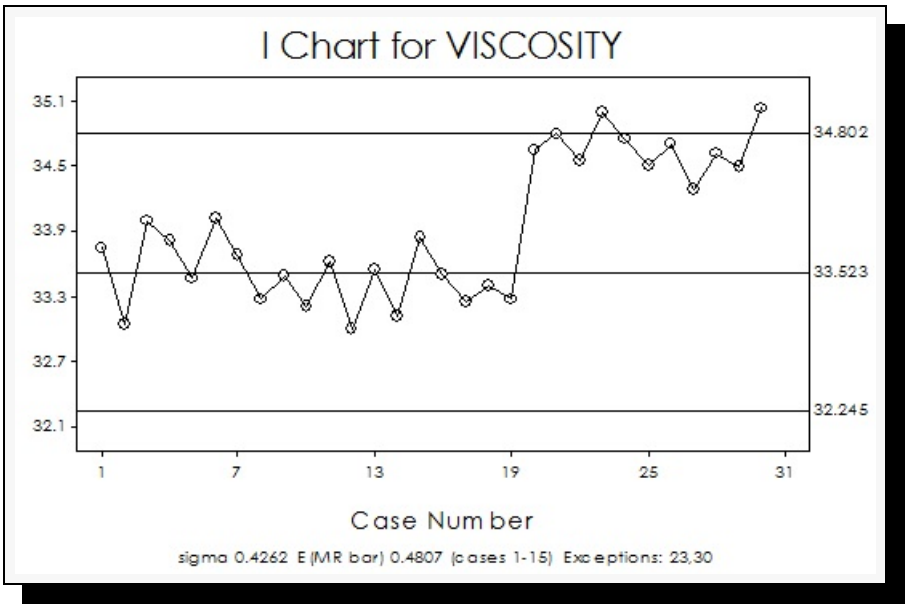
Check the ***Test Special Causes*** check box to have eight tests for special causes performed. These tests are described on page 345.

Example

We use data from Montgomery (1991, p. 242) for our example. The viscosity of batches of aircraft paint primer was measured. Because it takes several hours to make one batch of primer, it wasn't practical to accumulate samples of more than one batch. The data from 30 batches of primer are listed below. The data are available from the file Sample Data\paint.sx.

CASE	VISCOSITY	CASE	VISCOSITY
1	33.75	16	33.50
2	33.05	17	33.25
3	34.00	18	33.40
4	33.81	19	33.27
5	33.46	20	34.65
6	34.02	21	34.80
7	33.68	22	34.55
8	33.27	23	35.00
9	33.49	24	34.75
10	33.20	25	34.50
11	33.62	26	34.70
12	33.00	27	34.29
13	33.54	28	34.61
14	33.12	29	34.49
15	33.84	30	35.03

The dialog box on the preceding page specifies the I chart model. Cases 1-15 are used to estimate the process mean and standard deviation. The results are as follows:



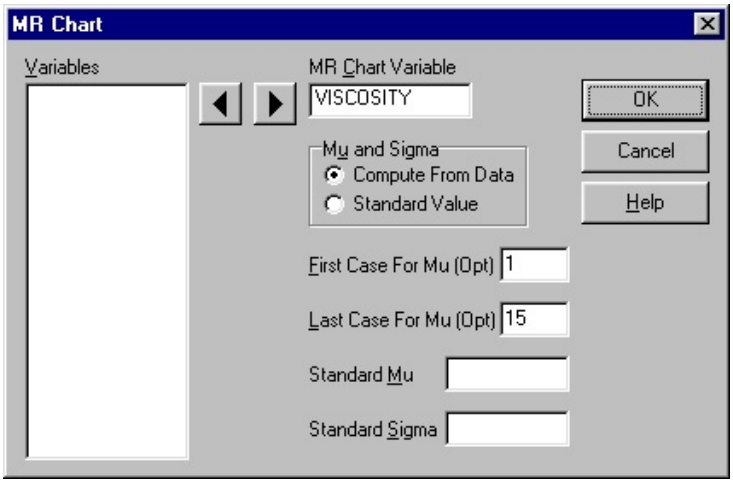
The individual observed values of primer viscosity are plotted for each case.

Note the shift in the process mean at case 20. Two points are plotted above the UCL. This process is clearly out of control.

MR Chart

The **MR Chart** is the control chart for individuals used to control process variability. The process variability is estimated from the moving range, the difference between two successive observations: MR = $|x_i - x_{i-1}|$.

Specification



Select the name of the variable that contains the individual observations and move it to the ***MR Chart Variable*** box.

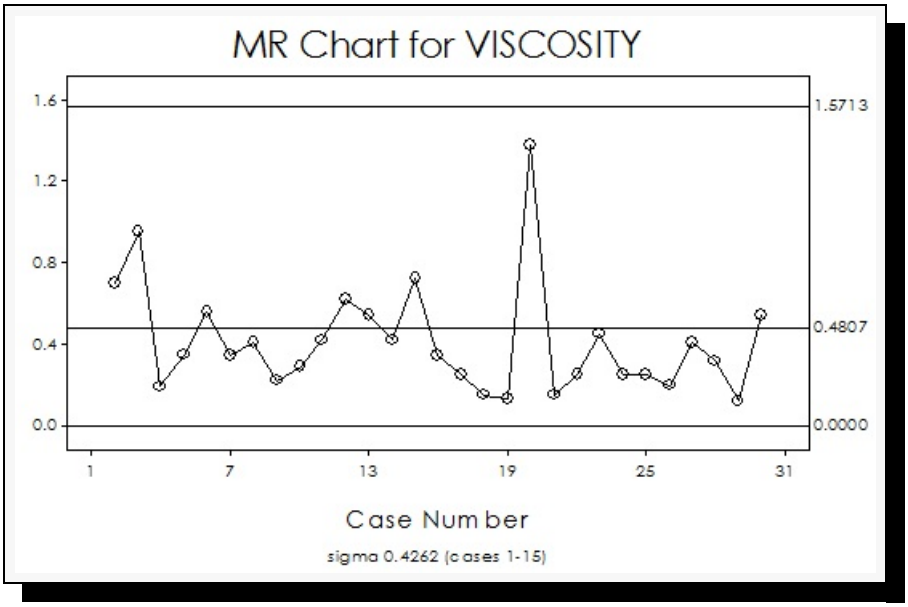
The center line and control limits are computed from the standard deviation (sigma) of the process. You can choose to enter a standard or historical value for sigma, or you can have an estimate computed from the data. If you select the ***Compute From Data*** radio button, you can specify the ***First Case*** and ***Last Case*** to use to compute sigma. If these case numbers are left blank, all cases are used.

If you select the ***Standard Value*** method, you must enter a value in the ***Standard Mu*** edit control.

Example

We use data from Montgomery (1991, p. 242) for our example. The viscosity of batches of aircraft paint primer was measured. Because it takes several hours to make one batch of primer, it wasn't practical to accumulate samples of more than one batch. The example uses 30 batches of primer. This data set was also used for the I Chart example and is listed on page 352. The data are available from the file Sample Data\paint.sx.

The dialog box on the preceding page specifies the MR chart model. Cases 1-15 are used to estimate the process standard deviation. The results are presented below.



The moving ranges are plotted for each case, starting with the second case. The center line and 3-sigma UCL are labeled on the right side. The estimates for sigma and expected MR bar, based on the first 15 cases, are given in the footnote.

The spike at case 20 signals the shift in the process mean observed in the example I chart shown on page 352.

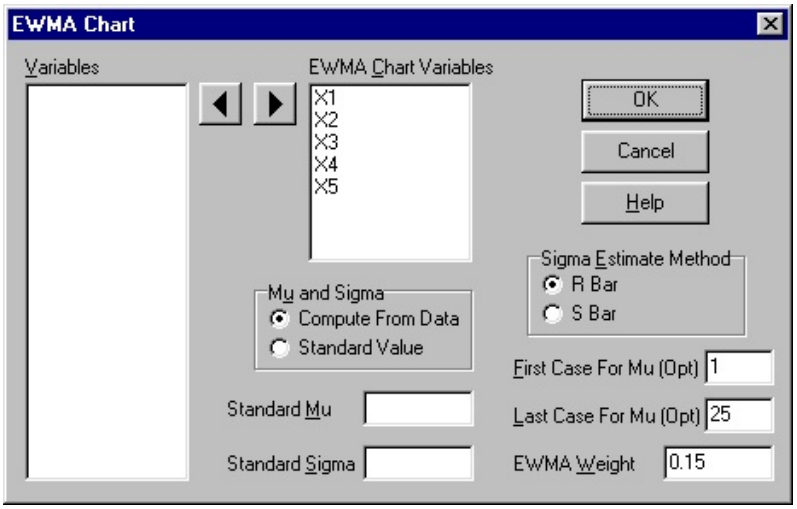
Montgomery (1991) warns that caution should be used when examining patterns in MR charts. Because the moving ranges are correlated, runs and patterns arise naturally in the charts.

