STATISTICA™

Quick Reference

Chapter

Appendixes

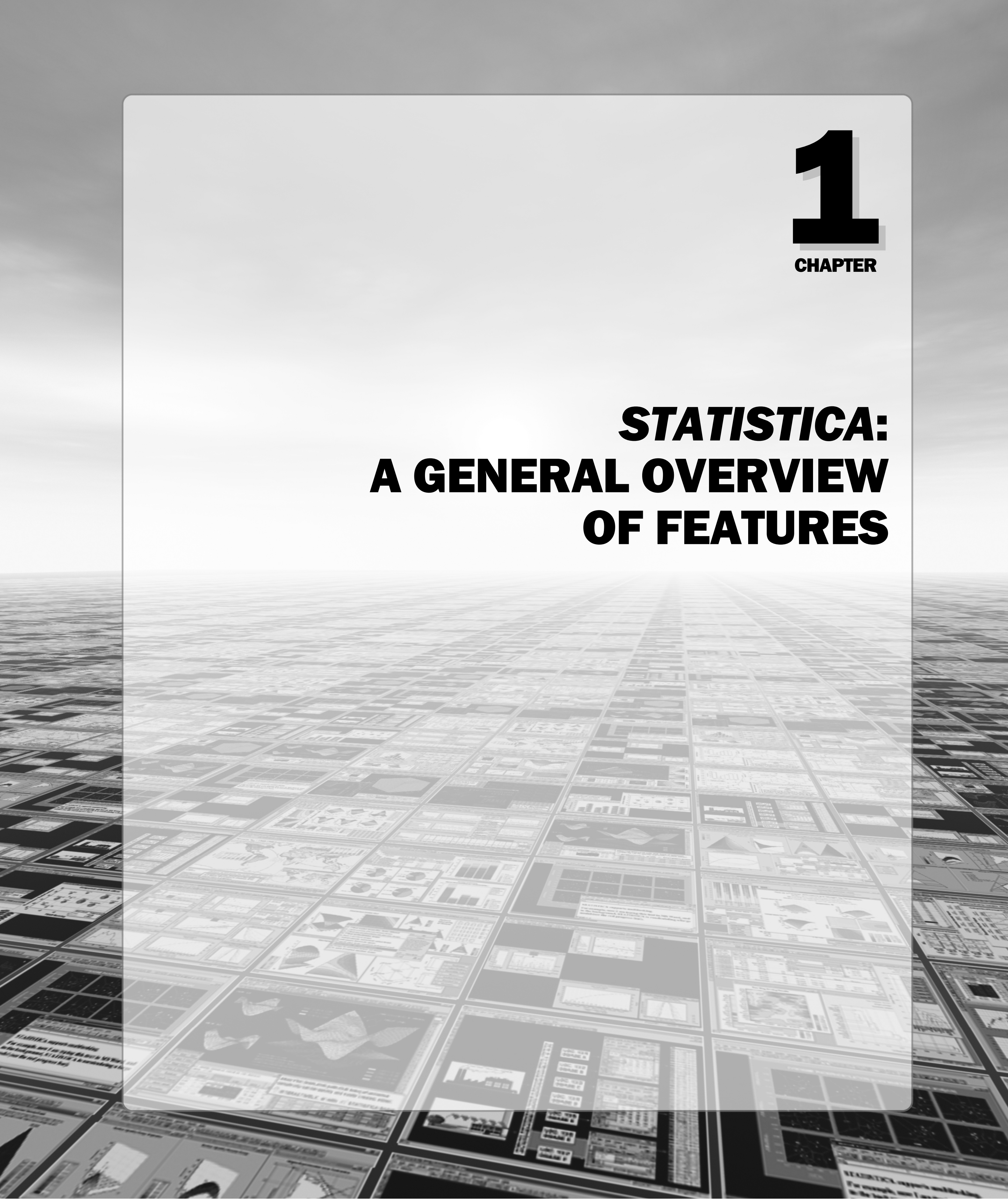
StatSoft®

www.statsoft.com

STATISTICA Quick Reference

Contents

APPENDIXES



1

CHAPTER

STATISTICA: A GENERAL OVERVIEW OF FEATURES

1

CHAPTER

STATISTICA: A GENERAL OVERVIEW OF FEATURES

STATISTICA is a comprehensive analytic, research, and business intelligence tool. It is an integrated data management, analysis, mining, visualization, and custom application development system featuring a wide selection of basic and advanced analytic procedures for business, data mining, science, and engineering applications.

Analytic Facilities

STATISTICA includes not only general purpose analytic, graphical, and database management procedures, but also comprehensive implementations of specialized methods for data analysis (e.g., predictive data mining; business, social sciences, and biomedical research; or engineering applications). All analytic tools offered in the *STATISTICA* line of software are available as part of an integrated package. These tools can be controlled through a selection of alternative user interfaces including:

- a highly optimized interactive user interface (with options to execute *STATISTICA* from within Microsoft Office and other applications),
- a complete thin-client, browser-based user interface (in *STATISTICA Enterprise Server*) that enables you to offload time-consuming tasks to the server and work collaboratively, and

- a comprehensive, industry standard, .NET-compatible programming interface (including the built-in, .NET-compatible Visual Basic), offering access to more than 14,000 externally callable functions.

Interactive user interfaces can be easily automated via macros and customized using a variety of methods, and they are recordable in the form of industry standard VB scripts. The built-in development environment can be used to interface *STATISTICA* with other applications and enterprise-wide infrastructures or to build custom extensions of any complexity, from simple shortcuts to advanced, large-scale development projects.

Unique Features

Some of the unique features of the *STATISTICA* line of software include:

- the breadth of selection and comprehensiveness of implementation of analytical procedures,
- the unparalleled selection, quality, and customizability of graphics integrated seamlessly with every computational procedure,
- a selection of efficient and user-friendly user interfaces,
- the ease of customizability using the truly open architecture compatible with virtually all enterprise and development environments (including .NET), that exposes *STATISTICA*'s more than 14,000 functions,
- a wide selection of advanced software technologies (see *Software Technology*, page 6) that is responsible for *STATISTICA*'s practically unlimited capacity, performance (speed, responsiveness), and application customization options,
- native R scripts can be run directly within *STATISTICA* and R output can be retrieved as native *STATISTICA* Spreadsheets and Graphs.

One of the most unique and important features of the *STATISTICA* family of applications is that these technologies enable even inexperienced users to tailor *STATISTICA* to their specific preferences. You can customize practically every aspect of *STATISTICA,* including even the low-level procedures of its user interface. The same version of *STATISTICA* can be used:

- By novices to perform routine tasks using the default analysis Startup dialog ***Quick*** tab (containing just a few, self-explanatory buttons), or even by accessing *STATISTICA* with their Web browsers (and a highly simplified "front end"), and
- By experienced analysts, professional statisticians, and advanced application developers who can integrate any of *STATISTICA*'s highly optimized procedures (more than 14,000 functions) into custom applications or computing environments, using any of the cutting edge .NET and Web-compatible technologies.

The General "Philosophy" of the *STATISTICA* Approach

STATISTICA's default configuration (its general user interface and system options) is a result of years of listening carefully to our users.

We have received feedback from tens of thousands of our users, representing hundreds of thousands of our users from all continents and, practically speaking, "all walks of life." One of the most important facts that we have learned from these users is how different their needs and preferences are (both across individuals and projects or applications). In order to meet those differentiated needs, *STATISTICA* is designed to offer perhaps one of the most flexible and easily customizable user interfaces of any contemporary application.

Although *STATISTICA* provides access to a powerful arsenal of advanced software technologies (see *Software Technology*, page 6), you do not even need to know about them, because they are designed to work automatically and intuitively. A novice user may never see more than a few self-explanatory buttons. Advanced options, however, are only one tab or mouse click away. Practically every aspect of *STATISTICA* (from the startup configuration, to the way the output is generated and managed by the system, to how *STATISTICA* prompts you to choose your next step) can be changed with a mouse click. Moreover, *STATISTICA* remembers your selections until you change your mind. Practically all dialogs used to select an analysis or perform a routine operation can be easily replaced (e.g., simplified, enhanced, or combined with custom, user-designed procedures). *STATISTICA* will always look and work the way you want.

Software Technology (A Technical Note)

The performance, customizability, and wide selection of options that can be tailored to your needs mentioned in the previous section would not be possible if *STATISTICA* did not feature the advanced technologies that drive all functions of the application.

STATISTICA uses and/or supports virtually all the relevant leading edge software technologies available today. Every one of the more than 14,000 *STATISTICA* functions is accessible to external applications. Practically no limitations are imposed in terms of either the amount or complexity of data that can be stored and accessed. *STATISTICA* also is optimized for Web and multimedia applications. Computational and graphics procedures are driven by countless proprietary optimizations such as, for example, the "quadruple precision" computational technology that enables us to overcome the limitations of the IEEE floating point storage standards and delivers computational accuracy normally found only in designated math applications (that feature arbitrary-precision options) but not in high volume data processing applications such as statistical or data mining programs.

As a result, *STATISTICA* offers unmatched speed, numerical precision, and responsiveness, which is aided by multithreading (and the advanced "supercomputer-like" distributed/parallel processing architecture offered in the Client-Server version, i.e., *STATISTICA Enterprise Server*).

Data access is based on a flexible streaming technology that enables *STATISTICA* to work effortlessly with both the simple input data files stored on the local drive and queries of multidimensional databases containing terabytes of data and stored in remote data warehouses and processed in-place (i.e., without having to import them to a local storage; this feature is available in enterprise versions of *STATISTICA*).

For example, you can simultaneously run multiple instances of *STATISTICA* [in any combination of local, network, and Client-Server (Web-based) environments], each running multiple analyses of data from multiple and simultaneously open input data files and queries, and the results can be organized into separate projects. *STATISTICA*'s input and output data files and graphs can be of practically unlimited size, comprising hierarchies of documents of various types. The output can be directed to a multitude of output channels such as multimedia tables, high performance workbooks, reports (including *.pdf* files and Microsoft Office documents), and the Internet, as well as the optional *STATISTICA Document Management System*, which can be seamlessly integrated with any *STATISTICA* application.

Web Enablement

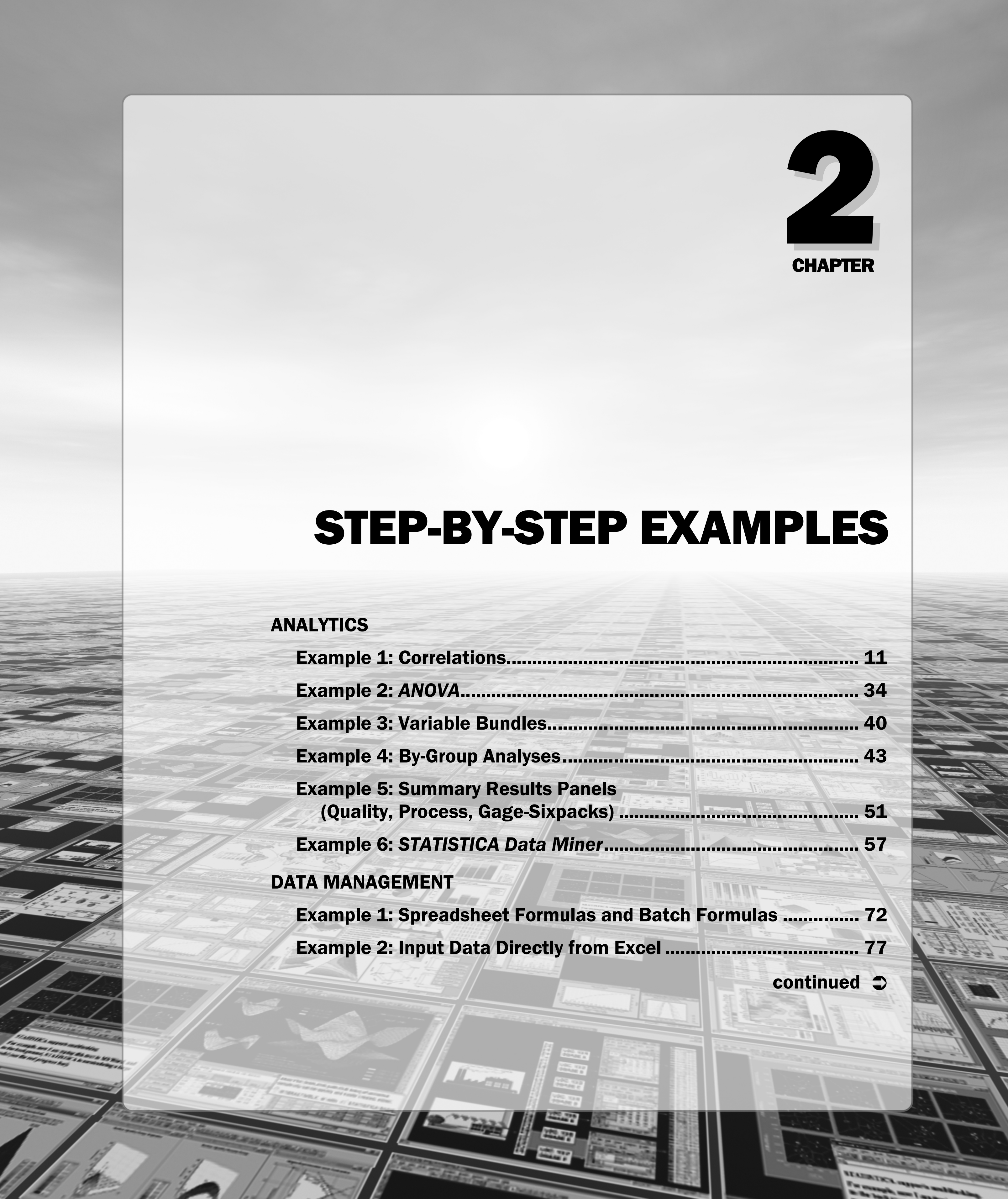
One of the unique features of the *STATISTICA* family of applications is that it is fully Web enabled, and if *STATISTICA Enterprise Server* is installed, you can not only offload time-consuming tasks to the server, but also access the comprehensive functionality of the *STATISTICA* system using a thin-client (browser) interface. This includes the option to execute prepared scripts and a plethora of interactive functionality, including such operations as interactively building predictive data mining models by dragging arrows in the interactive workspace of *STATISTICA Data Miner* (using only the browser, without any client software installed). For more information, please refer to *Appendix B – STATISTICA Enterprise Server*, page 263.

Note that most features described in this manual are available in all *STATISTICA* products, although some sections of the manual refer only to specific products such as the *STATISTICA Enterprise Server* facilities or the *STATISTICA Data Miner* line of products.

Record of Recognition

We are pleased to report that, as of this printing, *STATISTICA* has received the highest rating in every published independent comparative review in which it has been featured. In the history of the software industry, very few products have ever achieved such a record.

For more information about StatSoft and *STATISTICA*'s record of recognition, please visit our Web site at ***www.StatSoft.com***.