EWMA Chart

The **EWMA Chart** is a control chart for variables used to control the process mean. The value plotted on the chart is an exponentially weighted moving average incorporating data from all previous samples as well as the sample mean itself. This chart reacts faster than the X bar chart to small shifts in the process mean but reacts slower to large shifts.

Specification

The EWMA chart requires that the data be arranged so that each case represents one sample and the individual measurements of a sample are recorded in separate variables. If your sample size is greater than one but your data are presented in a single column, use the **Unstack** command to rearrange the data (see Chapter 2).



Move the names of the variables containing the individual measurements of the quality characteristic to the ***EWMA Chart Variables*** list box. The number of variables selected equals the sample size, which must be constant.

The center line and control limits are computed from the mean (mu) and standard deviation (sigma) of the process. You can choose to enter standard or historical values for mu and sigma, or you can have estimates computed from the data. If you select the ***Compute From Data*** radio button, you must also select a method for estimating sigma—the ***R-Bar*** or ***S-Ba***r method. You can also specify the ***First Case*** and ***Last Case*** to use to compute mu

and sigma. If these case numbers are left blank, all cases are used.

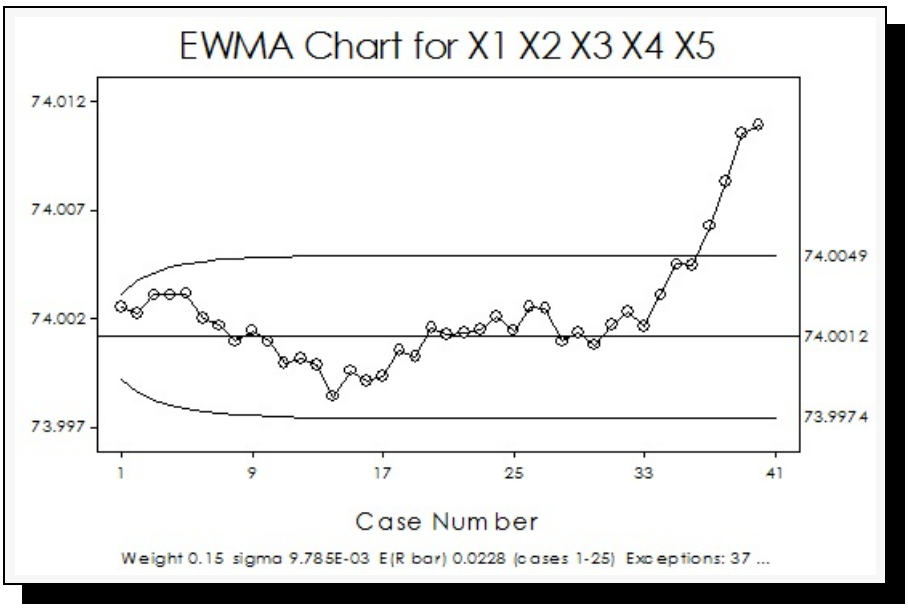
If you select the ***Standard Value*** method, you must enter values in the ***Standard Mu*** and ***Standard Sigma*** edit controls.

The EWMA chart requires a ***EWMA Weight*** that determines the extent to which past observations influence the exponentially weighted moving average plotted on the chart. A small weight results in a slow response to new values; a large weight results in a fast response to new values. Values used for the weight are normally in the range 0.05 - 0.25.

Example

We use data from Montgomery (1991, p. 206) for our example. Automotive piston rings were sampled from a forge, five rings per sample. The inside diameter of the piston ring was recorded. These data were also used for the X Bar Chart example and are listed on page 344. The data are available from the file Sample Data\pistons.sx.

The EWMA chart is specified in the dialog box on the preceding page. The results are as follows:

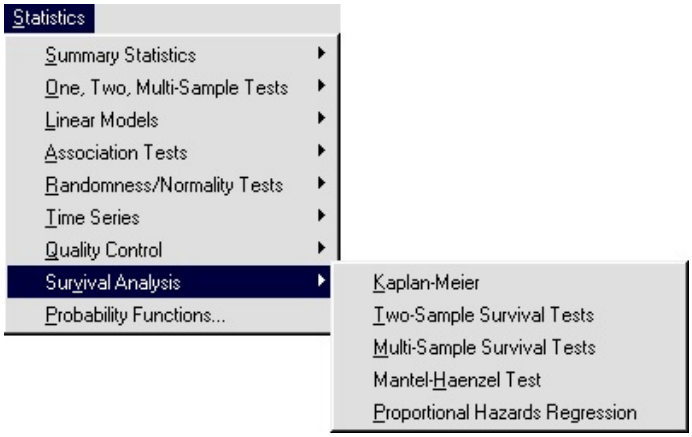


The process mean and 3-sigma control limits are reported on the right, and the process standard deviation (sigma) is reported at the bottom.

C H A P T E R

12

Survival Analysis



The statistical procedures in the chapter are used to analyze survival time data. Survival time is defined as the time to the occurrence of a specific event, which may be the development of a disease, response to a treatment, relapse, or death. Survival analysis has been extended to fields beyond biomedical studies to include electrical engineering, sociology, and marketing. For example of survival time in sociology might be the duration of first marriage.

A common complication of survival data are censored observations. A censored observation is one where the given event of interest was not recorded, either because the subject was lost to the study, or because the

study ended before the event occurred. The emphasis of the procedures in the chapter are those that can handle censored observations.

The distribution of survival times are described using the survivorship function (or survivor function) and the hazard function. The survivorship function S(t) is defined as the probability that an individual survives longer than t. The hazard function h(t) gives the conditional failure rate. It's the probability of failure during a small time interval assuming that the individual has survived to the beginning of the interval.

The **Kaplan-Meier** procedure computes the Kaplan-Meier product limit estimates of the survival functions. This method of estimating the survival functions can handle censored data and does not require any assumptions about the form of the survival function. It is appropriate for small and large data sets. The survivorship and hazard functions can be plotted.

The **Two-Sample Survival Tests** procedure computes five nonparametric tests for comparing two survival distributions: Gehan-Wilcoxon Test, Cox-Mantel Test, Logrank Test, Peto-Wilcoxon Test, and Cox's F Test. These tests are based on the ranks of the survival times and work for censored or uncensored observations.

The **Multi-Sample Survival Tests** procedure computes three nonparametric tests for comparing three or more survival distributions: Gehan-Wilcoxon Test, Logrank Test, and the Peto-Wilcoxon Test. These tests are extensions of the Kruskal-Wallis test discussed in Chapter 5 and the two-sample survival tests. These tests can be used to compare survival times for censored data.

The **Mantel-Haenzel Test** is used to compare survival experience between two groups when adjustments for other prognostic factors are needed. It's often used in clinical and epidemiologic studies as a method of controlling the effects of confounding variables.

The **Proportional Hazards Regression** procedure computes Cox's proportional hazards regression for survival data. It can be used to establish the statistical relationship between survival time and independent variables or covariates measured on the subjects. The reports include a regression coefficient table, likelihood test for the overall model, and a variance-covariance matrix of the coefficients. The regression coefficients can be used to compute relative risk, a measure of the effect of a factor on a subject's survival time.

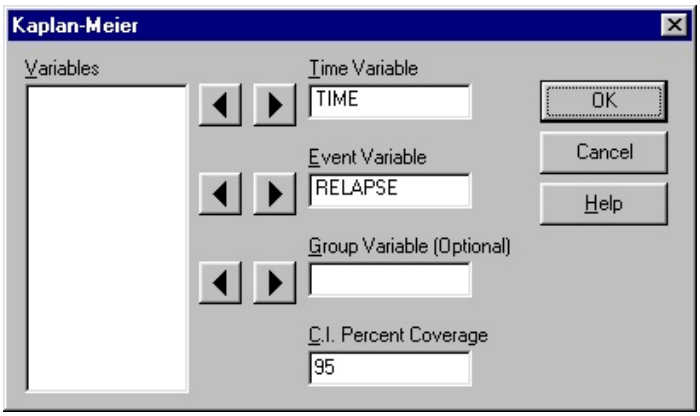
The procedures in this chapter concentrate on survival time analyses that can handle censored data. Many other procedures in *Statistix* are useful for analyzing uncensored survival time data. These include the **Wilcoxon Rank Sum Test** and **Kruskal-Wallis AOV** in Chapter 5, **Logistic Regression** in Chapter 6, and **Two By Two Tables** and **Spearman Rank Correlations** in Chapter 8.

Kaplan-Meier

This procedure computes the Kaplan-Meier product limit estimates of the survival functions. This method of estimating the survival functions can handle censored data and does not require any assumptions about the form of the survival function. It is appropriate for small and large data sets.

The procedure produces a survival function table, a percentile with confidence limits reports, survivorship function plot, and hazard function plot.

Specification



Select the name of the variable containing the survival times and move it to the ***Time Variable*** box. Select the variable used to indicate whether or not the event of interest (e.g., death) occurred or not and move it to the ***Event Variable*** box. The event variable must be coded 0 if the event did not occur (i.e., censored observation) and 1 if the event did occur (i.e., uncensored observation).

You can specify a grouping variable using the ***Group Variable*** box. The values of the group variable are used to divide the survival times into groups of interest (e.g., treatment). If you specify a group variable, survival functions are computed separately for each group.

The product-limit table displays confidence intervals for the survivorship function. You can change the ***C.I. Percent Coverage*** for the confidence intervals.

Data Restrictions

The event variable must be an integer or real variable and may only contain the values 0 and 1. The group variable, if used, may be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters.

Example

The example data are invented remission durations for 10 patients with solid tumors used for illustration by Lee (1992, p. 71). Six patients relapse at 3.0, 6.5, 6.5, 10, 12, and 15 months; 1 patient is lost to follow-up at 8.4 months; and 3 patients are still in remission at the end of the study after 4.0, 5.7, and 10 months. The remission times and relapse event indicators are stored in the variables TIME and RELAPSE.

CASE	TIME	RELAPSE
1	3.0	1
2	4.0	0
3	5.7	0
4	6.5	1
5	6.5	1
6	8.4	0
7	10.0	1
8	10.0	0
9	12.0	1
10	15.0	1

The analysis is specified using the dialog box of the preceding page. The first report displayed after pressing the OK button is the product-limit survival function table displayed below.

Kaplan-Meier Product-Limit Survival Distribution
Time Variable: TIME
Event Variable: RELAPSE

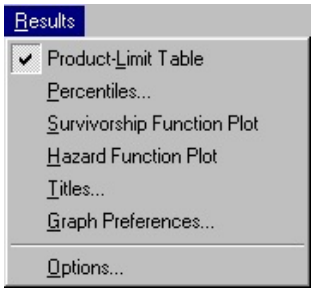
Time	Died	Cen-sored	At Risk	Lower 95% C.I.	S(t)	Upper 95% C.I.	SE S(t)	H(t)
3.0	1	0	10	0.6137	0.9000	0.9808	0.0949	0.1054
4.0	0	1	9					
5.7	0	1	8					
6.5	2	0	7	0.3581	0.6429	0.8531	0.1679	0.4418
8.4	0	1	5					
10.0	1	1	4	0.2066	0.4821	0.7690	0.1877	0.7295
12.0	1	0	2	0.0688	0.2411	0.5772	0.1946	1.4227
15.0	1	0	1	0.0000	0.0000	0.0000	0.0000	M

The table has one row for each distinct survival time. The column labeled "DIED" gives the number of subjects that had the event of interest recorded at the survival time for the row. In this example, it's the number of patients that relapsed. The column labeled "CENSORED" gives the number of censored observations at the survival time for the row. The "AT RISK" column gives the number of subjects still in the study before the survival

time for the row. The values for the survivorship function are given in the column labeled "S(t)" for each survival time that a death (relapse in this example) occurred. The lower and upper confidence limits, and the standard error of the survivorship function are also given. The last column gives the value of the hazard function (H(t)) for each survival time.

Kaplan-Meier Results Menu

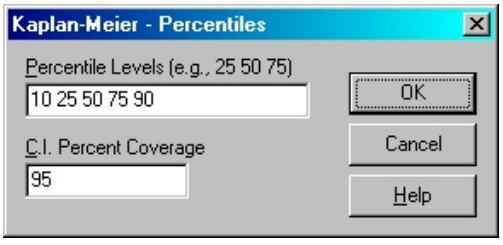
Once you've specified and computed a Kaplan-Meier analysis, a Results menu appears on the menu at the top of the *Statistix* window. Select the Results menu to access the pull-down menu displayed below.



Select Product-Limit Table from the menu to redisplay the results presented on the preceding page. Select Options from the menu to return to the Kaplan-Meier dialog box used to generate these results. The remaining options are discussed below.

Percentiles

One of the statistics that we want to find when analyzing survival analysis data is the median survival time. The Percentile report gives the median survival time with confidence limits, plus the values for other percentiles that may interest you. Select Percentiles from the Results menu and a dialog box like the one shown below appears.



List one or more percentile values in the ***Percentiles Levels*** box. You can change the ***C.I. Percent Coverage*** value. Press ***OK*** to display the report.

The report for the example tumor remission data is given below.

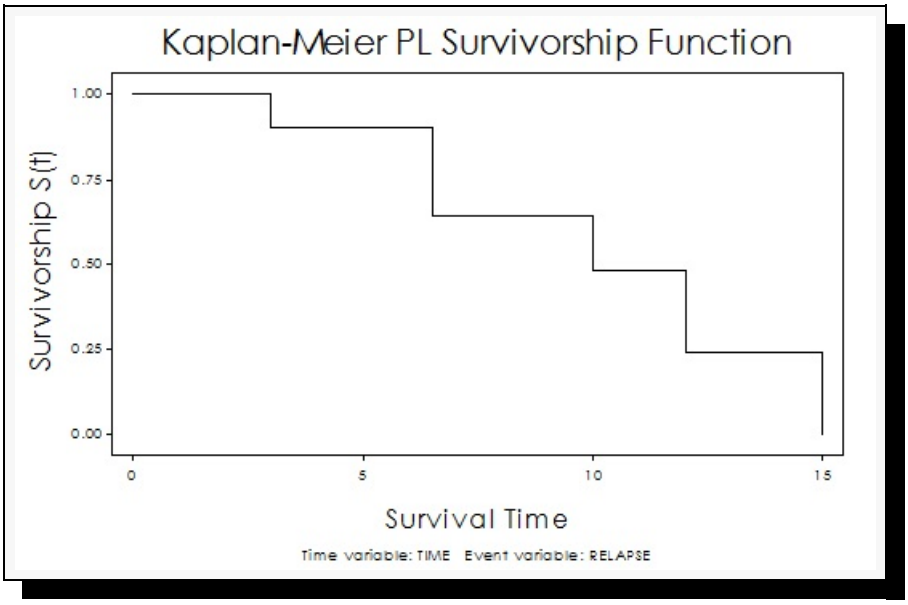
Kaplan-Meier Survivorship Percentiles
Time Variable: TIME
Event Variable: RELAPSE

Percentile	Lower 95% C.I.	Time	Upper 95% C.I.
90	3.000	5.700	6.500
75	3.000	6.500	12.000
50	6.500	9.822	M
25	10.000	11.926	M
10	12.000	13.756	M

The median remission duration for the example data is 9.822 with a 95% lower limit of 6.500. The upper limit can't be computed because of the insufficient number of uncensored observations, so an M is displayed.