2

CHAPTER

STEP-BY-STEP EXAMPLES

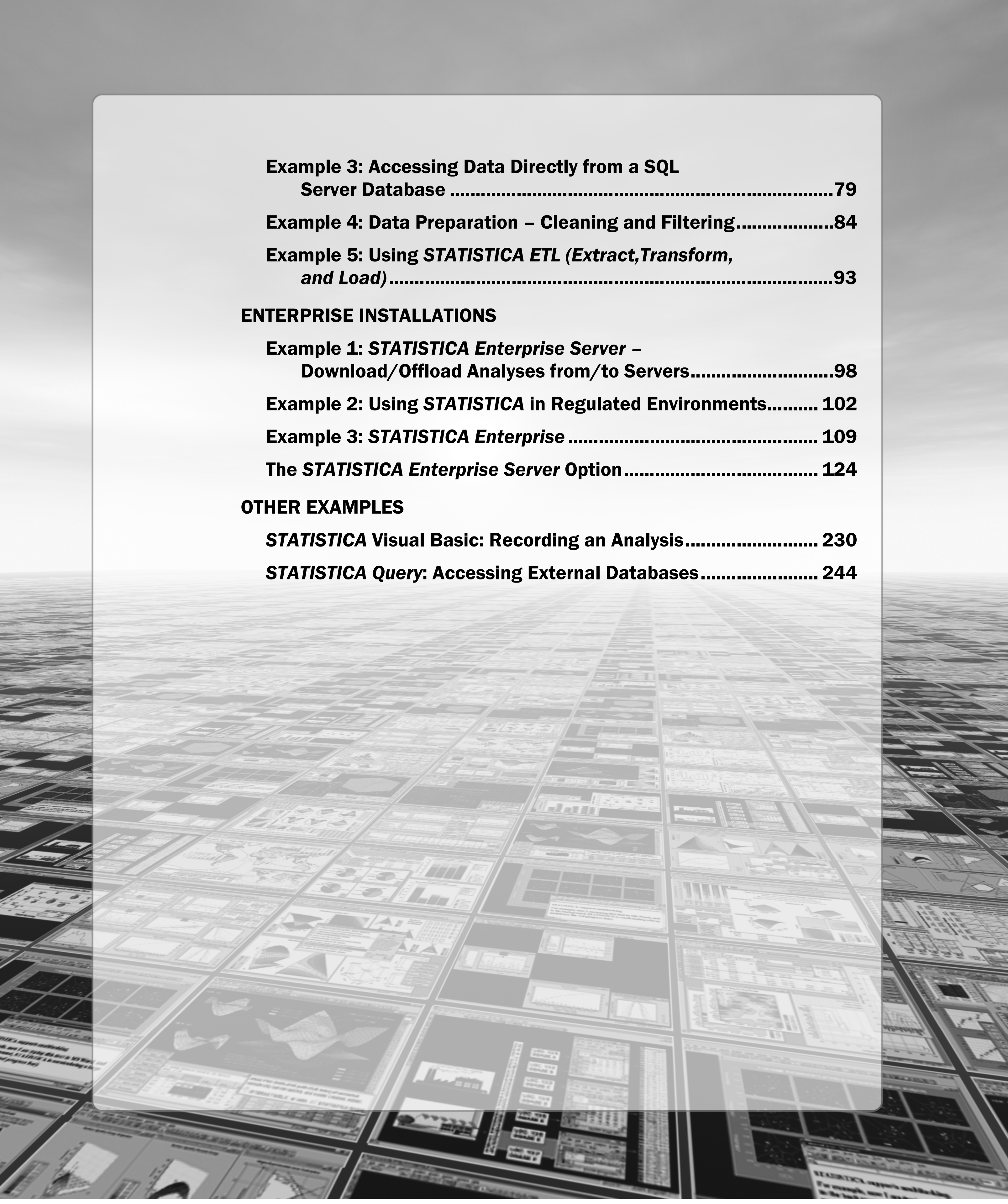
ANALYTICS

DATA MANAGEMENT

continued ➲

ENTERPRISE INSTALLATIONS

OTHER EXAMPLES



2

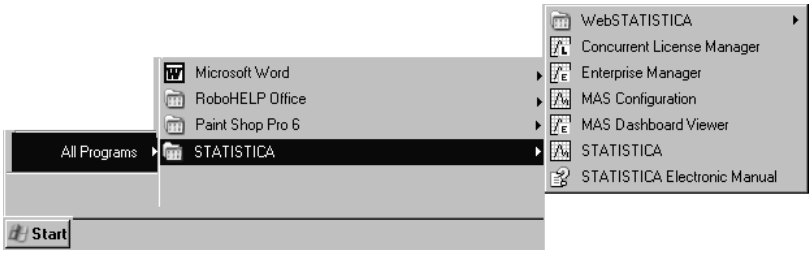
CHAPTER

STEP-BY-STEP EXAMPLES

ANALYTICS

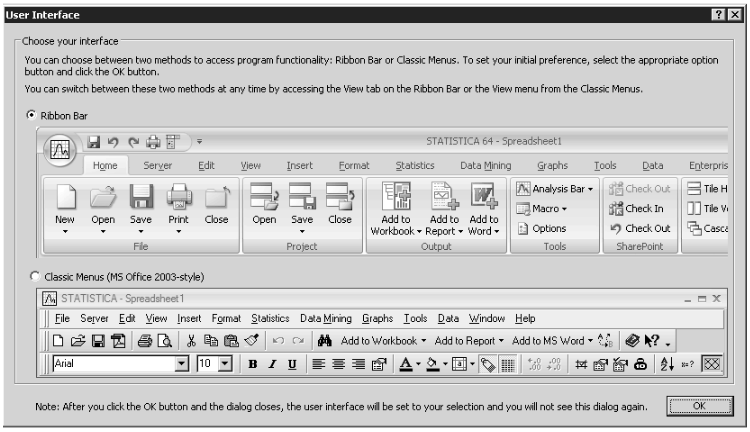
Example 1: Correlations

Starting *STATISTICA*. After installing *STATISTICA*, you can start the program by selecting ***STATISTICA*** from the Windows *Start - All Programs* submenu.



You can also double-click on either *STATIST.exe* in Windows Explorer or the icon of any *STATISTICA* file, e.g., a spreadsheet, to start the program.

When you start *STATISTICA* for the first time, the ***User Interface*** dialog is displayed, where you can choose to use the ribbon bar or the classic drop-down menus. All examples in this manual use the ribbon bar.



User Interface

Choose your interface

You can choose between two methods to access program functionality: Ribbon Bar or Classic Menus. To set your initial preference, select the appropriate option button and click the OK button.

You can switch between these two methods at any time by accessing the View tab on the Ribbon Bar or the View menu from the Classic Menus.

● Ribbon Bar

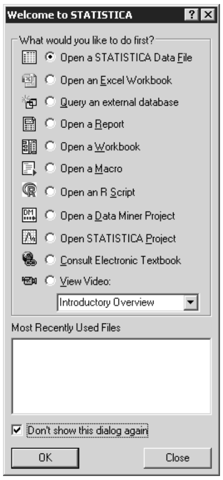
○ Classic Menus (MS Office 2003-style)

Note: After you click the OK button and the dialog closes, the user interface will be set to your selection and you will not see this dialog again.

OK

Note that it is easy to switch between the ribbon bar and the classic menus at any time. When the ribbon bar is displayed, click the menu icon on the Quick Access toolbar (located in the upper-left corner of the ribbon bar) to display the classic menus. When the classic menus are displayed, select ***Ribbon Bar*** from the ***View*** menu to display the ribbon bar.

To create more space in the application window, you can minimize the ribbon bar. Either double-click on the selected tab header, or right-click on the right side of the row of tabs and from the shortcut menu, select ***Minimize the Ribbon***.



Welcome to STATISTICA

What would you like to do first?

● Open a STATISTICA Data File
○ Open an Excel Workbook
○ Query an external database
○ Open a Report
○ Open a Workbook
○ Open a Macro
○ Open an R Script
○ Open a Data Miner Project
○ Open STATISTICA Project
○ Consult Electronic Textbook
○ View Video:

Introductory Overview

Most Recently Used Files

☑ Don't show this dialog again

OK Close

After you click ***OK*** in the ***User Interface*** dialog, the ***Welcome to STATISTICA*** dialog is displayed, which contains options that are useful to access common functions in *STATISTICA*.

If you prefer, you can select the ***Don't show this dialog again*** check box located near the bottom of the dialog, and this dialog will not be displayed when you start *STATISTICA*. Depending on the version of *STATISTICA* you have, there may be other dialogs displayed as well.

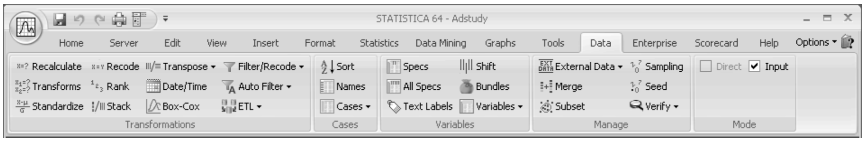
Customization of *STATISTICA*. Practically all aspects of the behavior and appearance of *STATISTICA* (even many elementary features illustrated in this example, such as where output is directed) can be permanently customized to match your preferences. For example, even the first step (opening *STATISTICA*) can be customized; you can change the default full-screen opening mode, the appearance of the data spreadsheet, and many other aspects of *STATISTICA*, which will be illustrated throughout this manual.

Selecting a data file. For this example, open *Adstudy.sta:* on the ***Home*** tab in the ***File*** group, click the ***Open*** arrow. From the drop-down menu, select ***Open Examples*** to display the ***Open a STATISTICA Data File*** dialog. Double-click on the *Datasets* folder, and double-click on *Adstudy*. You can also open data files by 1) selecting ***Open Document*** from the ***Open*** drop-down menu to display the ***Open*** dialog where you can browse to the appropriate location, 2) clicking the Open Data button located on each Startup Panel (the first dialog displayed when starting analysis or graph specifications), or 3) clicking the folder icon above ***Open*** on the ***Home*** tab.

Data spreadsheets (multimedia tables). *STATISTICA* data files are displayed in a spreadsheet (i.e., one spreadsheet is one data file). All *STATISTICA* Spreadsheets are displayed using StatSoft's powerful multimedia table technology, and they can contain not only practically unlimited amounts of data, but also sound, video, embedded documents, automation scripts, and custom user interfaces.

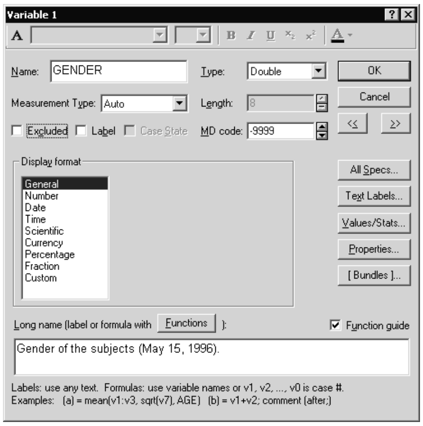
It is possible to have more than one data spreadsheet open at a time (with each spreadsheet connected to a different analysis).

Data management facilities are available on the ***Data*** tab, which is displayed whenever a spreadsheet is open. Commands on the tabs are organized in logical groups; e.g., the ***Data*** tab contains the ***Transformations***, ***Cases***, ***Variables***, ***Manage***, and ***Mode*** groups.



All the commands on the ribbon bar and classic menus are described in *STATISTICA* Help; point to (highlight) a command, and press F1 on your keyboard to display the respective Help topic.

Variable specifications. The variable (column) headers in the spreadsheet contain the variable names. Double-click on the first variable header – *GENDER* – to display its ***Variable*** specifications dialog.



Name: GENDER

Type: Double

Measurement Type: Auto

Length: 8

□ Excluded □ Label □ Case State

MD code: -9999

Display format: General, Number, Date, Time, Scientific, Currency, Percentage, Fraction, Custom

Long name (label or formula with Functions):

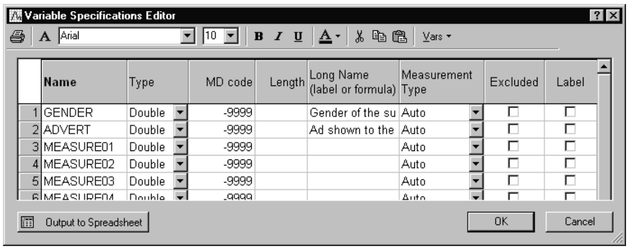
☑ Function guide

Gender of the subjects (May 15, 1996).

Labels: use any text. Formulas: use variable names or v1, v2, ..., v0 is case #.
Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after;)

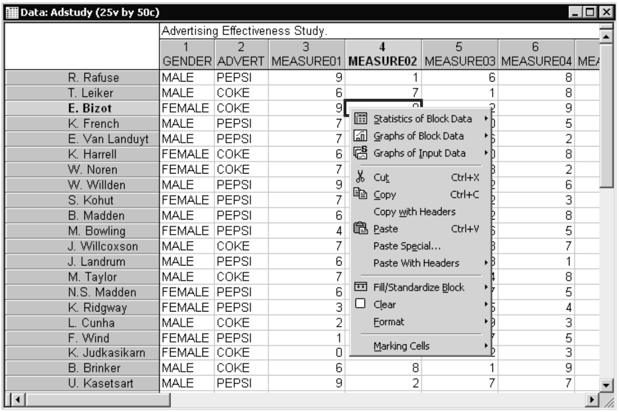
Spreadsheet formulas. Using the options in this dialog, you can change the variable name and/or format, enter a formula to recalculate the values of the variable, etc. If the entry in the ***Long name (label or formula with Functions)*** box starts with an equal sign (**=**), *STATISTICA* interprets it as a formula [a comment can follow after a semicolon (;)]. For example, if you enter into the ***Long name…*** box (of variable one) *=(v2+v3+v4)/3* or *=mean(v2:v4)*, the current values of that variable will be replaced by the average of variables two through four, separately for each case (row) of the spreadsheet.

Specifications of all variables can also be reviewed and edited together in a "combined" ***Variable Specifications Editor*** dialog, accessed by clicking the ***All Specs*** button in the ***Variable*** specifications dialog.



	Name	Type	MD code	Length	Long Name (label or formula)	Measurement Type	Excluded	Label
1	GENDER	Double	-9999		Gender of the su	Auto	□	□
2	ADVERT	Double	-9999		Ad shown to the	Auto	□	□
3	MEASURE01	Double	-9999			Auto	□	□
4	MEASURE02	Double	-9999			Auto	□	□
5	MEASURE03	Double	-9999			Auto	□	□
6	MEASURE04	Double	-9999			Auto	□	□

Shortcut menus accessed from spreadsheets. A useful feature of the spreadsheet is the list of commands available from its shortcut menus. Shortcut menus are dynamic menus that are displayed by right-clicking on an item (e.g., a cell in the spreadsheet, as shown in the illustration below). The spreadsheet shortcut menus include a selection of specific data management operations and other options related to the currently selected variable (column), case (row), block of cells, or other item.



Data: Adstudy (25v by 50c)

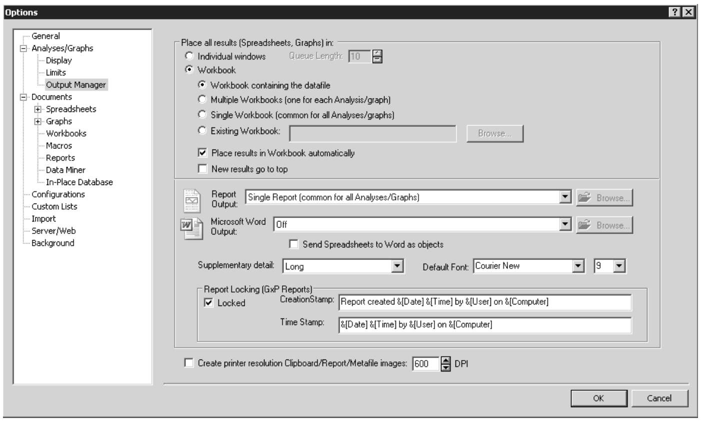
Advertising Effectiveness Study.