Survivorship Function Plot

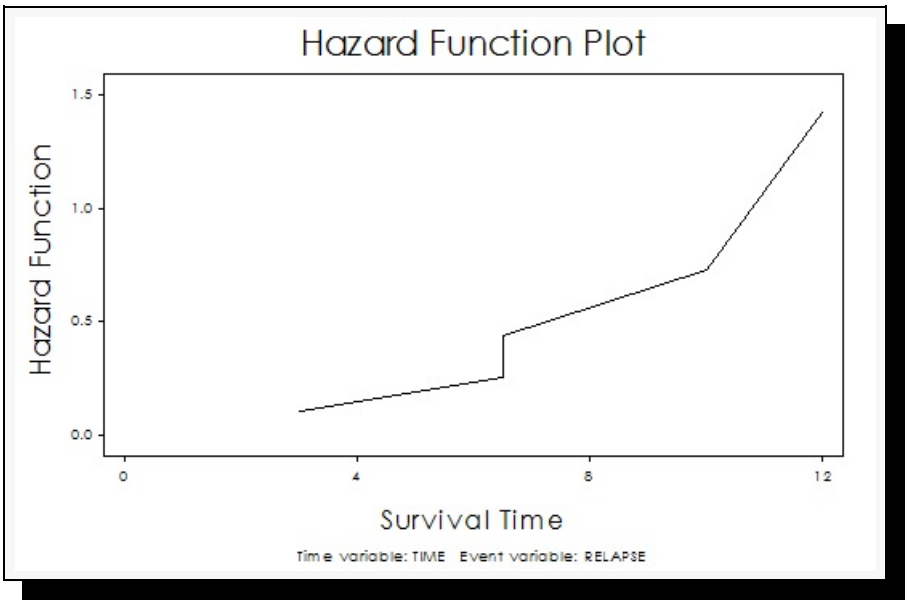
Select Survivorship Function Plot from the Results menu to plot the survivorship function. If a group variable was specified on the Kaplan-Meier dialog box, a separate line is plotted for each group. The survivorship function for the example remission data is displayed below.



Hazard Function Plot

Select Hazard Function Plot from the Results menu to plot the hazard function. If a group variable was specified on the Kaplan-Meier dialog box,

a separate line is plotted for each group. The hazard function for the example remission data is displayed below.



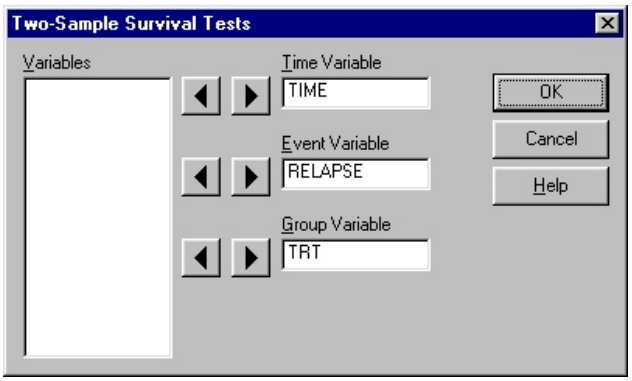
Computational Notes

Computations follow Lee (1992). The confidence intervals are computed using the technique described by Simon and Lee (1982).

Two-Sample Survival Tests

This procedure computes five nonparametric tests for comparing two survival distributions: Gehan-Wilcoxon Test, Cox-Mantel Test, Logrank Test, Peto-Wilcoxon Test, and Cox's F Test. These tests are based on the ranks of the survival times and work for censored or uncensored observations.

Specification



Select the variable that contains the survival times and move it to the ***Time Variable*** box. Select the variable used to indicate whether or not the event of interest (e.g., death) occurred or not and move it to the ***Event Variable*** box. Move the variable used to identify the two groups to the ***Group Variable*** box.

Data Restrictions

The event variable must be an integer or real variable and may only contain the values 0 and 1. The group variable may be of any data type but must have exactly two values. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. Cox's F-Test is only computed for complete of singly censored data.

Example

The example data are from Lee (1992, p. 107). Ten female patients with breast cancer are randomized to receive either CMF (cyclic administration of cyclophosphamide, methatrexate, and fluorouracil) or no treatment after a radical mastectomy. The remission times recorded at the end of two years are given on the next page.

CASE	TIME	RELAPSE	TRT
1	23	1	CMF
2	16	0	CMF
3	18	0	CMF
4	20	0	CMF
5	24	0	CMF
6	15	1	Control
7	18	1	Control
8	19	1	Control
9	19	1	Control
10	20	1	Control

The analysis is specified in the dialog box on the preceding page. The results are displayed below.

```
Two-Sample Survival Tests
Time Variable:  TIME
Event Variable: RELAPSE
Group Variable: TRT      (CMF, Control)

Gehan-Wilcoxon Test           Cox-Mantel Test
  W          18.000             U          2.7500
  Var(W)     57.778             I          1.0875
  Z            2.37             C            2.64
  P          0.0179             P          0.0084

Logrank Test                  Peto-Wilcoxon Test
  S          2.7500             S          2.1313
  Var(S)     1.2106             Var(S)     0.7651
  L            2.50             Z            2.44
  P          0.0124             P          0.0148
```

All four tests provide strong evidence that the two treatments are different. The positive sign of the test statistics (Z, C, L, and Z respectively) indicate that the first treatment, CMF, is more effective than the second.

The Cox's F-Test can only be used for singly censored or complete samples. The results above to not include Cox's F-Test because the data are progressively censored. Consider a second example containing singly censored data. In an experiment comparing two treatments for solid tumor, six mice are assigned to treatment A and six to treatment B (Lee, 1992, p. 115). The experiment is terminated after 30 days. The following survival times are recorded (+ indicates censored observations).

Treatment A: 8, 8, 10, 12, 12, 13
Treatment B: 9, 12, 15, 20, 30+, 30+

All of the mice die except for two mice that were still alive at the end of the study. The analysis is specified in the same manner as before. The results are displayed on the next page.

```
Two-sample Survival Tests
Time Variable:  TIME
Event Variable: DIED
Group Variable: TRT   (A, B)

Gehan-Wilcoxon Test           Cox-Mantel Test
  W          -24.000            U          -2.8306
  VAR(W)      152.73            I           1.6037
  Z            -1.94            C            -2.24
  P           0.0521            P           0.0254

Logrank Test                  Peto-Wilcoxon Test
  S          -2.8306            S          -2.1667
  VAR(S)      2.0583            VAR(S)      1.0795
  L            -1.97            Z            -2.09
  P           0.0485            P           0.0370

Cox's F Test
  F             0.25
  DF           12, 8
  P           0.0305
```

The results for the five different tests are similar. Lee (1992) discusses under what circumstances one test may be more powerful than another.

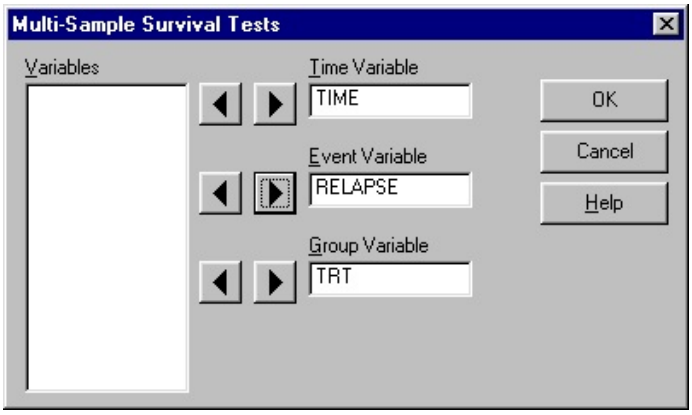
Computational Notes

The computations for these tests can be found in Lee (1992).

Multi-Sample Survival Tests

This procedure computes three nonparametric tests for comparing three or more survival distributions: Gehan-Wilcoxon Test, Logrank Test, and the Peto-Wilcoxon Test. These tests are extensions of the Kruskal-Wallis test discussed in Chapter 5 and the two-sample survival tests discussed in the preceding section. These tests can be used to compare survival times for censored data.

Specification



Select the variable that contains the survival times and move it to the ***Time Variable*** box. Select the variable used to indicate whether or not the event of interest (e.g., death) occurred or not and move it to the ***Event Variable*** box. Move the variable used to identify the various groups to the ***Group Variable*** box.

Data Restrictions

The event variable must be an integer or real variable and may only contain the values 0 and 1. The group variable may be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters.

Example

The example data are from Lee (1992, p. 127). Three different treatments are given to leukemia patients. We're interested in determining whether differences in remission times exist between the three groups. The remission times are given in the table on the next page, and are stored in the file Sample Data\leukemia.sx. The variable RELAPSE indicates whether

each patient relapsed or not.

CASE	TIME	RELAPSE	TRT	CASE	TIME	RELAPSE	TRT
1	4	1	1	34	75	1	2
2	5	1	1	35	99	1	2
3	9	1	1	36	103	1	2
4	10	1	1	37	162	1	2
5	12	1	1	38	169	1	2
6	13	1	1	39	195	1	2
7	10	1	1	40	220	1	2
8	23	1	1	41	161	0	2
9	28	1	1	42	199	0	2
10	28	1	1	43	217	0	2
11	28	1	1	44	245	0	2
12	29	1	1	45	8	1	3
13	31	1	1	46	10	1	3
14	32	1	1	47	11	1	3
15	37	1	1	48	23	1	3
16	41	1	1	49	25	1	3
17	41	1	1	50	25	1	3
18	57	1	1	51	28	1	3
19	62	1	1	52	28	1	3
20	74	1	1	53	31	1	3
21	100	1	1	54	31	1	3
22	139	1	1	55	40	1	3
23	20	0	1	56	48	1	3
24	258	0	1	57	89	1	3
25	269	0	1	58	124	1	3
26	8	1	2	59	143	1	3
27	10	1	2	60	12	0	3
28	10	1	2	61	159	0	3
29	12	1	2	62	190	0	3
30	14	1	2	63	196	0	3
31	20	1	2	64	197	0	3
32	48	1	2	65	205	0	3
33	70	1	2	66	219	0	3

The analysis is specified using the dialog box on the preceding page. The results are shown below.

Multi-Sample Survival Tests
Time Variable: TIME
Event Variable: RELAPSE

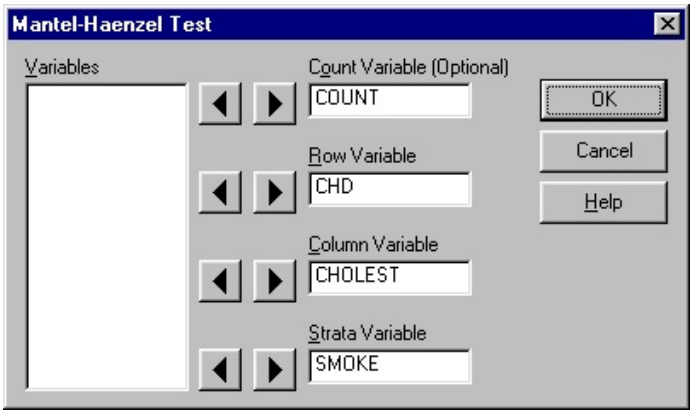
		Gehan-Wilcoxon Test		Logrank Test		Peto-Wilcoxon Test	
TRT	N	Sum	Mean	Sum	Mean	Sum	Mean
1	25	-273.00	-10.920	6.6349	0.2654	4.1072	0.1643
2	19	170.00	8.9474	-3.6934	-0.1944	-2.6586	-0.1399
3	22	103.00	4.6818	-2.9415	-0.1337	-1.4486	-0.0658
Chi-Square		3.61		3.81		3.46	
DF		2		2		2	
P		0.1643		0.1485		0.1769	

The p-values for all three tests are larger than 0.05. The data do not show significant differences among the three treatments.

Mantel-Haenzel Test

The Mantel-Haenzel Test is used to compare survival experience between two groups when adjustments for other prognostic factors are needed. It's often used in clinical and epidemiologic studies as a method of controlling the effects of confounding variables. The data are stratified by the confounding variable and cast into a sequence of 2 X 2 tables.

Specification



The test builds a series of 2 X 2 contingency tables. The data can be presented as raw data where each case represents one subject. Or the data can be tabulated where each case in the data set represents one cell in the contingency table. In the later case, a count variable is needed to supply *Statistix* with the counts for each cell of the table. If your data are already tabulated, move the variable containing the counts to the ***Count Variable*** box. If your data are not tabulated, leave the Count Variable box empty.

Move the categorical variable that you want to use to identify the rows of the 2 X 2 tables to the ***Row Variable*** box. Move the categorical variable that you want to use to identify the columns to the ***Column Variable*** box. Move the categorical variable that you want to use to identify the levels of the confounding factor to the ***Strata Variable*** box.

Data Restrictions

The row, column, and strata variables may be of any data type. Real values are truncated to whole numbers and must be no larger than 99,999. Strings are truncated to ten characters. The maximum number of strata is 500.

Example

The example data are from Lee (1992). Five hundred and ninety-five people participate in a case control study of the association of cholesterol and coronary heart disease (CHD). Among them, 300 people are known to have CHD and 295 do not. To find out if elevated cholesterol is significantly associated with CHD, the investigator decides to control the effects of smoking. The study subjects are divided into two strata: smokers and nonsmokers. The data are presented below (see Sample Data\CHD.sx).

CASE	COUNT	CHD	CHOLEST	SMOKE
1	120	With CHD	Elevated	Smokers
2	20	W/O CHD	Elevated	Smokers
3	80	With CHD	Normal	Smokers
4	60	W/O CHD	Normal	Smokers
5	30	With CHD	Elevated	Nonsmokers
6	60	W/O CHD	Elevated	Nonsmokers
7	70	With CHD	Normal	Nonsmokers
8	155	W/O CHD	Normal	Nonsmokers

The data are already tabulated with the counts in the variable COUNT. The analysis is specified on the preceding page. The results are displayed below. Cholesterol

Mantel-Haenzel Test

SMOKE	CHD	CHOLEST Normal	CHOLEST Elevated	Total	Percent Elevated
Nonsmokers	W/O CHD	155	60	215	27.9
	With CHD	70	30	100	30.0
	TOTAL	225	90	315	28.6
Smokers	W/O CHD	60	20	80	25.0
	With CHD	80	120	200	60.0
	TOTAL	140	140	280	50.0

Chi-Square	16.22
DF	1
P	0.0001

The results display the frequencies for the stratified 2 X 2 tables including row and column subtotals. The rightmost column displays the percent of the second level for the column variable for each row. The chi-square for the Mantel-Haenzel Test is 16.22 with an associated p-value of 0.0001. We conclude that elevated cholesterol is significantly associated with CHD after adjusting for the effects of smoking.

Computational Notes

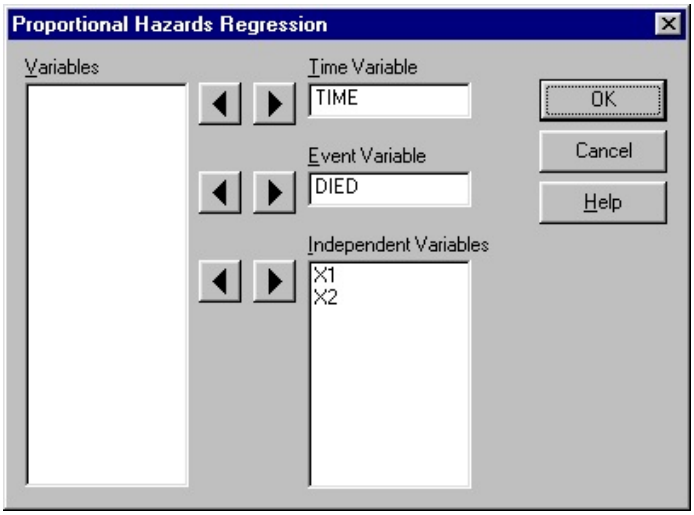
The computations follow Lee (1992). We do not use the correction for continuity.

Proportional Hazards Regression

This procedure computes Cox's proportional hazards regression for survival data. The model assumes that individuals have hazard functions that are proportional to one another, that is, that the ratio of the hazard functions for two individuals does not vary with time.

Proportional hazards regression is computed using the ranks of the survival times. While it is useful for studying the relationships among the covariates, it can't be used to build prediction equations.

Specification



Select the name of the variable containing the survival times and move it to the ***Time Variable*** box. Select the variable used to indicate whether or not the event of interest (e.g., death) occurred or not and move it to the ***Event Variable*** box. The event variable must be coded 0 if the event did not occur (i.e., censored observation) and 1 if the event did occur (i.e., uncensored observation). Move one or more independent variables to the ***Independent Variables*** box. The independent variables can be continuous or discrete variables. Discrete variable must entered using indicator (0 or 1) variables.