	1 GENDER	2 ADVERT	3 MEASURE01	4 MEASURE02	5 MEASURE03	6 MEASURE04	MEA
R. Rafuse	MALE	PEPSI	9	1	6	8	
T. Leiker	MALE	COKE	6	7	1	8	
E. Bizot	FEMALE	COKE	9			9	
K. French	MALE	PEPSI	7			5	
E. Van Landuyt	MALE	PEPSI	7			2	
K. Harrell	FEMALE	COKE	6			8	
W. Noren	FEMALE	COKE	7			2	
W. Willden	MALE	PEPSI	9			6	
S. Kohut	FEMALE	PEPSI	7			3	
B. Madden	MALE	PEPSI	6			8	
M. Bowling	FEMALE	PEPSI	4			5	
J. Willcoxson	MALE	COKE	7			7	
J. Landrum	MALE	PEPSI	6			1	
M. Taylor	MALE	COKE	7			8	
N.S. Madden	FEMALE	PEPSI	6			5	
K. Ridgway	FEMALE	PEPSI	3			4	
L. Cunha	MALE	COKE	2			3	
F. Wind	FEMALE	PEPSI	1			5	
K. Judkasikarn	FEMALE	COKE	0			3	
B. Brinker	MALE	COKE	6	8	1	9	
U. Kasetsart	MALE	PEPSI	9	2	7	7	

Six ways of handling output. You can customize the way output is managed in *STATISTICA* (see *Five Channels for Output from Analyses*, page 147). You can direct all output to five basic channels:

- Workbooks, see page 148,
- Stand-alone windows, see page 150,
- Reports, see page 151,
- Microsoft Word, see page 154,
- The Web, see page 155, and
- SharePoint or *STATISTICA Document Management System (SDMS)*, see page 163

The first four output channels listed above are controlled by the options in the ***Output Manager*** options pane of the ***Options*** dialog [accessible by selecting the ***Tools*** tab and clicking ***Options***; in the ***Options*** dialog, select ***Output Manager*** in the tree view (the left pane) to view related specifications in the options pane (the right pane)]. SharePoint options are located on the ***Home*** tab in the ***SharePoint*** group. *STATISTICA Document Management System (SDMS*), a complete solution for managing documents, is available from StatSoft. See *Appendix C – STATISTICA Family of Products* for more information.



Options

Place all results (Spreadsheets, Graphs) in:
○ Individual windows Queue Length: 10
● Workbook
● Workbook containing the datafile
○ Multiple Workbooks (one for each Analysis/graph)
○ Single Workbook (common for all Analyses/graphs)
○ Existing Workbook: Browse...
☑ Place results in Workbook automatically
☐ New results go to top

Report Output: Single Report (common for all Analyses/Graphs) Browse...
Microsoft Word Output: Off Browse...
☐ Send Spreadsheets to Word as objects

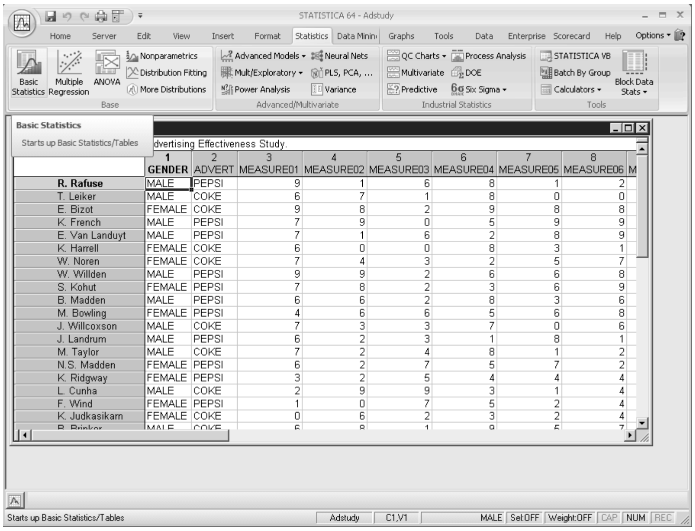
Supplementary detail: Long Default Font: Courier New 9

Report Locking (GxP Reports)
☑ Locked CreationStamp: Report created &[Date] &[Time] by &[User] on &[Computer]
Time Stamp: &[Date] &[Time] by &[User] on &[Computer]

☐ Create printer resolution Clipboard/Report/Metafile images: 600 DPI

There are a number of ways to output to the Web, depending on the version of *STATISTICA* you have. These means for output can be used in many combinations (e.g., a workbook and report simultaneously), and each output channel can be customized in a variety of ways. Also, all output objects (spreadsheets and graphs) can contain other embedded and linked objects and documents, so *STATISTICA* output can be hierarchically organized in a variety of ways.

Calculating a correlation matrix. Now, let's compute a correlation matrix for the variables in the *Adstudy.sta* data file. To display the ***Basic Statistics and Tables*** Startup Panel, select the ***Statistics*** tab, and in the ***Base*** group, click ***Basic Statistics***,



Basic Statistics

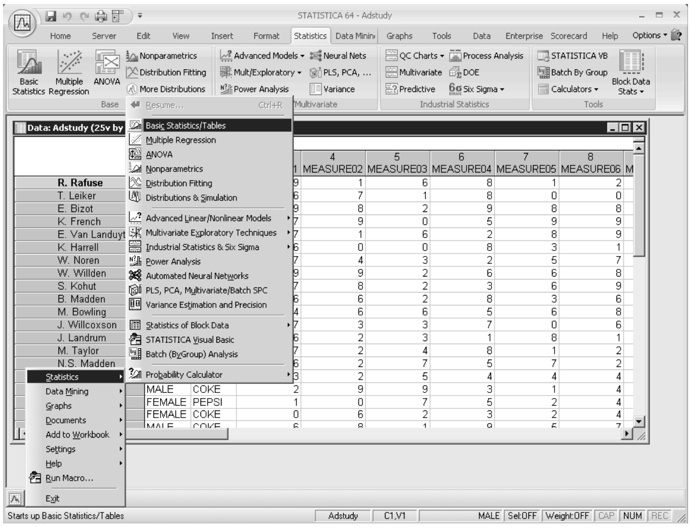
Starts up Basic Statistics/Tables

dvertising Effectiveness Study.

	1 GENDER	2 ADVERT	3 MEASURE01	4 MEASURE02	5 MEASURE03	6 MEASURE04	7 MEASURE05	8 MEASURE06	M
R. Rafuse	MALE	PEPSI	9	1	6	8	1	2	
T. Leiker	MALE	COKE	6	7	1	8	0	0	
E. Bizot	FEMALE	COKE	9	8	2	9	8	8	
K. French	MALE	PEPSI	7	9	0	5	9	9	
E. Van Landuyt	MALE	PEPSI	7	1	6	2	8	9	
K. Harrell	FEMALE	COKE	6	0	0	8	3	1	
W. Noren	FEMALE	COKE	7	4	3	2	5	7	
W. Willden	MALE	PEPSI	9	9	2	6	6	8	
S. Kohut	FEMALE	PEPSI	7	8	2	3	6	9	
B. Madden	MALE	PEPSI	6	6	2	8	3	6	
M. Bowling	FEMALE	PEPSI	4	6	6	5	6	8	
J. Willcoxson	MALE	COKE	7	3	3	7	0	6	
J. Landrum	MALE	PEPSI	6	2	3	1	8	1	
M. Taylor	MALE	COKE	7	2	4	8	1	2	
N.S. Madden	FEMALE	PEPSI	6	2	7	5	7	2	
K. Ridgway	FEMALE	PEPSI	3	2	5	4	4	4	
L. Cunha	MALE	COKE	2	9	9	3	1	4	
F. Wind	FEMALE	PEPSI	1	0	7	5	2	4	
K. Judkasikarn	FEMALE	COKE	0	6	2	3	2	4	
R. Brinker	MALE	COKE	6	8	1	9	5	7	

Starts up Basic Statistics/Tables

Adstudy | C1,V1 | MALE | Sel:OFF | Weight:OFF | CAP | NUM | REC or select ***Statistics - Basic Statistics/Tables*** from the *STATISTICA* Start menu in the lower-left corner of the screen.

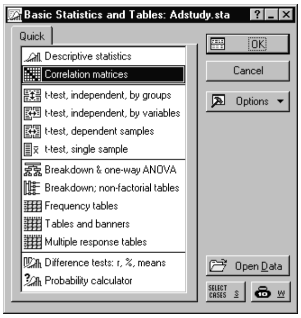


				4 MEASURE02	5 MEASURE03	6 MEASURE04	7 MEASURE05	8 MEASURE06
R. Rafuse			9	1	6	8	1	2
T. Leiker			6	7	1	8	0	0
E. Bizot			9	8	2	9	8	8
K. French			7	9	0	5	9	9
E. Van Landuyt			7	1	6	2	8	9
K. Harrell			6	0	0	8	3	1
W. Noren			7	4	3	2	5	7
W. Willden			9	9	2	6	6	8
S. Kohut			7	8	2	3	6	9
B. Madden			6	6	2	8	3	6
M. Bowling			4	6	6	5	6	8
J. Willcoxson			7	3	3	7	0	6
J. Landrum			6	2	3	1	8	1
M. Taylor			7	2	4	8	1	2
N.S. Madden			6	2	7	5	7	2
			3	2	5	4	4	4
	MALE	COKE	2	9	9	3	1	4
	FEMALE	PEPSI	1	0	7	5	2	4
	FEMALE	COKE	0	6	2	3	2	4

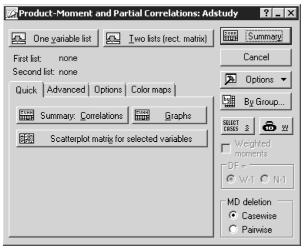
Starts up Basic Statistics/Tables

At this point, ensure that a block (a group of selected cells) is *not* selected in the spreadsheet. To deselect a block, click in any cell in the spreadsheet. If a block is selected, *STATISTICA* assumes that the variables corresponding to the block are intentionally preselected for the analysis, and when you later click the ***OK*** or ***Summary*** button to produce the analysis results, instead of prompting you to select variables, *STATISTICA* will automatically produce the correlations for the selected block variables.

In the ***Basic Statistics and Tables*** Startup Panel (shown in the next illustration),


Basic Statistics and Tables: Adstudy.sta
Quick
Descriptive statistics
Correlation matrices
t-test, independent, by groups
t-test, independent, by variables
t-test, dependent samples
t-test, single sample
Breakdown & one-way ANOVA
Breakdown; non-factorial tables
Frequency tables
Tables and banners
Multiple response tables
Difference tests: r, %, means
Probability calculator
OK
Cancel
Options
Open Data
SELECT CASES
S
W

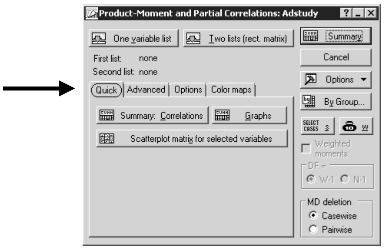
select ***Correlation matrices*** and click the ***OK*** button (or double-click ***Correlation matrices***) to display ***the Product-Moment and Partial Correlations*** dialog.



First list: none
Second list: none
□ Weighted moments
DF = ● W-1 ○ N-1
MD deletion ● Casewise ○ Pairwise

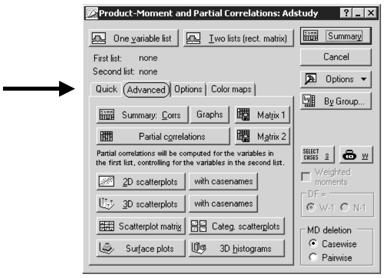
Quick vs. advanced analyses. As with most analysis specification dialogs (and other types of *STATISTICA* dialogs), the ***Product-Moment and Partial Correlations*** dialog is organized by tabs according to the type of options available. Typically, at least two categories of options are available.

The ***Quick*** tab of a dialog contains the most commonly used options, enabling you to quickly specify a basic analysis without having to search through numerous options.



First list: none
Second list: none
□ Weighted moments
DF = ● W-1 ○ N-1
MD deletion ● Casewise ○ Pairwise

The ***Advanced*** tab typically contains the same options available on the ***Quick*** tab as well as a variety of less commonly used options (e.g., in this case, options to save matrices, produce less commonly requested statistics, and create a variety of plots). Additional tabs are often available as well, depending on the type of analysis being specified.



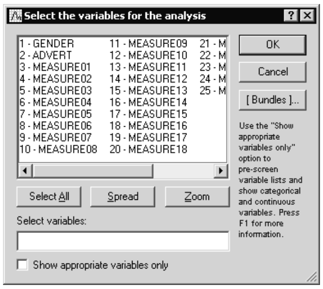
First list: none
Second list: none
Partial correlations will be computed for the variables in the first list, controlling for the variables in the second list.
□ Weighted moments
DF = ● W-1 ○ N-1
MD deletion ● Casewise ○ Pairwise

Note that in some cases, only a ***Quick*** tab is available. As with all dialogs in *STATISTICA*, you can press F1 on your keyboard or click the ? button in the upper-right corner to display a Help topic containing information about the options available on the currently selected tab.

The "self-prompting" nature of *STATISTICA* dialogs. All dialogs in *STATISTICA* follow the "self-prompting" dialog convention, which means that whenever you are not sure what to select next, simply click the ***OK*** button or the ***Summary*** button and *STATISTICA* will proceed to the next logical step, prompting you for the specific input needed (e.g., variables to be analyzed).

Variables* button.** Every analysis specification dialog in *STATISTICA* contains one or more ***Variables buttons used to display the variable selection dialog to specify variables to be analyzed.

Variable selection dialog. For this example, click the ***One variable list*** button (or press ALT+V on your keyboard) to display the ***Select the variables for the analysis*** dialog. Note that the variable selection dialog is also displayed if you click the ***Summary*** button before variables are selected. (As mentioned previously, if a block of variables is selected in the data file, those variables will be specified automatically for the analysis, and when you click the ***Summary*** button, a correlation matrix will be produced for the variables selected in the block, not all variables in the data file.)



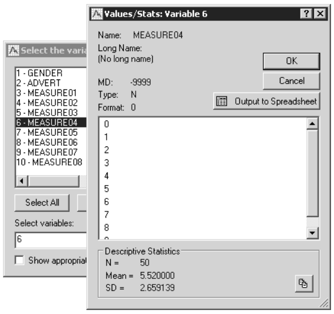
```
1 - GENDER       11 - MEASURE09   21 - M
2 - ADVERT       12 - MEASURE10   22 - M
3 - MEASURE01    13 - MEASURE11   23 - M
4 - MEASURE02    14 - MEASURE12   24 - M
5 - MEASURE03    15 - MEASURE13   25 - M
6 - MEASURE04    16 - MEASURE14
7 - MEASURE05    17 - MEASURE15
8 - MEASURE06    18 - MEASURE16
9 - MEASURE07    19 - MEASURE17
10 - MEASURE08   20 - MEASURE18
```

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

□ Show appropriate variables only

The variable selection dialog supports various ways of selecting variables (including the standard Windows SHIFT+click and CTRL+click conventions to select ranges and discontinuous lists of variables).

You can also use various shortcuts and options in the variable selection dialog to review the contents of the data file. For example, you can spread the variable list to review the variables' long names or formulas (click the ***Spread*** button), or you can zoom in on a variable (click the ***Zoom*** button) to review a sorted list of all values and descriptive statistics for the selected variable (see the next illustration).



Values/Stats: Variable 6

Name: MEASURE04
Long Name:
(No long name)

MD: -9999
Type: N
Format: 0

0
1
2
3
4
5
6
7
8

Descriptive Statistics
N = 50
Mean = 5.520000
SD = 2.659139

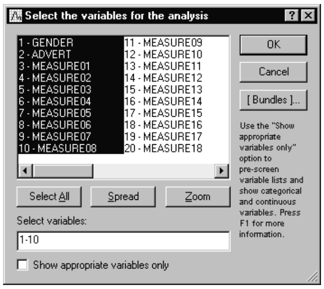
Select the varia

1 - GENDER
2 - ADVERT
3 - MEASURE01
4 - MEASURE02
5 - MEASURE03
6 - MEASURE04
7 - MEASURE05
8 - MEASURE06
9 - MEASURE07
10 - MEASURE08

Select variables: 6

□ Show appropria

For this example, select variables *1* through *10* in the variable selection dialog.



Select the variables for the analysis

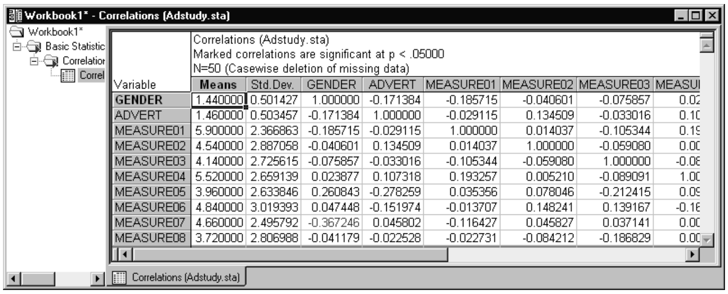
1 - GENDER
2 - ADVERT
3 - MEASURE01
4 - MEASURE02
5 - MEASURE03
6 - MEASURE04
7 - MEASURE05
8 - MEASURE06
9 - MEASURE07
10 - MEASURE08
11 - MEASURE09
12 - MEASURE10
13 - MEASURE11
14 - MEASURE12
15 - MEASURE13
16 - MEASURE14
17 - MEASURE15
18 - MEASURE16
19 - MEASURE17
20 - MEASURE18

Select variables: 1-10

□ Show appropriate variables only

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

Click the ***OK*** button. A message will be displayed informing you that there are text variables selected. Click the ***Continue with current selection*** button to return to the ***Product-Moment and Partial Correlations*** dialog. Next, click the ***Summary*** button to generate a correlation matrix for the selected variables.



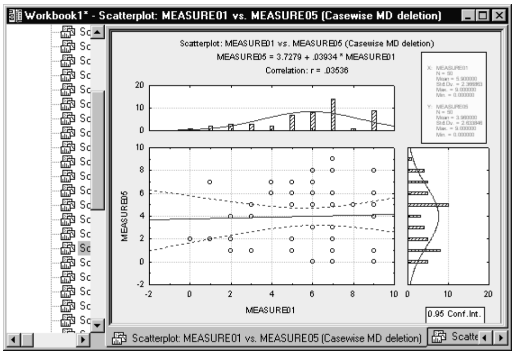
Workbook1* - Correlations (Adstudy.sta)

Correlations (Adstudy.sta)
Marked correlations are significant at p < .05000
N=50 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	GENDER	ADVERT	MEASURE01	MEASURE02	MEASURE03	MEASU
GENDER	1.440000	0.501427	1.000000	-0.171384	-0.185715	-0.040601	-0.075857	0.02
ADVERT	1.460000	0.503457	-0.171384	1.000000	-0.029115	0.134509	-0.033016	0.10
MEASURE01	5.900000	2.366863	-0.185715	-0.029115	1.000000	0.014037	-0.105344	0.19
MEASURE02	4.540000	2.887058	-0.040601	0.134509	0.014037	1.000000	-0.059080	0.00
MEASURE03	4.140000	2.725615	-0.075857	-0.033016	-0.105344	-0.059080	1.000000	-0.08
MEASURE04	5.520000	2.659139	0.023877	0.107318	0.193257	0.005210	-0.089091	1.00
MEASURE05	3.960000	2.633846	0.260843	-0.278259	0.035356	0.078046	-0.212415	0.09
MEASURE06	4.840000	3.019393	0.047448	-0.151974	-0.013707	0.148241	0.139167	-0.16
MEASURE07	4.660000	2.495792	-0.367246	0.045802	-0.116427	0.045827	0.037141	0.00
MEASURE08	3.720000	2.806988	-0.041179	-0.022528	-0.022731	-0.084212	-0.186829	0.00

Note that instead of clicking the ***Summary*** button, you could have clicked the ***Summary: Correlations*** button on the ***Quick*** tab or on the ***Advanced*** tab with the same results. Also, depending on the defaults you have specified for handling output (in the ***Output Manager*** options pane of the ***Options*** dialog), the *Correlations* spreadsheet can be displayed in a report or a stand-alone window or sent to a Word document, rather than in a workbook as shown above.

Summary graphs. *STATISTICA* provides extremely flexible tools and methods for summarizing key results in graphs and/or tables. For example, resume the analysis by clicking the ***Product-Moment and…*** button on the Analysis bar in the lower-left corner of the screen or by pressing CTRL+R on your keyboard, and click the Graphs button to display summary graphs for each pair of variables in the correlation matrix.



Workbook1* - Scatterplot: MEASURE01 vs. MEASURE05 (Casewise MD deletion)

Scatterplot: MEASURE01 vs. MEASURE05 (Casewise MD deletion)
MEASURE05 = 3.7279 + .03934 * MEASURE01
Correlation: r = .03536

X: MEASURE01
N = 50
Mean = 5.900000
Std.Dv. = 2.366963
Max. = 9.000000
Min. = 0.000000

Y: MEASURE05
N = 50
Mean = 3.980000
Std.Dv. = 2.633846
Max. = 9.000000
Min. = 0.000000

These graphs not only show the scatterplot of points for each correlation, but also the distributions (histograms) for each variable, as well as the respective correlation coefficient and regression equation.

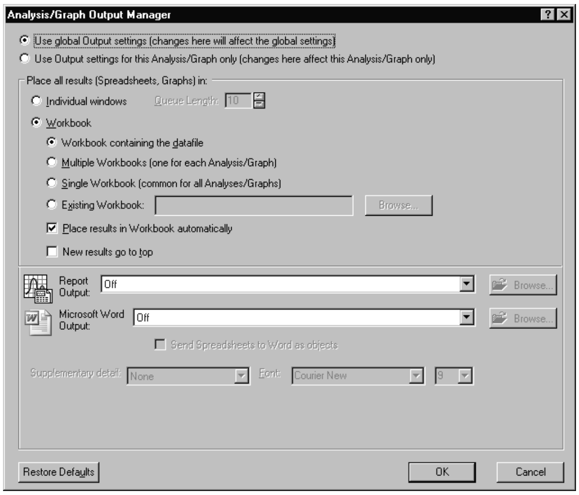
STATISTICA incorporates many such displays to summarize basic descriptive statistics, correlations, the results of Gage or Process capability studies, or other types of data analyses.

Results spreadsheets (multimedia tables). In addition to storing data, spreadsheets are used in *STATISTICA* to display most of the numeric output. Note that spreadsheets offer many display features and options, and in this example, significant correlations are marked with a different format to help distinguish them; by default, the color is red (in the *Correlations* spreadsheet, see the cell adjacent to *MEASURE07* under *GENDER*). Spreadsheets can hold anywhere from a short line to gigabytes of output, and they offer a variety of options to facilitate reviewing the results and visualizing them in predefined and custom-defined graphs, as will be seen later in this example. Also, as mentioned previously, *STATISTICA* Spreadsheets are managed using StatSoft's powerful multimedia table technology. They can handle not only virtually unlimited amounts of data, but also video, sound, custom user interfaces, and auto-executing scripts, as well as offer virtually unlimited customization options (see page 173 for further details on spreadsheets).

Spreadsheet options. Most spreadsheet facilities are accessible via options on the ***Data*** tab and the shortcut menus (displayed by right-clicking in the spreadsheet). You can try these options to see how they work, or you can review their descriptions by pressing the Help key (F1). You can change all aspects of the display formats for each spreadsheet column, edit the output, or append blank cases and variables to make room for notes or output pasted from other sources. Spreadsheets can be printed in a variety of ways (by default, in presentation-quality tables with grid lines). Also, since spreadsheets are used for input, you can easily specify an analysis using the results from a previous analysis (for example, you could use this correlation matrix to specify a multidimensional scaling analysis). To use a results spreadsheet as an input spreadsheet, select the ***Input*** check box (located on the ***Data*** tab in the ***Mode*** group) when that spreadsheet is active.

Analysis workbooks and other output options. All results can be displayed (and stored) in stand-alone windows, reports, Word documents, or workbooks, which represent the default (and perhaps the most versatile) way of handling output from analyses (see page 148 and page 169 for further details on workbooks). Depending on your selections in the ***Output Manager*** (accessible by selecting the ***Home*** tab and clicking ***Options*** in the ***Tools*** group, and then selecting ***Output Manager***, located under ***Analyses/Graphs***), results can be put in a single workbook that holds the results from all analyses, a separate analysis workbook that holds the results (spreadsheets and graphs) from a single analysis, the workbook that contains the original data file, or a preexisting workbook. Additionally, you can choose to have the results sent to a workbook automatically, or you can send them to the workbook yourself by clicking ***Add to Workbook*** on the ***Home*** tab in the ***Output*** group to send selected stand-alone spreadsheets or graphs to a workbook.

Output Manager. Which type of workbook you choose, or whether you choose to use a workbook, depends entirely on how you prefer to store your data and results. To change the output destination for the results of a particular analysis only, click the Options ▾ button on any analysis or graph specification dialog, and select ***Output*** to display the ***Analysis/Graph Output Manager*** dialog.



Analysis/Graph Output Manager

● Use global Output settings (changes here will affect the global settings)
○ Use Output settings for this Analysis/Graph only (changes here affect this Analysis/Graph only)

Place all results (Spreadsheets, Graphs) in:
○ Individual windows Queue Length: 10
● Workbook
● Workbook containing the datafile
○ Multiple Workbooks (one for each Analysis/Graph)
○ Single Workbook (common for all Analyses/Graphs)
○ Existing Workbook:
☑ Place results in Workbook automatically
□ New results go to top

Report Output: Off
Microsoft Word Output: Off
□ Send Spreadsheets to Word as objects
Supplementary detail: None Font: Courier New 9

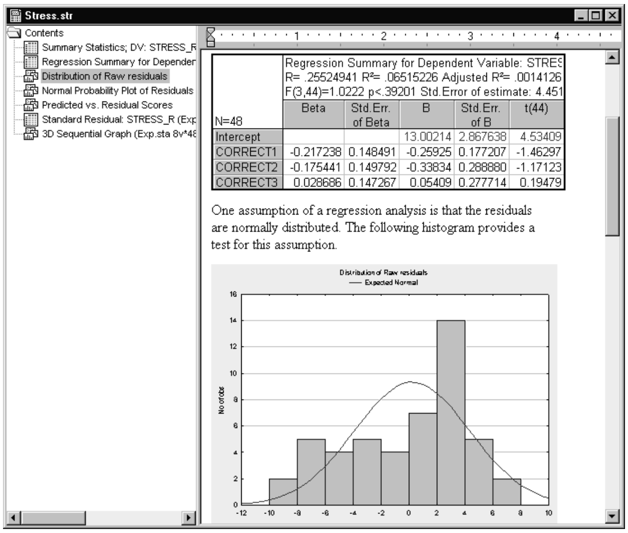
To change output options for all analyses, use the (global) ***Output Manager*** (the ***Output Manager*** options pane of the ***Options*** dialog, accessible by selecting the ***Home*** tab and clicking ***Options*** in the ***Tools*** group), or select the ***Use global Output settings (changes here will affect the global settings)*** option button in the ***Analysis/Graph Output Manager*** dialog.

As with all workbooks, individual documents (e.g., spreadsheets or graphs) or groups of documents can be printed, extracted, copied, and deleted from an analysis workbook. See the overview of *Workbooks* on page 169 for more details; see also the *Electronic Manual* (*STATISTICA* Help).

Copy vs. Copy with Headers. Contents of spreadsheets can be copied to the Clipboard by pressing CTRL+C (which copies only the contents of the selected block). To copy the block along with its respective variable and case names, select the ***Edit*** tab, and in the ***Clipboard/Data*** group, click the ***Copy*** arrow and select ***Copy with Headers*** from the drop-down menu. When spreadsheets are pasted into a word processor document, they will be active (in-place editable) *STATISTICA* objects, standard RTF-formatted tables, unformatted text, pictures, or HTML (depending on your choice in the *Paste Special* dialog of the word processor).

Printing spreadsheets. To produce a hard copy of an output spreadsheet, select the ***Home*** tab, and in the ***File*** group, click ***Print*** (or press CTRL+P) to display the ***Print Spreadsheet*** dialog, in which you specify printing options. You can also use the shortcut method of clicking the printer icon 🖨 located on the Quick Access toolbar in the upper-left corner of the ribbon bar. This shortcut method does not display the ***Print Spreadsheet*** dialog, but prints the entire current document. If you want to print a document from within a workbook, ensure that the document is selected in the workbook, and select the ***Selection*** option button in the ***Print Spreadsheet*** dialog. You can also extract a copy of the document from the workbook (drag it from the tree pane, or select the document and click ***Move*** on the ***Workbook*** tab in the ***Extract*** group) and then print it.

Optional reports of all output. Workbooks offer perhaps the most flexible options to manage your output (see pages 148 and 169). In some circumstances, however, it may be useful to automatically produce a log of all results (contents of all spreadsheets and/or graphs) in a traditional word processor style report format where comments and annotations can be inserted in arbitrary locations, objects can be placed side by side, etc. (see page 151 and page 180 for further details on reports).


Stress.str
Contents
Summary Statistics; DV: STRESS_R
Regression Summary for Depender
Distribution of Raw residuals
Normal Probability Plot of Residuals
Predicted vs. Residual Scores
Standard Residual: STRESS_R (Exp
3D Sequential Graph (Exp.sta 8v*48
1
2
3
4

Regression Summary for Dependent Variable: STRES
R= .25524941 R²= .06515226 Adjusted R²= .0014126
F(3,44)=1.0222 p<.39201 Std.Error of estimate: 4.451

N=48	Beta	Std.Err. of Beta	B	Std.Err. of B	t(44)
Intercept			13.00214	2.867638	4.53409
CORRECT1	-0.217238	0.148491	-0.25925	0.177207	-1.46297
CORRECT2	-0.175441	0.149792	-0.33834	0.288880	-1.17123
CORRECT3	0.028686	0.147267	0.05409	0.277714	0.19479

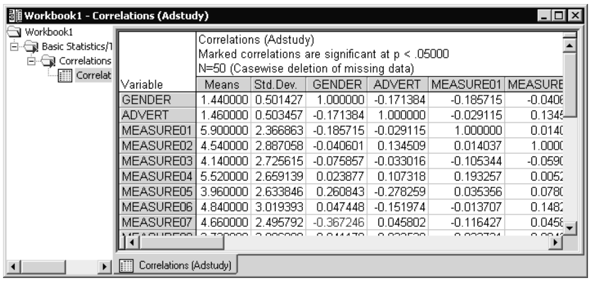
One assumption of a regression analysis is that the residuals are normally distributed. The following histogram provides a test for this assumption.

Use the options in the ***Output Manager*** to create such a report. To display the ***Output Manager***, select the ***Tools*** tab, click ***Options***, and in the ***Options*** dialog, select ***Output Manager*** located under ***Analyses/Graphs*** (for global changes). To display the ***Analysis/Graph Output Manager*** dialog, click the Options ▾ button in any analysis or graph specification dialog, and select ***Output*** (for local changes).

In the ***Output Manager*** options pane of the ***Options*** dialog or in the ***Analysis/Graph Output Manager*** dialog, click the ***Report Output*** arrow. From the drop-down menu, select either ***Send to Multiple Reports (one for each Analysis/Graph)***, ***Single Report (common for all Analyses/graphs)***, or ***[Select File]*** (which will display the ***Open*** dialog where you can select an already established report).