Data Restrictions

You can include up to 50 independent variables in the model. Discrete variables must be coded using indicator (0 or 1) variables.

Example

The example survival data in the table below are from 30 patients with AML (Lee, 1992, p. 257). Two possible prognostic factors X1 and X2 are considered. X1 is coded 1 if the patient was 50 years old or older, and 0 otherwise. X2 is coded 1 if cellularity of marrow clot section is 100%, and 0 otherwise.

CASE	TIME	DIED	X1	X2	CASE	TIME	DIED	X1	X2
1	18	1	0	0	16	8	1	1	0
2	9	1	0	1	17	2	1	1	1
3	28	0	0	0	18	26	0	1	0
4	31	1	0	1	19	10	1	1	1
5	39	0	0	1	20	4	1	1	0
6	19	0	0	1	21	3	1	1	0
7	45	0	0	1	22	4	1	1	0
8	6	1	0	1	23	18	1	1	1
9	8	1	0	1	24	8	1	1	1
10	15	1	0	1	25	3	1	1	1
11	23	1	0	0	26	14	1	1	1
12	28	0	0	0	27	3	1	1	0
13	7	1	0	1	28	13	1	1	1
14	12	1	1	0	29	13	1	1	1
15	9	1	1	0	30	35	0	1	0

The analysis is specified on the preceding page. The results are displayed below.

```
Proportional Hazards Regression

Time Variable:  TIME
Event Variable: DIED

Variable   Coefficient   Std Error        Z          P  Rel Risk
X1             1.01317     0.45740     2.22     0.0268      2.75
X2             0.35025     0.43917     0.80     0.4252      1.42

Log Likelihood, No Variables   -65.83
Log Likelihood, Model          -63.12
Chi-Square, Overall Model        5.43
DF                                  2
P                              0.0663
```

The coefficient table lists the regression coefficients, standard errors of the coefficients, z-statistics (coefficient/standard error), p-values, and relative risks ($e^{coefficient}$) for each independent variable. The z-statistic and the associated p-value tests the null hypothesis that the coefficient equals zero. The test for the independent variable X1 is significant at the .05 level. The positive sign for the coefficient indicate that the older patients have a higher risk of dying. The estimated risk of dying for patients at least 50 years of age is 2.75 times higher than that for patients less than 50.

The statistics below the list of regression coefficients are used to test the fit of the overall model. The chi-square value given is called the likelihood

test. It tests whether the independent variables in the model contribute to the prediction of survivorship. The test for this example is almost significant at the .05 level.

Regression Results Menu

Once the proportional hazards regression analysis is computed and displayed, a ***Results*** pull-down menu appears on the menu at the top of the Statistix window. Click on the Results menu to display the proportional hazards regression results menu show below.

Select Coefficient Table from the menu to redisplay the regression coefficient table displayed on the preceding page. Select Options to return to the dialog box used to specify the regression model. The Var-covar of Betas menu item is described below.

Variance-Covariance of Betas

Select this option to obtain the variance-covariance matrix of the regression coefficient estimates. The matrix for the AML example are displayed below.

Variance-Covariance Matrix for Coefficients

	X1	X2
X1	0.20921	
X2	0.16525	0.19287

Computational Notes

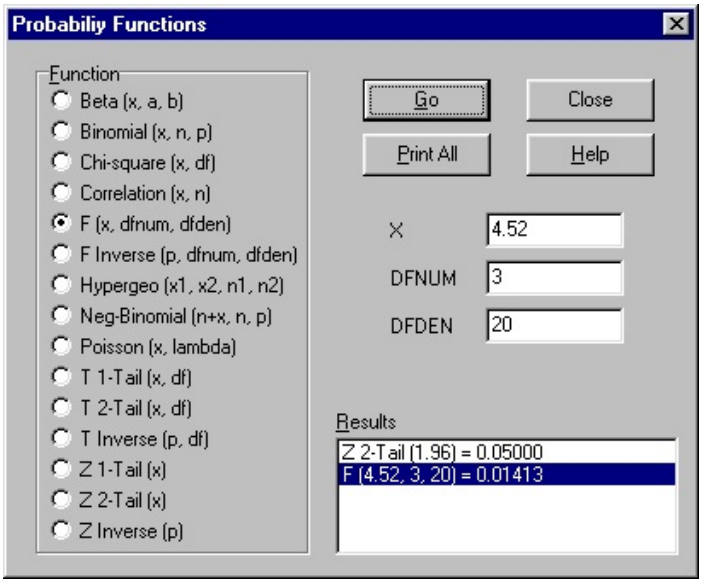
Computations follow Kalbfleisch and Prentice (1980).

C H A P T E R

13

Probability Functions

Statistix offers a number of useful procedures to calculate the probabilities for various probability distributions and the inverse functions for the standard normal, the t-distribution, and the F-distribution. The function names and arguments are displayed at the left of the dialog box shown below, which appears when **Probability Functions** is selected from the **Statistics** menu.



First select the probability function or inverse function you want to use. Once you've selected a function, the prompts for the arguments appear next to the four edit controls. Enter numbers for each of the arguments displayed. In the list of function names, the position of the random variable is usually represented as an **x**, although there are some minor variations on this which are described below. Parameters are indicated in the dialog box with logical abbreviations.

After you've entered all of the values, press the ***Go*** button and the result will be computed and displayed in the ***Results*** list box.

Press the ***Print All*** button to print the results. The entire contents of the Results list box are printed, then the Results box is emptied. Press the ***Close*** button to exit the Probability Functions procedure.

The functions are described in detail below. We use standard notation to describe the region of the distribution for which the probability is being calculated. For example, the expression "Pr ($y \leq X$)" represents the probability of a value of a random variable y equal to or less than some specified value X. The inverse functions compute the test statistic for a lower-tail probability.

BETA (X, A, B)
Lower-Tail Beta Probability Distribution

This function computes Pr ($y \leq X$) for a beta random variable y with parameters A and B. The beta distribution is very flexible. The parameters A and B control the shape of the distribution, the mean of the distribution is given by A / (A + B). The beta distribution is sometimes used itself for tests and several other important distributions are easily derived from it, such as the t- and F-distributions (Kennedy and Gentle 1980).

The values permitted for X range from 0 to 1. A and B must be positive, but they don't need to be integer values. The beta distribution can be used to compute probability values for generalized t and F random variables with noninteger degrees of freedom—T 1-Tail, T 2-Tail, and F functions require integer degrees of freedom.

BINOMIAL (X, N, P)
Lower-Tail Binomial Probability Distribution

This function computes Pr ($y \leq X$) for a random variable y from a binomial distribution with N trials and parameter P. In many situations, the random variable is thought of as the number of "successes" out of N independent trials (i.e., there were N-X "failures"). The parameter P is the probability of a success on a particular trial. In other words, Binomial computes the probability of observing X or fewer successes out of N trials when the success rate per trial is P.

X and N must be specified as integers. The parameter P must be between zero and one. To find the upper-tail distribution, you can use the relationship Pr ($y > X$) = 1 - Pr ($y \leq X$).

CHI-SQUARE (X, DF)
Upper-Tail Chi-Square Probability Distribution

This function computes Pr ($y \geq X$) for a random variable y from a central chi-square distribution with DF degrees of freedom. In other words, it computes the probability of a value equal to or larger than X. The typical chi-square tests in goodness-of-fit analyses use the upper-tail probabilities. Use the relationship Pr ($y < X$) = 1 - Pr ($y \geq X$) if you want a lower-tail probability.

X must be a positive number and the degrees of freedom must be a positive integer.

CORRELATION (X, N)
Two-Tailed Probability Distribution for the Correlation Coefficient

This function computes Pr ($*y* \geq *X*$) for a random variable y, where y is a simple (Pearson) correlation coefficient computed from N pairs of data. This distribution is appropriate for testing the null hypothesis that the correlation coefficient is equal to zero; the distribution being computed assumes the true correlation coefficient is zero. Snedecor and Cochran (1980, sect. 10.5) discuss the application of this procedure and the assumptions required. This procedure is equivalent to testing the hypothesis that the slope of the line is equal to zero in simple linear regression; the assumptions required are the same.

The values permitted for X range between -1 and 1. The number of pairs N must be a positive integer greater than two.