In the ***Output Manager***, you can also specify the amount of supplementary information to be included with the spreadsheet results. Use the ***Supplementary detail*** option to specify either ***Brief*** (includes only the selected spreadsheets and graphs), ***Medium*** (includes the selected spreadsheets and graphs as well as the current data file name, information on case selection conditions and case weights if any were specified, a list of all variables selected for each analysis, and the missing data values for each variable), ***Long*** [includes all information from the ***Medium*** format and the long variable labels (e.g., formulas), reserving one line of output (or more) for each variable], or ***Comprehensive*** (includes all information included in the ***Long*** report format as well as a complete list of all of the text labels for each selected variable).

Interpretation of the results – *STATISTICA Electronic Manual* (Help) and the Electronic Statistics Textbook. Now let's return to the example and the correlation matrix that has been produced.

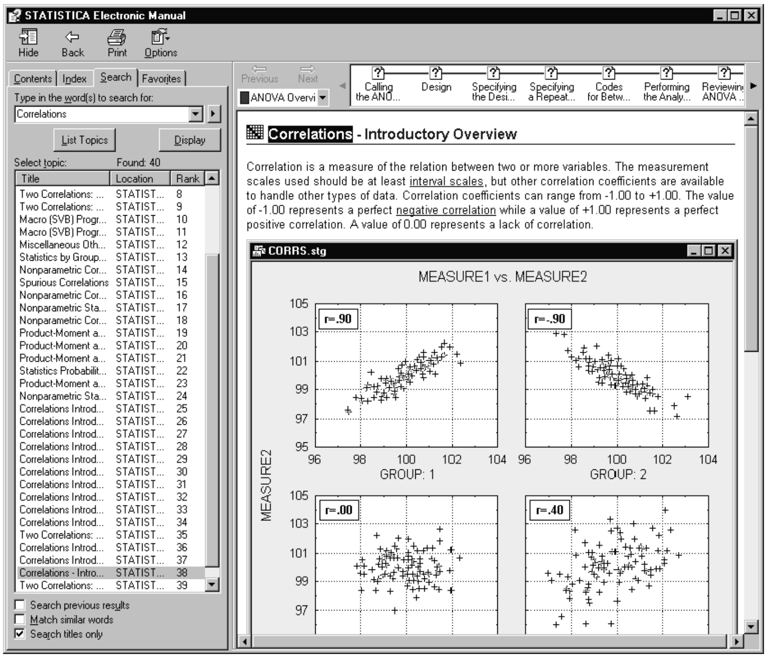


Correlations (Adstudy)
Marked correlations are significant at p < .05000
N=50 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	GENDER	ADVERT	MEASURE01	MEASURE
GENDER	1.440000	0.501427	1.000000	-0.171384	-0.185715	-0.0406
ADVERT	1.460000	0.503457	-0.171384	1.000000	-0.029115	0.1345
MEASURE01	5.900000	2.366863	-0.185715	-0.029115	1.000000	0.0140
MEASURE02	4.540000	2.887058	-0.040601	0.134509	0.014037	1.0000
MEASURE03	4.140000	2.725615	-0.075857	-0.033016	-0.105344	-0.0590
MEASURE04	5.520000	2.659139	0.023877	0.107318	0.193257	0.0052
MEASURE05	3.960000	2.633846	0.260843	-0.278259	0.035356	0.0780
MEASURE06	4.840000	3.019393	0.047448	-0.151974	-0.013707	0.1482
MEASURE07	4.660000	2.495792	-0.367246	0.045802	-0.116427	0.0458

Each of the cells of the correlation matrix represents a value (in the range of –1.00 to +1.00) that reflects the relation between the variables (see the respective variable and case headers). The higher the absolute value of the correlation coefficient, the closer the relation; if the value is positive, the relation is "positive" (high values of one variable correspond to high values of the other variable; likewise, low values of one variable correspond to low values of the other variable). If the value is negative, the opposite is true (low values of one variable correspond to high values of the other variable).

To learn more about how to interpret values of correlations, you can review a comprehensive, illustrated discussion of the topic in the *Electronic Manual* (*STATISTICA* Help), which features the complete contents of the StatSoft *Electronic Statistics Textbook*. To display the *Electronic Manual*, select the ***Help*** tab, and in the ***Help*** group, click ***Help***. On the ***Search*** tab of the *Electronic Manual*, enter the respective term (e.g., *Correlations*) into the ***Type in the word(s) to search for*** box, click ***the List Topics*** button, and then select the desired topic in the ***Select topic*** box (in this case, *Correlations - Introductory Overview*):



Type in the word(s) to search for: Correlations

Select topic: Found: 40

Title	Location	Rank
Two Correlations: ...	STATIST...	8
Two Correlations: ...	STATIST...	9
Macro (SVB) Progr...	STATIST...	10
Macro (SVB) Progr...	STATIST...	11
Miscellaneous Oth...	STATIST...	12
Statistics by Group...	STATIST...	13
Nonparametric Cor...	STATIST...	14
Spurious Correlations	STATIST...	15
Nonparametric Cor...	STATIST...	16
Nonparametric Sta...	STATIST...	17
Nonparametric Cor...	STATIST...	18
Product-Moment a...	STATIST...	19
Product-Moment a...	STATIST...	20
Product-Moment a...	STATIST...	21
Statistics Probabilit...	STATIST...	22
Product-Moment a...	STATIST...	23
Nonparametric Sta...	STATIST...	24
Correlations Introd...	STATIST...	25
Correlations Introd...	STATIST...	26
Correlations Introd...	STATIST...	27
Correlations Introd...	STATIST...	28
Correlations Introd...	STATIST...	29
Correlations Introd...	STATIST...	30
Correlations Introd...	STATIST...	31
Correlations Introd...	STATIST...	32
Correlations Introd...	STATIST...	33
Correlations Introd...	STATIST...	34
Two Correlations: ...	STATIST...	35
Correlations Introd...	STATIST...	36
Correlations Introd...	STATIST...	37
Correlations - Intro...	STATIST...	38
Two Correlations: ...	STATIST...	39

☐ Search previous results
☐ Match similar words
☑ Search titles only

Correlations - Introductory Overview

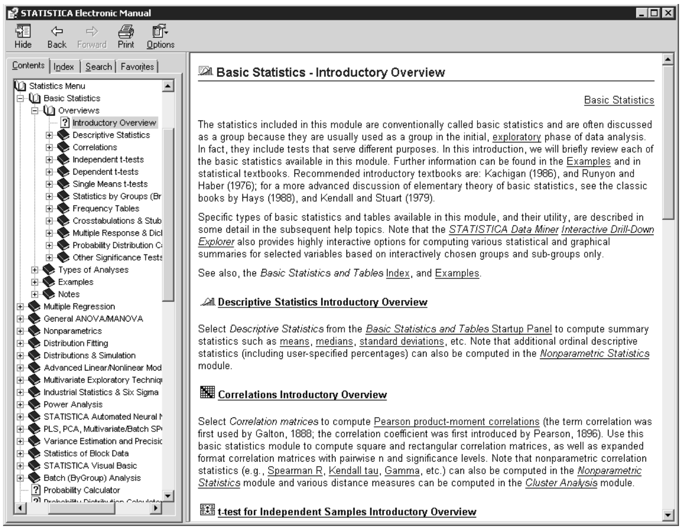
Correlation is a measure of the relation between two or more variables. The measurement scales used should be at least interval scales, but other correlation coefficients are available to handle other types of data. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of correlation.

Another valuable reference tool is the StatSoft *Electronic Statistics Textbook* (an award-wining, Web-based general resource on statistics that has been recommended by *Encyclopedia Britannica* for its “Quality, Accuracy, Presentation, and Usability”).

StatSoft web site praised for Quality, Accuracy, Presentation and Usability.

To open the textbook, select the ***Help*** tab, and in the ***Help*** group, click ***Electronic Stats Textbook***.

Also, many topics in *STATISTICA* Help contain a link to the textbook.



Basic Statistics - Introductory Overview

Basic Statistics

The statistics included in this module are conventionally called basic statistics and are often discussed as a group because they are usually used as a group in the initial, exploratory phase of data analysis. In fact, they include tests that serve different purposes. In this introduction, we will briefly review each of the basic statistics available in this module. Further information can be found in the Examples and in statistical textbooks. Recommended introductory textbooks are: Kachigan (1986), and Runyon and Haber (1976); for a more advanced discussion of elementary theory of basic statistics, see the classic books by Hays (1988), and Kendall and Stuart (1979).

Specific types of basic statistics and tables available in this module, and their utility, are described in some detail in the subsequent help topics. Note that the *STATISTICA Data Miner Interactive Drill-Down Explorer* also provides highly interactive options for computing various statistical and graphical summaries for selected variables based on interactively chosen groups and sub-groups only.

See also, the *Basic Statistics and Tables* Index, and Examples.

Descriptive Statistics Introductory Overview

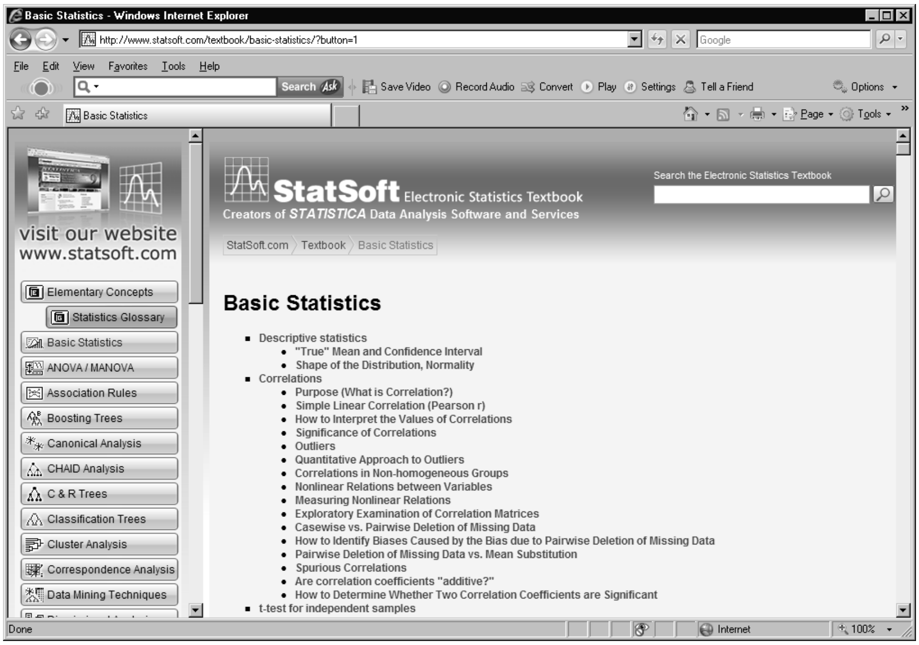
Select *Descriptive Statistics* from the *Basic Statistics and Tables* Startup Panel to compute summary statistics such as means, medians, standard deviations, etc. Note that additional ordinal descriptive statistics (including user-specified percentages) can also be computed in the *Nonparametric Statistics* module.

Correlations Introductory Overview

Select *Correlation matrices* to compute Pearson product-moment correlations (the term correlation was first used by Galton, 1888; the correlation coefficient was first introduced by Pearson, 1896). Use this basic statistics module to compute square and rectangular correlation matrices, as well as expanded format correlation matrices with pairwise n and significance levels. Note that nonparametric correlation statistics (e.g., Spearman R, Kendall tau, Gamma, etc.) can also be computed in the *Nonparametric Statistics* module and various distance measures can be computed in the *Cluster Analysis* module.

t-test for Independent Samples Introductory Overview

Click the link in the upper-right corner of the topic to display the respective page in the *Electronic Textbook*.

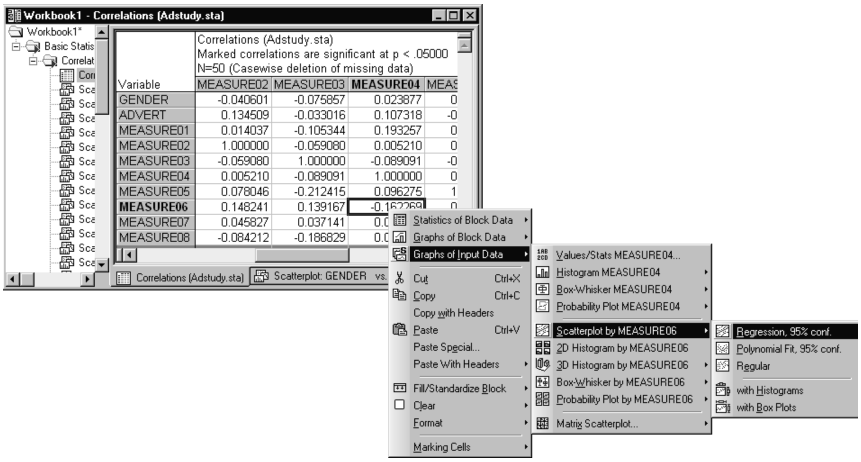


Basic Statistics

- Descriptive statistics
  - "True" Mean and Confidence Interval
  - Shape of the Distribution, Normality
- Correlations
  - Purpose (What is Correlation?)
  - Simple Linear Correlation (Pearson r)
  - How to Interpret the Values of Correlations
  - Significance of Correlations
  - Outliers
  - Quantitative Approach to Outliers
  - Correlations in Non-homogeneous Groups
  - Nonlinear Relations between Variables
  - Measuring Nonlinear Relations
  - Exploratory Examination of Correlation Matrices
  - Casewise vs. Pairwise Deletion of Missing Data
  - How to Identify Biases Caused by the Bias due to Pairwise Deletion of Missing Data
  - Pairwise Deletion of Missing Data vs. Mean Substitution
  - Spurious Correlations
  - Are correlation coefficients "additive?"
  - How to Determine Whether Two Correlation Coefficients are Significant
- t-test for independent samples

Producing graphs from spreadsheets. One of the important (and often overlooked) issues discussed in the *Electronic Manual* is the importance of scatterplots in examining correlations. For example, even very large and highly statistically significant correlation coefficients can be entirely due to one unusual data point (“outlier”), and if that is the case, then the correlation coefficient (even if statistically significant) would have no value to us (i.e., it would have no “predictive validity”). Following this concern, and the advice of the *Electronic Manual*, let’s examine a scatterplot that will visualize a relation between the variables and, thus, visualize a particular correlation coefficient from the table.

While examining the spreadsheet, you can view the correlations graphically, for example, to visualize the correlation between variables *Measure06* and *Measure04*. To produce a scatterplot for these two variables, right-click on the respective correlation coefficient *(-0.162269*). In the resulting shortcut menu, select ***Graphs of Input Data - Scatterplot by MEASURE06 - Regression, 95% conf.***, as shown in the next image.

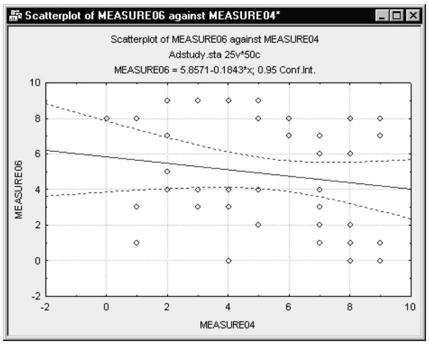


Workbook1 - Correlations (Adstudy.sta)

Correlations (Adstudy.sta)
Marked correlations are significant at p < .05000
N=50 (Casewise deletion of missing data)

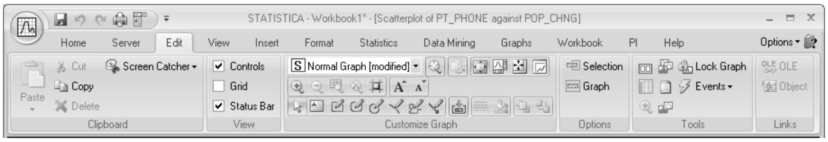
Variable	MEASURE02	MEASURE03	MEASURE04	MEAS
GENDER	-0.040601	-0.075857	0.023877	0
ADVERT	0.134509	-0.033016	0.107318	-0
MEASURE01	0.014037	-0.105344	0.193257	0
MEASURE02	1.000000	-0.059080	0.005210	0
MEASURE03	-0.059080	1.000000	-0.089091	-0
MEASURE04	0.005210	-0.089091	1.000000	0
MEASURE05	0.078046	-0.212415	0.096275	1
MEASURE06	0.148241	0.139167	-0.162269	0
MEASURE07	0.045827	0.037141	0.0	
MEASURE08	-0.084212	-0.186829	0.0	

The specified graph will be displayed.