F (X, DFNUM, DFDEN)
Upper-Tail F Probability Distribution

This function computes Pr ($y \geq X$) for a random variable y from a central F-distribution with DFNUM numerator degrees of freedom and DFDEN denominator degrees of freedom. In other words, it computes the probability of a value equal to or larger than the observed X. The typical F tests in regression and analysis of variance use the upper-tail probabilities. Use the relationship Pr ($y < X$) = 1 - Pr ($y \geq X$) if you want a lower-tail probability.

X must be a positive number and both degrees of freedom must be positive integers.

F INVERSE (P, DFNUM, DFDEN)
Inverse of the F-Distribution

This function computes the F test statistic for which the probability of a smaller value is P.

HYPERGEO (X1, X2, N1, N2)
Lower-Tail Hypergeometric Probability Distribution

This function computes Pr ($y1 \leq X1$), where y1 is a random variable drawn from a hypergeometric distribution. Using the traditional "urn" model, N1 corresponds to the number of red balls initially in the urn, and N2 corresponds to the number of black balls. Then, y1+y2 balls are randomly drawn from the urn. HYPER computes the probability of observing X1 or fewer red balls in such a sample.

All four parameters must be nonnegative integers.

NEG-BINOMIAL (N+X, N, P)
Lower-Tail Negative Binomial Probability Distribution

This function computes Pr ($N + y \leq N + X$) or equivalently Pr ($y \leq X$) for a random variable y which follows a negative binomial distribution. Typically, N+X is referred to as the number of trials required to get N successes. The probability of a success on a particular trial is the parameter p. In other words, Neg-Binomial computes the probability of requiring N+X or fewer trials to get N successes. X is the number of failures observed before the Nth success.

N+X and N must be positive integers, and N+X must be greater than N. The parameter p may range from zero to one.

POISSON (X, LAMBDA)
Lower-Tail Poisson Probability Distribution

This function computes Pr ($y \leq X$) for a random variable y from a Poisson distribution with rate parameter LAMBDA. In some situations, the random variable is thought of as the number of random events in some interval of time or space, and the rate parameter LAMBDA is the average number of such events expected in the interval. In other words, POISSON computes the probability of observing X or fewer events in an interval if the expected number of events is LAMBDA. You can find the upper-tail distribution by using the relationship Pr ($y > X$) = 1 - Pr ($y \leq X$).

X must be an integer value and LAMBDA must be greater than zero.

T 1-TAIL (X, DF)
One-Tailed Probability Value for Student's T-Distribution

This function computes Pr ($y \leq X$) for $X \leq 0$, and Pr ($y \geq X$) for $X > 0$ for a random variable y from a central t-distribution with DF degrees of freedom. In other words, T 1-Tail computes the probability of a t value equal to or more extreme than X, taking into account the sign of X. This is often called the one-tailed significance of X.

DF must be a positive integer value.

T 2-TAIL (X, DF)
Two-Tailed Probability Value for Student's T-Distribution

This function computes Pr (*y* ≥ *X*) for a random variable y from a central t-distribution with DF degrees of freedom. In other words, T 2-Tail computes the probability of a t value with an absolute value equal to or larger than the absolute value of X. This is often called the two-tailed significance of X.

DF must be a positive integer value.

T INVERSE (P, DF)
Inverse of the Student's T-Distribution

This function computes the Students's t test statistic for which the probability of a smaller value is P.

Z 1-TAIL (X)
One-Tailed Probability Value for the Standard Normal Distribution

This function computes Pr (y ≤ X) for X ≤ 0, and Pr (y > X) for X > 0 for a standard normal random variable y. In other words, Z 1-Tail computes the probability of a value equal to or more extreme than X, taking into account the sign of X. This is often called the one-tailed significance of X. A statistic with a standard normal distribution is often referred to as a Z statistic, and hence the function name.

A standard normal distribution has a mean of zero and a variance of one. A normally distributed statistic can be transformed to standard normal form by subtracting the mean and then dividing by the standard deviation.

Z 2-TAIL (X)
Two-Tailed Probability Value for the Standard Normal Distribution

This function computes Pr (*y* ≥ *X*) for a standard normal random variable y. In other words, Z 2-Tail computes the probability of a value with an absolute value equal to or larger than the absolute value of X. This is often called the two-tailed significance of X. A statistic with a standard normal distribution is often referred to as a Z statistic, and hence the

function name.

A standard normal distribution has a mean of zero and a variance of one. A normally distributed statistic can be transformed to standard normal form by subtracting the mean and then dividing by the standard deviation.

Z INVERSE (P)
Inverse of the Standard Normal Distribution

This function computes the standard normal value z for which the probability of a smaller value is P.

Computational Notes

Three key functions are used to generate the various probabilities—the error function, the log gamma function, and the incomplete beta function. The error function is patterned after a routine suggested by Kennedy and Gentle (1980). The method used to calculate the log gamma function is similar to that used at the University of Wisconsin computer center (Reference Manual 1410 - Probability Distribution Functions). Representing the gamma function as G(X), an asymptotic expansion is used directly when $X \geq 8$. Otherwise, the relationship $G(X + 1) = XG(X)$ is applied until the expansion can be used. The procedure for computing the incomplete beta function is patterned after the IMSL routine MDBETA, which is discussed in Kennedy and Gentle (1980). To speed up computation, a large sample approximation for the incomplete beta function is used for certain "safe" parameter values (Abramowitz and Stegun, eq. 26.5.21).

All the probability functions are based on relatively simple functions of these three functions. Consult Kennedy and Gentle (1980) for further detail.

R E F E R E N C E S

Abraham, B. and J. Ledolter. 1983. *Statistical Methods for Forecasting*. Wiley. New York.

Abramowitz, M. and I. A. Stegun. 1977. *Handbook of Mathematical Functions*. National Bureau of Standards. Washington, D.C.

Begun, J. M. and K. R. Gabriel. 1981. Closure of the Newman-Keuls multiple comparisons procedure. *Journal of the American Statistical Association*. 76:374.

Bickel, P. J. and K. D. Doksum. 1977. *Mathematical Statistics*. Holden-Day. San Francisco, California.

Bingham, C. and S. E. Fienberg. 1982. Textbook analysis of covariance - is it correct? *Biometrics*. 38:747-753.

Bishop. Y. M. M., S. E. Fienberg and P. W. Holland. 1975. *Discrete Multivariate Analysis*. MIT Press. Cambridge, Massachusetts.

BMDP-83. Dixon, W. J. (ed). 1983. *BMDP Statistical Software - 1983 Printing with Additions*. Berkeley, California.

Bowdler, H., R. S. Martin, C. Reinsch and J. H. Wilkinson. 1968. The QR and QL algorithms for symmetric matrices. *Numerical Mathematics*. 11:293-306.

Bowker, A. H. 1948. A test for symmetry in contingency tables. *Journal of the American Statistical Association*. 43:572-574.

Box, G. E. P. and G. W. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*. Holden-Day. San Francisco, California.

Bradley, J. V. 1968. *Distribution-free Statistical Tests.* Prentice-Hall. Englewood Cliffs, New Jersey.

Brown, M. B., and J. K. Benedetti. 1977. Sampling behavior of tests for correlation in two-way contingency tables. *Journal of the American Statistical Association*. 72:309-315.

Chatterjee, S. and Price B. 1991. *Regression Analysis by Example*. 2nd ed. Wiley. New York.

Clarke, M. R. B. 1981. A Givens algorithm for moving from one linear model to another without going back to the data. *Applied Statistics*. 30:198-203.

Conover, W. J. 1980. *Practical Nonparametric Statistics.* 2nd ed. Wiley. New York.

Conover, W. J. and R. L. Iman. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*. 35:124-129.

Cook, R. D. 1977. Detection of influential observations in linear regression. *Technometrics*. 19:15-18.

Cook, R. D. 1979. Influential observations in linear regression. *Journal of the American Statistical Association*. 74:169-174.

Cook, R. D. and S. Weisberg. 1982. *Residuals and Influence in Regression.* Chapman and Hall, New York.

Cooper. B. E. 1968. The use of orthogonal polynomials. *Applied Statistics*. 17:283-287.

Cox, D. R. 1970. *The Analysis of Binary Data*. Chapman and Hall, London.

Crowder, M. J. and D. J. Hand. 1990. *Analysis of Repeated Measures.* Chapman and Hall. New York.