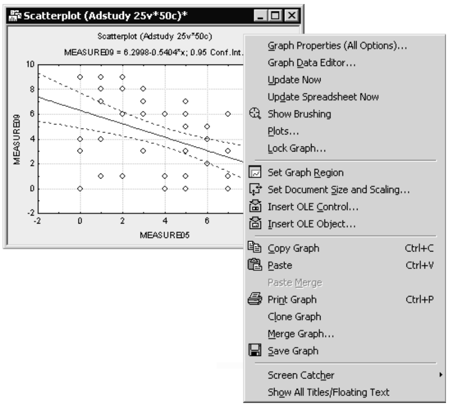


As we can learn from the graph, there are no unusual patterns of data, thus, there is no reason to be concerned about outliers (see the short discussion of outliers on page 28; see also the topic on outliers in the *Electronic Manual*).

Graph customization. Note that now, when the focus is on the graph window, the ***Edit*** tab contains different options than it did for the spreadsheets.



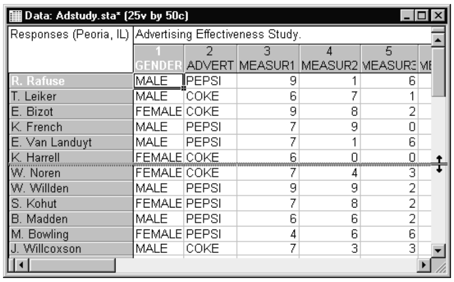
It contains a variety of graph customization and drawing tools. Many of these options are also available from shortcut menus accessed by right-clicking on specific parts of the graph. Note that the options on shortcut menus are hierarchical, meaning that the first one or two options apply specifically to the graph element you have selected, while lower options will display dialogs that offer more options on a greater variety of graph elements related to the element you have selected. If you right-click anywhere in the space outside the graph axes, a menu of global options is displayed (as shown in the next image).



For more information on graph customization, see page 190 and the *Electronic Manual*.

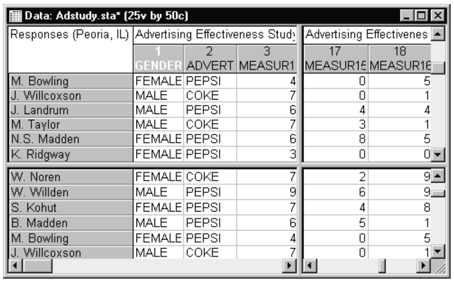
Now let’s return to the spreadsheet.

Split scrolling in spreadsheets. Spreadsheets can be split into up to four sections (panes) by dragging the split box (the small rectangle at the top of the vertical scrollbar or to the left of the horizontal scrollbar). This is useful if you have a large amount of information and you want to review results from different parts of the spreadsheet. When you move the mouse pointer to the split box, the mouse pointer changes to ⟺ or ⇳. Now, to position the split, drag it to the desired position.



Responses (Peoria, IL)	Advertising Effectiveness Study.				
	1 GENDER	2 ADVERT	3 MEASUR1	4 MEASUR2	5 MEASUR3
R. Rafuse	MALE	PEPSI	9	1	6
T. Leiker	MALE	COKE	6	7	1
E. Bizot	FEMALE	COKE	9	8	2
K. French	MALE	PEPSI	7	9	0
E. Van Landuyt	MALE	PEPSI	7	1	6
K. Harrell	FEMALE	COKE	6	0	0
W. Noren	FEMALE	COKE	7	4	3
W. Willden	MALE	PEPSI	9	9	2
S. Kohut	FEMALE	PEPSI	7	8	2
B. Madden	MALE	PEPSI	6	6	2
M. Bowling	FEMALE	PEPSI	4	6	6
J. Willcoxson	MALE	COKE	7	3	3

You can change the position of the split by dragging the split box (now located between panes) to a new position.

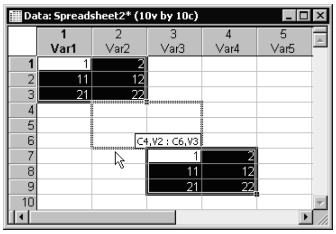


Data: Adstudy.sta* (25v by 50c)

Responses (Peoria, IL)	Advertising Effectiveness Study			Advertising Effectivenes	
	1 GENDER	2 ADVERT	3 MEASUR1	17 MEASUR15	18 MEASUR16
M. Bowling	FEMALE	PEPSI	4	0	5
J. Willcoxson	MALE	COKE	7	0	1
J. Landrum	MALE	PEPSI	6	4	4
M. Taylor	MALE	COKE	7	3	1
N.S. Madden	FEMALE	PEPSI	6	8	5
K. Ridgway	FEMALE	PEPSI	3	0	0
W. Noren	FEMALE	COKE	7	2	9
W. Willden	MALE	PEPSI	9	6	9
S. Kohut	FEMALE	PEPSI	7	4	8
B. Madden	MALE	PEPSI	6	5	1
M. Bowling	FEMALE	PEPSI	4	0	5
J. Willcoxson	MALE	COKE	7	0	1

Note that vertically split panes scroll together when you scroll horizontally; horizontally split panes scroll together when you scroll vertically. For information about highlighting blocks of data across split panes and about variable-speed highlighting of blocks of data, see *How can I expand a block in the spreadsheet outside the current screen?* in the *Electronic Manual*.

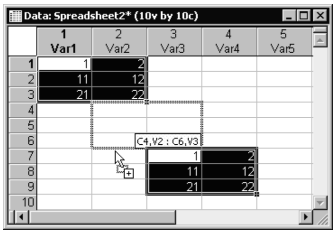
Drag-and-drop. *STATISTICA* supports the complete set of standard spreadsheet (Microsoft Excel-style) drag-and-drop facilities. For example, in order to move a block, point to the border of the selection (the mouse pointer changes to an arrow) and drag it to the new location.



Data: Spreadsheet2* (10v by 10c)

	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5
1	1	2			
2	11	12			
3	21	22			
4					
5					
6					
7			1	2	
8			11	12	
9			21	22	
10					

To copy a block of data, point to the border of the selection (the mouse pointer changes to an arrow), and drag the selection to a new location while pressing the CTRL key. Note that when you are dragging the selection, a plus sign (+) is displayed next to the mouse pointer to indicate you are copying the text rather than moving it (see the next image).



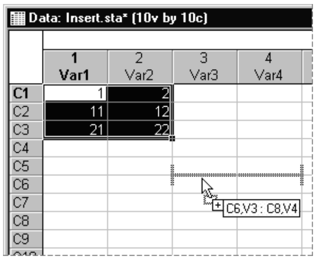
Data: Spreadsheet2* (10v by 10c)

	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5
1	1	2			
2	11	12			
3	21	22			
4					
5					
6					
7			1	2	
8			11	12	
9			21	22	
10					

To insert a block between columns or rows, point to the border of the selection (the mouse pointer changes to an arrow) and then drag the selection while pressing the SHIFT key.

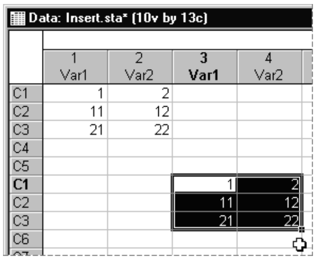
If you point between rows, an insertion bar is displayed between the rows, and when you release the mouse button, the block is inserted between those two rows [creating new case(s)]. If you point between columns, an insertion bar is displayed between the columns, and when you release the mouse button, the block is inserted between those two columns [creating new variable(s)].

Note that if you also press the CTRL key while you are dragging the selection, the block will be copied and inserted instead of moved and inserted; a plus will appear next to the mouse pointer (as shown in the next illustration).



Data: Insert.sta* (10v by 10c)

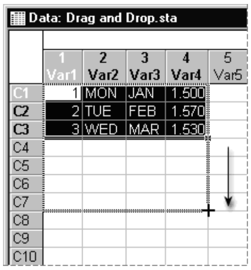
	1 Var1	2 Var2	3 Var3	4 Var4
C1	1	2		
C2	11	12		
C3	21	22		
C4				
C5				
C6				
C7				
C8				
C9				



Data: Insert.sta* (10v by 13c)

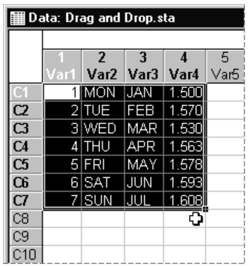
	1 Var1	2 Var2	3 Var1	4 Var2
C1	1	2		
C2	11	12		
C3	21	22		
C4				
C5				
C1			1	2
C2			11	12
C3			21	22
C6				

Additionally, a series of values within a block can be extrapolated (AutoFilled) by dragging the Fill Handle (the small, solid square located on the lower-right corner of the block border).



Data: Drag and Drop.sta

	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5
C1	1	MON	JAN	1.500	
C2	2	TUE	FEB	1.570	
C3	3	WED	MAR	1.530	
C4					
C5					
C6					
C7					
C8					
C9					
C10					



Data: Drag and Drop.sta

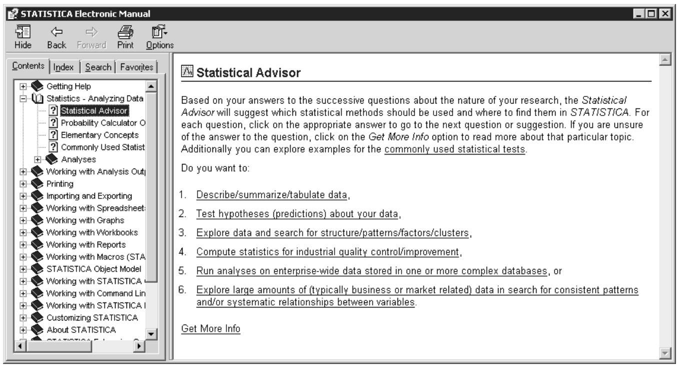
	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5
C1	1	MON	JAN	1.500	
C2	2	TUE	FEB	1.570	
C3	3	WED	MAR	1.530	
C4	4	THU	APR	1.563	
C5	5	FRI	MAY	1.578	
C6	6	SAT	JUN	1.593	
C7	7	SUN	JUL	1.608	
C8					
C9					
C10					

Electronic Manual. *STATISTICA* provides an *Electronic Manual* with comprehensive documentation on all program procedures and all options, available in a context-sensitive manner (there is a total of more than 100 megabytes of compressed documentation included). To access the manual, select the ***Help*** tab and click ***Help*** in the ***Help*** group, or click the ? icon in the upper-right corner of the ribbon bar. You can also point to (highlight) a menu command or select a tab in a dialog for which you want information, and press F1 on your keyboard to display the respective Help topic, or click the Help button ? that is on the caption bar of all dialogs.

Due to its dynamic hypertext organization, organizational tabs (***Contents***, ***Index***, ***Search***, and ***Favorites***), and various facilities used to customize the Help system, it is faster to use the *Electronic Manual* than to look for information in the traditional manuals.

Also, ToolTips display short explanations of the commands when the mouse pointer hovers over them.

Statistical Advisor. A Statistical Advisor facility is built into the *STATISTICA Electronic Manual*. On the ***Help*** tab in the ***Help*** group, click ***Statistical Advisor*** to display a set of simple questions about the nature of the research problem and the type of your data. Click on the appropriate links to answer the questions, and suggestions for the statistical procedures that appear most relevant will be displayed, containing links to guide you to the specific procedures in the *STATISTICA* system.



Statistical Advisor

Based on your answers to the successive questions about the nature of your research, the *Statistical Advisor* will suggest which statistical methods should be used and where to find them in *STATISTICA*. For each question, click on the appropriate answer to go to the next question or suggestion. If you are unsure of the answer to the question, click on the *Get More Info* option to read more about that particular topic. Additionally you can explore examples for the commonly used statistical tests.

Do you want to:

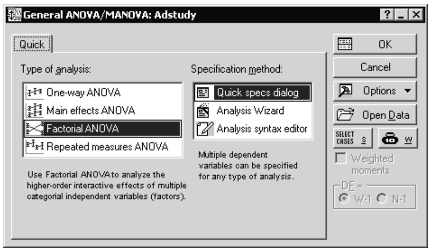
1. Describe/summarize/tabulate data,
2. Test hypotheses (predictions) about your data,
3. Explore data and search for structure/patterns/factors/clusters,
4. Compute statistics for industrial quality control/improvement,
5. Run analyses on enterprise-wide data stored in one or more complex databases, or
6. Explore large amounts of (typically business or market related) data in search for consistent patterns and/or systematic relationships between variables.

Get More Info

Direct jumps (hypertext links) in the ***Statistical Advisor*** topics guide you to corresponding *Introductory Overviews*, which discuss in detail the respective statistical methods and procedures.

Example 2: *ANOVA*

Calling the ANOVA module. For this example of a 2 x 2 (between) x 3 (repeated measures) design, open the *Adstudy.sta* data file. Then, to start the *ANOVA/MANOVA* analysis, select the ***Statistics*** tab, and in the ***Base*** group, click ***ANOVA*** to display the ***General ANOVA/MANOVA*** Startup Panel.

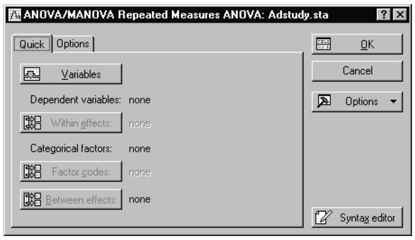


Use Factorial ANOVAto analyze the higher-order interactive effects of multiple categorial independent variables (factors).

Multiple dependent variables can be specified for any type of analysis.

This dialog is used to specify very simple analyses (e.g., via ***One-way ANOVA*** – designs with only one between-group factor) and more complex analyses (e.g., via ***Repeated measures ANOVA*** – designs with between-group factors and a within-subject factor).

Design. Select ***Repeated measures ANOVA*** as the ***Type of analysis*** and ***Quick specs dialog*** as the ***Specification method***, and then click the ***OK*** button in the ***General ANOVA/MANOVA*** Startup Panel to display the ***ANOVA/MANOVA Repeated Measures ANOVA*** dialog.



Dependent variables: none

Within effects: none

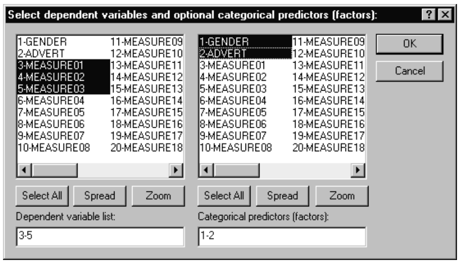
Categorical factors: none

Factor codes: none

Between effects: none

Specifying the design (variables). The first (between-group) factor is *Gender* (with 2 levels: *Male* and *Female*). The second (between-group) factor is *Advert* (with 2 levels: *Pepsi* and *Coke*). The two factors are crossed, which means that there are both *Male* and *Female* subjects in the *Pepsi* and *Coke* groups. Each of those subjects responded to three questions (this repeated measure factor will be called *Response*; it has three levels represented by variables *Measure01*, *Measure02*, and *Measure03*).

Click the ***Variables*** button (in the ***ANOVA/MANOVA Repeated Measures ANOVA*** dialog) to display the variable selection dialog. Select *Measure01* through *Measure03* as dependent variables (from the ***Dependent variable list*** field) and *Gender* and *Advert* as factors [from the ***Categorical predictors (factors)*** field].


Select dependent variables and optional categorical predictors (factors):

Dependent variable list: 3-5

Categorical predictors (factors): 1-2

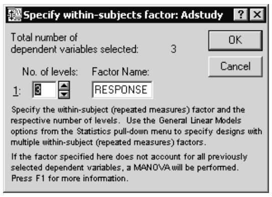
Then click the ***OK*** button to return to the ***ANOVA/MANOVA Repeated Measures ANOVA*** dialog.

The repeated measures design. The design of the experiment that we are going to analyze can be summarized as follows:

	Between-Group	Between-Group	Repeated Measure Factor: *Response*		
	Factor #1: *Gender*	Factor #2: *Advert*	Level #1: *Measure01*	Level #2: *Measure02*	Level #3: *Measure03*
Subject 1	Male	Pepsi	9	1	6
Subject 2	Male	Coke	6	7	1
Subject 3	Female	Coke	9	8	2
⋮	⋮	⋮	⋮	⋮	⋮

Specifying a repeated measures factor. The minimum necessary selections are now complete, and, if we did not want to select the repeated measures factor, we would be ready to click the ***OK*** button and see the results of the analysis. However, for our example, we need to specify that the three dependent variables we have selected be interpreted as three levels of a repeated measures (within-subject) factor. Unless we do so, *STATISTICA* assumes that those are three “different” dependent variables and runs a ***MANOVA*** (i.e., ***Multivariate ANOVA***).

In order to define the desired repeated measures factor, click the ***Within effects*** button on the ***Quick*** tab to display the ***Specify within-subjects factor*** dialog.



Specify within-subjects factor: Adstudy

Total number of dependent variables selected: 3

No. of levels: 1: 3

Factor Name: RESPONSE

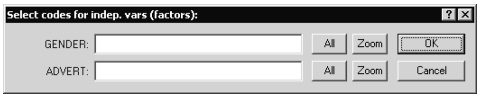
Specify the within-subject (repeated measures) factor and the respective number of levels. Use the General Linear Models options from the Statistics pull-down menu to specify designs with multiple within-subject (repeated measures) factors.

If the factor specified here does not account for all previously selected dependent variables, a MANOVA will be performed. Press F1 for more information.

Note that *STATISTICA* has suggested the selection of one repeated measures factor with *3* levels (default name *R1*). You can specify only one within-subject (repeated measures) factor via this dialog. To specify multiple within-subject factors, use the *General Linear Models* module (available in the optional *Advanced Linear/Nonlinear Models* package). Press the F1 key on your keyboard while the ***Specify within-subjects factor*** dialog is displayed (or click the ? button in the upper-right corner of the dialog) to display the *Electronic Manual* topic that describes all options in this dialog and contains links to comprehensive discussions of repeated measures and examples of designs.

For this example, edit the name for the factor: in the ***Factor Name*** box, change the default *R1* to *RESPONSE*, and click the ***OK*** button to exit the dialog.

Codes (defining the levels) for between-group factors. You do not need to manually specify codes for between-group factors [i.e., there is no need to instruct *STATISTICA* that variable *Gender* has two levels: *1* and *2* (or *Male* and *Female*)] unless you want to prevent *STATISTICA* from using, by default, all codes encountered in the selected grouping variables in the data file. To enter such custom code selection, click the ***Factor codes*** button to access the ***Select codes for indep. vars (factors)*** dialog.

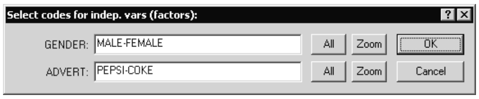


Select codes for indep. vars (factors):

GENDER:

ADVERT:

Before you make your selections, you can use the options in this dialog to review values of individual variables by clicking the ***Zoom*** button, scan the file, and fill in the codes fields (e.g., *Gender* and *Advert*) for an individual variable or all variables, etc. For now, click the ***OK*** button in the ***Select codes for indep. vars (factors)*** dialog; *STATISTICA* automatically fills in the codes fields with all distinctive values encountered in the selected variables,



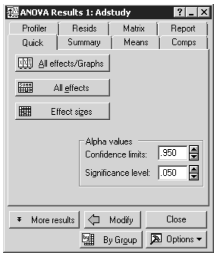
Select codes for indep. vars (factors):

GENDER: MALE-FEMALE

ADVERT: PEPSI-COKE

and closes the dialog.

Performing the analysis. Click the ***OK*** button in the ***ANOVA/MANOVA Repeated Measures ANOVA*** dialog. The analysis is performed and the ***ANOVA Results*** dialog is displayed, which contains various output spreadsheets and graphs options.



ANOVA Results 1: Adstudy

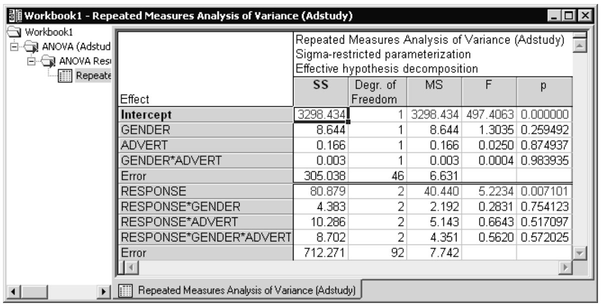
Alpha values

Confidence limits: .950

Significance level: .050

This dialog contains several tabs that enable you to quickly locate the desired results options. For example, if you want to perform planned comparisons, select the ***Comps*** tab. To view residual statistics, select the ***Resids*** tab. For this example, we will only use the results options available on the ***Quick*** tab.

Reviewing ANOVA results. Let's start by looking at the ANOVA summary of all effects table by clicking the ***All effects*** button (the one with the ***SUMM*** icon).



Workbook1 - Repeated Measures Analysis of Variance (Adstudy)

Repeated Measures Analysis of Variance (Adstudy)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	3298.434	1	3298.434	497.4063	0.000000
GENDER	8.644	1	8.644	1.3035	0.259492
ADVERT	0.166	1	0.166	0.0250	0.874937
GENDER*ADVERT	0.003	1	0.003	0.0004	0.983935
Error	305.038	46	6.631		
RESPONSE	80.879	2	40.440	5.2234	0.007101
RESPONSE*GENDER	4.383	2	2.192	0.2831	0.754123
RESPONSE*ADVERT	10.286	2	5.143	0.6643	0.517097
RESPONSE*GENDER*ADVERT	8.702	2	4.351	0.5620	0.572025
Error	712.271	92	7.742		

The only effect (ignoring the *Intercept*) in this analysis that is statistically significant (*p = .007*) is the *RESPONSE* effect. This result may be caused by many possible patterns of means of the *RESPONSE* effect (for more information, consult the

ANOVA Introductory Overview in the *Electronic Manual*). We will now look at the marginal means for this effect graphically to see what it means.

To display the ***ANOVA Results*** dialog again (that is, "resume" the analysis), press CTRL+R or click the ***ANOVA Results*** button on the analysis bar. Then, click the ***All effects/Graphs*** button to display the ***Table of All Effects*** dialog to review the means for individual effects.

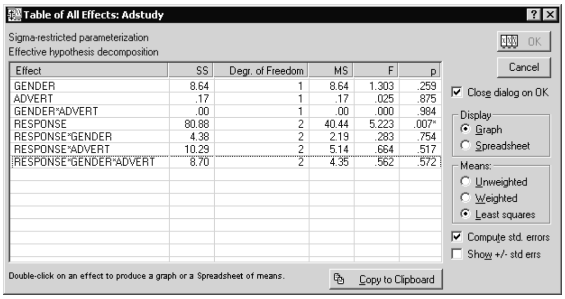


Table of All Effects: Adstudy

Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
GENDER	8.64	1	8.64	1.303	.259
ADVERT	.17	1	.17	.025	.875
GENDER*ADVERT	.00	1	.00	.000	.984
RESPONSE	80.88	2	40.44	5.223	.007*
RESPONSE*GENDER	4.38	2	2.19	.283	.754
RESPONSE*ADVERT	10.29	2	5.14	.664	.517
RESPONSE*GENDER*ADVERT	8.70	2	4.35	.562	.572

☑ Close dialog on OK

Display
● Graph
○ Spreadsheet

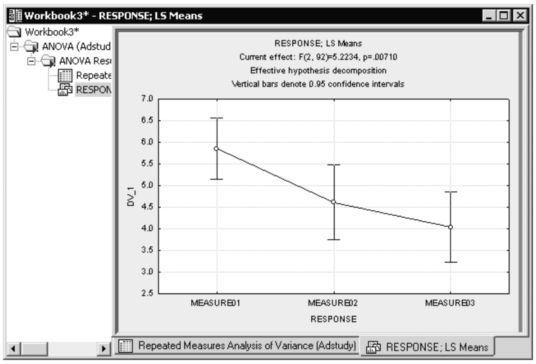
Means:
○ Unweighted
○ Weighted
● Least squares

☑ Compute std. errors
☐ Show +/- std errs

Double-click on an effect to produce a graph or a Spreadsheet of means.

This dialog contains a summary table of all effects (with most of the information you have seen in the all effects spreadsheet) and is used to review individual effects from that table in the form of the plots of the respective means (or, optionally, spreadsheets of the respective mean values).

Plot of means for a main effect. In the ***Table of All Effects*** dialog, double-click on the significant main effect *RESPONSE* (the one marked with an asterisk in the *p* column) to produce the respective plot.

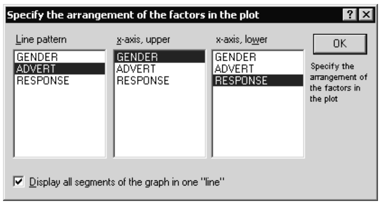


Workbook3* - RESPONSE; LS Means

RESPONSE; LS Means
Current effect: F(2, 92)=5.2234, p=.00710
Effective hypothesis decomposition
Vertical bars denote 0.95 confidence intervals

The graph indicates that there is a clear decreasing trend; the means for the consecutive three questions are gradually lower. Even though there are no significant interactions in this design (see the discussion of the *Table of all effects*, page 37), we will look at the highest-order interaction to examine the consistency of this strong decreasing trend across the between-group factors.

Plot of means for a three-way interaction. To see the plot of the highest-order interaction, in the ***Table of All Effects*** dialog, double-click on the row marked *RESPONSE*GENDER*ADVERT*, representing the interaction between factors 1 (*Gender*), 2 (*Advert*), and 3 (*Response*). An intermediate dialog, ***Specify the arrangement of the factors in the plot***, is displayed, which is used to customize the default arrangement of factors in the graph (note that, unlike the previous plot of a simple factor, the current effect can be visualized in a variety of ways).



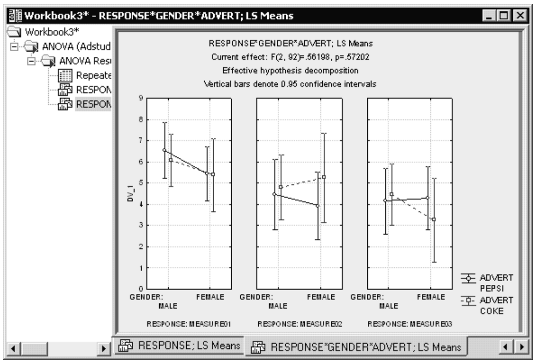
Specify the arrangement of the factors in the plot

Line pattern	x-axis, upper	x-axis, lower
GENDER	GENDER	GENDER
ADVERT	ADVERT	ADVERT
RESPONSE	RESPONSE	RESPONSE

Selected: Line pattern ADVERT; x-axis, upper GENDER; x-axis, lower RESPONSE

☑ Display all segments of the graph in one "line"

Click the ***OK*** button to accept the default arrangement and produce the plot of means.



Workbook3* - RESPONSE*GENDER*ADVERT; LS Means

RESPONSE*GENDER*ADVERT; LS Means
Current effect: F(2, 92)=.56198, p=.57202
Effective hypothesis decomposition
Vertical bars denote 0.95 confidence intervals

As you can see, this pattern of means (split by the levels of the between-group factors) does not indicate any salient deviations from the overall pattern revealed in the first plot (for the main effect, *RESPONSE*). Now you can continue to interactively examine other effects – run post-hoc comparisons, planned comparisons, extended diagnostics, etc. – to further explore the results.

Interactive data analysis in *STATISTICA*. This example illustrates the way in which *STATISTICA* supports interactive data analysis. You are not forced to specify all output to be generated before seeing any results. Even simple analysis designs can produce large amounts of output and countless graphs, but usually you cannot know what will be of interest until you have a chance to review the basic output. With *STATISTICA*, you can select specific types of output, interactively conduct follow-up tests, and run supplementary "what-if" analyses after the data are processed and basic output reviewed. *STATISTICA*'s flexible computational procedures and wide selection of options used to visualize any combination of values from numerical output offer countless methods to explore your data and verify hypotheses.

Automating analyses (macros and *STATISTICA* Visual Basic). Any selections that you make in the course of the interactive data analysis (including both specifying the designs and choosing the output options) are automatically recorded in the industry standard Visual Basic code. You can save such macros for repeated use (you can also assign them to toolbar buttons, modify or edit them, combine them with other programs, etc.). For more information, see Chapter 8 – *STATISTICA Visual Basic* on page 219 or the *STATISTICA Visual Basic Primer*.

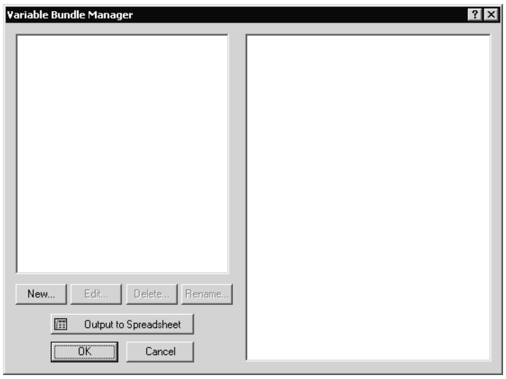
Example 3: Variable Bundles

STATISTICA offers a unique option – variable bundles – to locate a subset of data quickly and easily in a large data file. Bundles can be created to organize large sets of variables and to facilitate the repeated selection of the same set of variables.

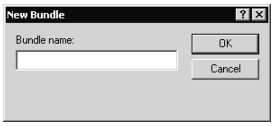
Open *EnginePerformance.sta*. This data set describes the performance of large engines and contains various process parameters recorded during their manufacture. It includes 128 engines; their *Efficiency*, *Fuel Economy*, and *Power* as measured during testing; and 74 process parameters collected during the manufacture of each engine.

For this example, we will proceed with the premise that we often need to generate analyses in which the same set of variables is repeatedly used.

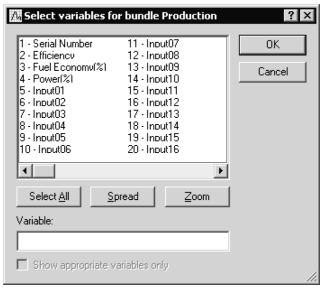
Select the ***Data*** tab, and in the ***Variables*** group, click ***Bundles*** to display the ***Variable Bundle Manager*** dialog.