Daniel, C. 1976. *Applications of Statistics to Industrial Experimentation.* Wiley. New York.

Daniel, C. and F. Wood. 1971. *Fitting Equations to Data.* Wiley. New York.

Daniel, W. W., 1990. *Applied Nonparametric Statistics*. 2nd ed. PWS-Kent.

Dineen, L. C. and B. C. Blakesley. 1973. A generator for the sampling distribution of the Mann-Whitney U statistic. *Applied Statistics*. 22:269-273.

Draper, N. R. and H. Smith. 1966. *Applied Regression Analysis.* Wiley. New York.

Durbin, J. and G. S. Watson. 1950. Testing for serial correlation in least squares regression I. *Biometrika*. 37:409-428.

Durbin, J. and G. S. Watson. 1951. Testing for serial correlation in least squares regression II. *Biometrika*. 38:159-178.

Durbin, J. and G. S. Watson. 1971. Testing for serial correlation in least squares regression III. *Biometrika*. 58:1-19.

Einot, I. and K. R. Gabriel. 1975. A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*. 70:351.

Federer, W. T. 1957. Variance and covariance analyses for unbalanced classifications. *Biometrics*. 13:333-362.

Fienberg, S. E. 1980. *The Analysis of Cross Classified Categorical Data.* MIT Press. Cambridge, Massachusetts.

Gentleman, W. M. 1973. Basic procedures for large, sparse or weighted least squares problems. *Applied Statistics*. 23:448-454.

Glantz, Stanton A. and Bryan K. Slinker. 1990. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill. New York.

Gomez, Kwanchai A. and Arturo A. Gomez. 1984. *Statistical Procedures for Agricultural Research.* 2nd ed. Wiley. New York.

Gordon, H. A. 1981. Errors in computer packages. Least squares regression through the origin. *The Statistician*. 30:23-29.

Griffiths, William E., R. Carter Hill, and George G. Judge. 1993. *Learning and Practicing Econometrics.* Wiley. New York.

Haberman, S. J. 1972. Log-linear fit for contingency tables. *Applied Statistics*. 21:218-225.

Haberman, S. J. 1978. *The Analysis of Qualitative Data.* Vols. I & II. Academic Press. New York.

Hald, A. 1952. *Statistical Theory with Engineering Applications.* Wiley. New York.

Hanke, J. E. and A. G. Reitsch. 1989. *Business Forecasting.* Allyn and Bacon. Boston, Massachusetts.

Heisey, D. M. 1985. Analyzing selection experiments with log-linear models. *Ecology*. 66:1744-1748.

Hill, G. W. 1970. Student's t quantiles. *CACM*. 13:619-620.

Hollander, M. and D. A. Wolfe. 1973. *Nonparametric Statistical Methods.* Wiley. New York.

Hosmer, D. W. and S. Lemeshow. 1989. *Applied Logistic Regression.* Wiley. New York.

Hsu, Jason C. 1996. *Multiple Comparisons: Theory and Methods.* Chapman and Hall. New York.

Iman, R. L. and J. M. Davenport. 1976. New approximations to the exact distribution of the Kruskal-Wallis test statistic. *Communications in Statistics*. Ser. A. 5:1335-1348.

Iman, R. L. and J. M. Davenport. 1980. Approximations of the critical region of the Friedman statistic. *Communications in Statistics*. Ser. A. 9:571-595.

IMSL - International Mathematical and Statistical Libraries, Inc. 1975. *IMSL Library 1 Reference Manual.* IMSL, 7500 Bellaire Blvd., Floor 6, GNB Building, Houston, Texas.

Jolliffe, I. T. 1982. A note on the use of principal components in regression. *Applied Statistics*. 31:300-303.

Kalbfleisch, J. D. and R. L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. Wiley. New York.

Kennedy, W. J. and J. E. Gentle. 1980. *Statistical Computing.* Dekker. New York.

Lee, E. T. 1992. *Statistical Methods for Survival Data Analysis*. 2nd. ed. Wiley. New York.

Lehmann, E. L. 1975. *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day. San Francisco, California.

Ljung, G. and G. E. P. Box. 1978. On a measure of lack of fit in time series models. *Biometrika*. 65:297-304.

Lund, R. E. and J. R. Lund. 1983. Algorithm 190. Probabilities and upper quantiles for the Studentized range. *Applied Statistics*. 32:204-210.

Majumder, K. L. and G. P. Bhattacharjee. 1973. Algorithm 64. Inverse of the incomplete beta function ratio. *Applied Statistics*. 22:411-414.

Manly, Bryan F. J. 1991. *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall, London.

Martin, R. S., C. Reinsch and J. H. Wilkinson. 1968. Householder's tri-diagonalization of a symmetric matrix. *Numerical Mathematics*. 11:181-195.

Maurice, S. Charles and Christopher R. Thomas. 2002. *Managerial Economics.* 7th ed. McGraw-Hill. New York.

Mercier, L. J. 1987. *Handbook for Time Series Analysis*. AFIT. Dayton, Ohio.

McCullagh, P. and J. A. Nelder. 1983. *Generalized Linear Models.* Chapman and Hall. London.

Montgomery, D. C. 1991. *Introduction to Statistical Quality Control.* 2nd. ed. Wiley. New York.

Morrison, D. F. 1977. *Multivariate Statistical Methods.* 2nd. ed. MacGraw-Hill. New York.

Nash, J. C. 1979. *Compact Numerical Methods For Computers: Linear Algebra and Function Minimisation.* Wiley. New York.

Nelder, J. A. and R. W. M. Wedderburn. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society*. 135:370-384.

Nelson, L. S. 1984. The Shewhart Control Chart - Tests for Special Causes. *Journal of Quality Technology*. 16:237-239.

Oliver, I. 1967. Analysis of factorial experiments using generalized matrix operations. *Journal of the Association of Computing Machinery*. 14:508-519.

Pearson, E. S., and H. O. Hartley. 1954. *Biometrika Tables for Statisticians*. Cambridge University Press.

Pregibon, D. 1981. Logistic regression diagnostics. *Annals of Statistics*. 9:705-724.

Royston, J. P. 1982. Expected normal order statistics (exact and approximate) *Applied Statistics*. 31:161-165.

Royston, J. P. 1995. A remark on algorithm AS 181: The W-test for normality. *Applied Statistics*. 44:547-551.

Ryan, T. A. 1960. Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychological Bulletin*. 57:318-328.

Scheffe, H. 1959. *The Analysis of Variance.* Wiley. New York.

Searle, Shayle R. 1987. *Linear Models for Unbalanced Data.* Wiley. New York.

Seber, G. A. F. 1977. *Linear Regression Analysis*. Wiley. New York.

Shapiro, S. S. and R. S. Francia. 1972. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*. 67:215-216.

Shapiro, S. S. and M. Wilk. 1965. An analysis of variance test for normality. *Biometrika*. 52:591-611.

Siegel, S., and N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences.* 2nd ed. McGraw-Hill. New York.

Simon, R. and Y. Lee. 1982. Non parametric confidence limits for survival probabilities and the median. *Cancer Treatment Reports*. 66:37-42.

Smirnov, N. V. 1939. Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*. 2(2):3-16.

Snedecor, G. W. and W. G. Cochran. 1980. *Statistical Methods.* 7th ed. The Iowa State University Press. Ames, Iowa.

Steel, Robert G. D. and James H. Torrie. 1980. *Principles and Procedures of Statistics: A Biometrical Approach.* 2nd ed. McGraw-Hill. New York.

Thomopoulos, N. T. 1980. *Applied Forecasting Methods*. Prentice-Hall. Englewood Cliffs, New Jersey.

Velleman, P. and D. Hoaglin. 1981. *ABC's of EDA.* Duxbury Press.

Weisberg, S. 1982. *MULTREG User's Manual.* Technical Report #298, School of Statistics, University of Minnesota. St. Paul, Minnesota.

Weisberg, S. 1985. *Applied Linear Regression.* 2nd ed. Wiley. New York.

Weisberg, S. and C. Bingham. 1975. An analysis of variance test for normality suitable for machine calculation. *Technometrics*. 17:133.

Welsch, R. E. 1977. Stepwise multiple comparison procedures. *Journal of the American Statistical Association*. 72:359.

Wichura, Michael J. 1988. The percentage points of the normal distribution. *Applied Statistics*. 37:477-484.

I N D E X