Click the ***New*** button to display the ***New Bundle*** dialog,

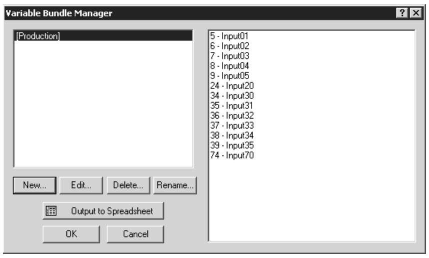


enter the name *Production* in the ***Bundle name*** field, and click the ***OK*** button. The ***Select variables for bundle*** dialog is displayed, which contains all the variables in the *EnginePerformance.sta* data set.



For our analyses, we need to select the variables *Input01-Input05*, *Input20*, *Input30-Input35*, and *Input70*. You can select these variables using the standard Windows SHIFT+click and CTRL+click conventions to select ranges and discontinuous lists of items, respectively.

Click the ***OK*** button to close the ***Select variables for bundle*** dialog and return to the ***Variable Bundle Manager***.



5 - Input01
6 - Input02
7 - Input03
8 - Input04
9 - Input05
24 - Input20
34 - Input30
35 - Input31
36 - Input32
37 - Input33
38 - Input34
39 - Input35
74 - Input70

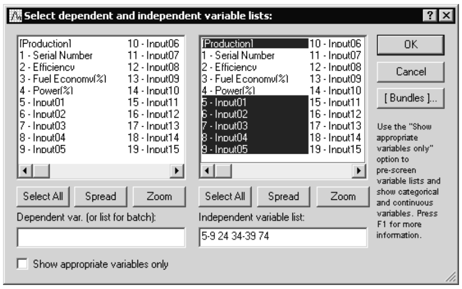
The left pane of this dialog displays the names of all bundles that have been defined for this spreadsheet (you can create numerous bundles in each spreadsheet if needed). The right pane displays the contents of the bundle that is currently selected in the left pane. If both of these panes are empty, no bundles have been created for this spreadsheet.

You can make changes to a bundle by clicking the ***Edit*** button, discard a bundle by clicking the ***Delete*** button, change the title of a bundle by clicking the ***Rename*** button, and produce a spreadsheet containing information about the bundles for the active data spreadsheet by clicking the ***Output to Spreadsheet*** button.

For this example, click the ***OK*** button to accept the bundle we created and close the ***Variable Bundle Manager*** dialog.

Then, select the ***Statistics*** tab, and in the ***Base*** group, click ***Multiple Regression*** to display the ***Multiple Linear Regression*** Startup Panel. On the ***Quick*** tab, click the ***Variables*** button to display the variable specification dialog.

Bundles are displayed in brackets and listed (in alphabetical order) at the top of the variable list. In the ***Independent variable list***, select the *Production* bundle to specify – with one click of the mouse button – *Input01-Input05*, *Input 20*, *Input 30-Input35*, and *Input 70* as the independent variables for the analysis.



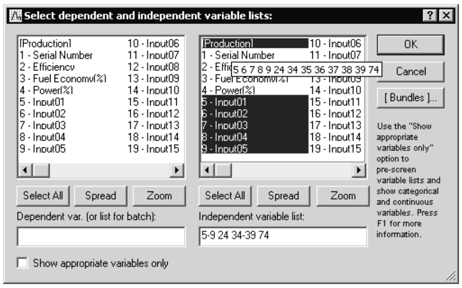
Dependent var. (or list for batch):

Independent variable list: 5-9 24 34-39 74

□ Show appropriate variables only

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

If you aren’t sure what variables are included in a bundle, move the mouse pointer over the bundle name in the variable selection dialog, and a ToolTip will display the variable numbers.



Select dependent and independent variable lists:

[Production]
1 - Serial Number
2 - Efficiency
3 - Fuel Economy(%)
4 - Power(%)
5 - Input01
6 - Input02
7 - Input03
8 - Input04
9 - Input05
10 - Input06
11 - Input07
12 - Input08
13 - Input09
14 - Input10
15 - Input11
16 - Input12
17 - Input13
18 - Input14
19 - Input15

5 6 7 8 9 24 34 35 36 37 38 39 74

Dependent var. (or list for batch):

Independent variable list: 5-9 24 34-39 74

□ Show appropriate variables only

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

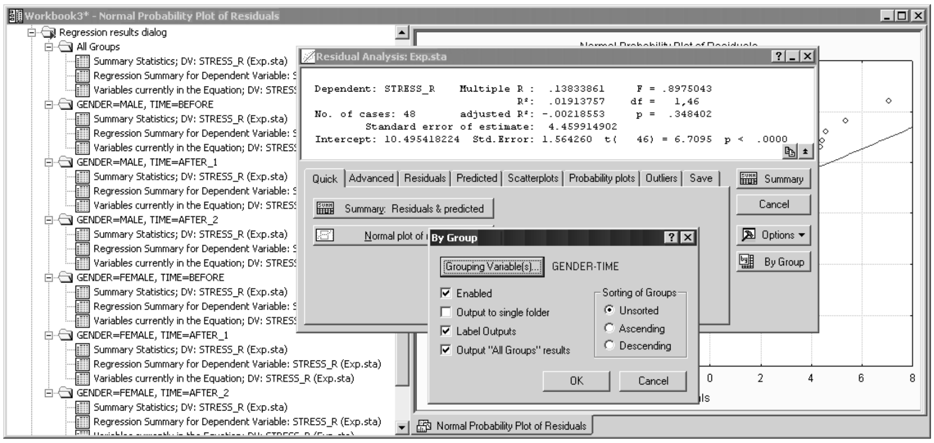
Additionally, you can view the list of variables (by name) by clicking the ***[Bundles]*** button in the variable specification dialog. This displays the ***Variable Bundles Manager***.

Note that bundles are defined for a single spreadsheet, and they are only used for variable selection. Hence, they are never listed in reports or other output.

As you can see with this example, you will save considerable time by selecting a bundle rather than looking for the correct variables to choose in a large data set.

Example 4: By-Group Analyses

STATISTICA offers a powerful option to turn every statistical or graphics analysis into an analysis by group. When reviewing results in the results dialog of practically any analysis, or using the graphs options, you can select one or more grouping variables, and then create results 1) for all cases in the data combined, and/or 2) broken down by each combination of unique values in the grouping variables.


Workbook3* - Normal Probability Plot of Residuals
Regression results dialog
All Groups
Summary Statistics; DV: STRESS_R (Exp.sta)
Regression Summary for Dependent Variable: S
Variables currently in the Equation; DV: STRESS
GENDER=MALE, TIME=BEFORE
Summary Statistics; DV: STRESS_R (Exp.sta)
Regression Summary for Dependent Variable: S
Variables currently in the Equation; DV: STRESS
GENDER=MALE, TIME=AFTER_1
Summary Statistics; DV: STRESS_R (Exp.sta)
Regression Summary for Dependent Variable: S
Variables currently in the Equation; DV: STRESS
GENDER=MALE, TIME=AFTER_2
Summary Statistics; DV: STRESS_R (Exp.sta)
Regression Summary for Dependent Variable: S
Variables currently in the Equation; DV: STRESS
GENDER=FEMALE, TIME=BEFORE
Summary Statistics; DV: STRESS_R (Exp.sta)
Regression Summary for Dependent Variable: S
Variables currently in the Equation; DV: STRESS
GENDER=FEMALE, TIME=AFTER_1
Summary Statistics; DV: STRESS_R (Exp.sta)
Regression Summary for Dependent Variable: STRESS_R (Exp.sta)
Variables currently in the Equation; DV: STRESS_R (Exp.sta)
GENDER=FEMALE, TIME=AFTER_2
Summary Statistics; DV: STRESS_R (Exp.sta)
Regression Summary for Dependent Variable: STRESS_R (Exp.sta)
Residual Analysis: Exp.sta
Quick
Advanced
Residuals
Predicted
Scatterplots
Probability plots
Outliers
Save
Summary
Cancel
Options
By Group
Summary: Residuals & predicted
Normal plot of r
0
2
4
6
8
Normal Probability Plot of Residuals

```
Dependent: STRESS_R    Multiple R :  .13833861     F = .8975043
                              R²:  .01913757    df =   1,46
No. of cases: 48       adjusted R²: -.00218553     p =  .348402
       Standard error of estimate:  4.459914902
Intercept: 10.495418224  Std.Error: 1.564260  t(   46) = 6.7095  p <  .0000
```

By Group

Grouping Variable(s)... GENDER-TIME

☑ Enabled
☐ Output to single folder
☑ Label Outputs
☑ Output "All Groups" results

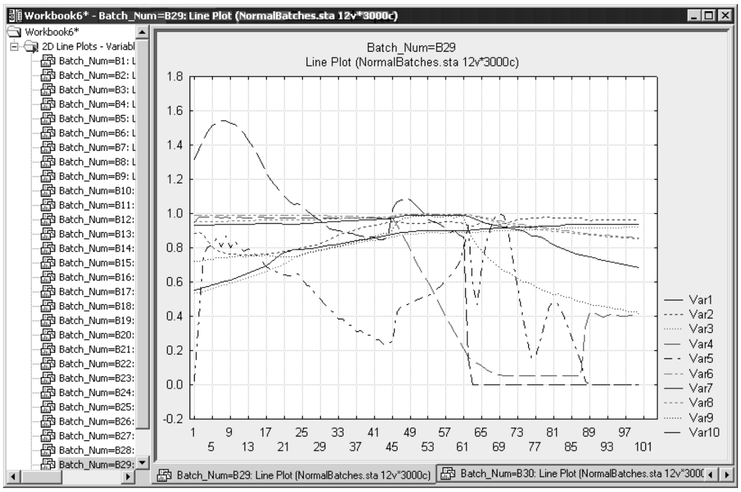
Sorting of Groups
● Unsorted
○ Ascending
○ Descending

OK Cancel

This is a very powerful tool for interactive and exploratory data analysis, allowing you to review quickly whether any patterns or specific results hold in all subgroups, samples, or strata in your data.

For example, you may be performing a multiple regression analysis and decide to review, without exiting the current dialog, the results broken down by Gender and another grouping variable in your data. After selecting (enabling) this option (by clicking the ***By Group*** button), every time you click any of the results buttons (e.g., to create a summary results spreadsheet or graph), all results are computed not only for all groups (optionally), but also for each unique combination of grouping variables that were specified (e.g., by Gender and another grouping variable).

The results of the By Group analysis can be placed either in the default results workbook into their own folder, labeled with the respective by-group condition (e.g., *Gender=Female; Time=After1*), or into the same folder with all other results.

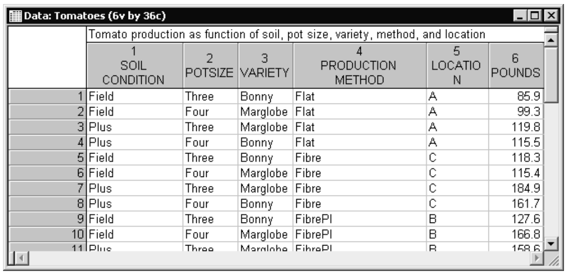


For example, you could create multiple line plots to describe a multivariate batch process, creating a separate graph (“trajectories”) for each batch.

Exploring Experimental Data Using the By Group Option

This example is based on the data file *Tomatoes.sta*, which is one of the example data files described in greater detail in the *Experimental Design* section of the *STATISTICA Electronic Manual* (see the example *Designing and Analyzing a $2^3 3^2$ Experiment*). Connor and Young (in McLean and Anderson, 1984) report an experiment (taken from Youden and Zimmerman, 1936) on various methods of producing tomato plant seedlings prior to transplanting in the field.

Start by opening the example *Tomatoes.sta* data set. Select the ***Home*** tab. In the ***File*** group, click the ***Open*** arrow and select ***Open Examples*** from the drop-down menu to display the ***Open a STATISTICA Data File*** dialog. Double-click on the *Datasets* folder, and then select and open the *STATISTICA* data set *Tomatoes.sta*.



Data: Tomatoes (6v by 36c)

Tomato production as function of soil, pot size, variety, method, and location

	1 SOIL CONDITION	2 POTSIZE	3 VARIETY	4 PRODUCTION METHOD	5 LOCATION	6 POUNDS
1	Field	Three	Bonny	Flat	A	85.9
2	Field	Four	Marglobe	Flat	A	99.3
3	Plus	Three	Marglobe	Flat	A	119.8
4	Plus	Four	Bonny	Flat	A	115.5
5	Field	Three	Bonny	Fibre	C	118.3
6	Field	Four	Marglobe	Fibre	C	115.4
7	Plus	Three	Marglobe	Fibre	C	184.9
8	Plus	Four	Bonny	Fibre	C	161.7
9	Field	Three	Bonny	FibrePl	B	127.6
10	Field	Four	Marglobe	FibrePl	B	166.8
11	Plus	Three	Marglobe	FibrePl	B	158.6

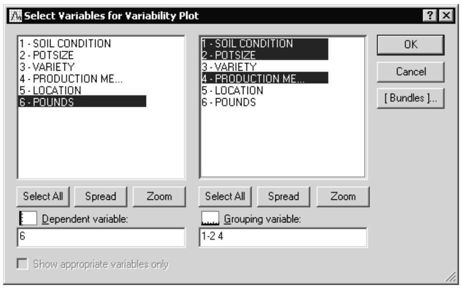
Shown here are a few rows (cases) of that data file. You can refer to the *Experimental Design Electronic Help* example topic for a complete analysis of these data.

Exploring Patterns by Variety

This example illustrates a typical "workflow" as it often applies to the analysis of discrete or batch-manufacturing data, i.e., the goal of the analysis is to verify (graphically or analytically) that some patterns or distributions equally apply to all samples, parts, or batches.

We will explore the effect of *Production Method*, *Soil Condition*, and *Potsize* on yield (*Pounds*), and evaluate whether any patterns hold for each *Variety* in the study. Instead of performing a complete analysis of variance (as is described in the *Experimental Design* example of the *Electronic Help*), we will use mostly graphical methods and visual inspection.

Specifying variability plots. Select the ***Graphs*** tab. In the ***More*** group, click ***2D***, and from the drop-down menu, select ***Variability Plots*** to display the ***Variability Plot*** dialog. Click the ***Variables*** button, and in the ***Select Variables for Variability Plot*** dialog, select *POUNDS* as the ***Dependent variable***, and *SOIL CONDITION*, *POTSIZE*, and *PRODUCTION METHOD* from the ***Grouping variable*** list.



Select Variables for Variability Plot

1 - SOIL CONDITION
2 - POTSIZE
3 - VARIETY
4 - PRODUCTION ME...
5 - LOCATION
6 - POUNDS

1 - SOIL CONDITION
2 - POTSIZE
3 - VARIETY
4 - PRODUCTION ME...
5 - LOCATION
6 - POUNDS

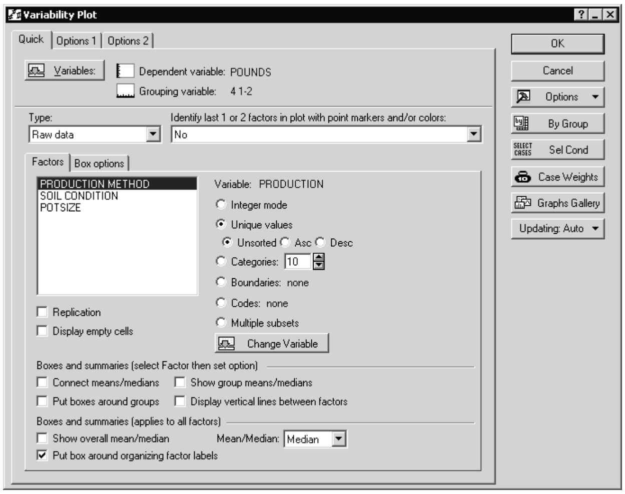
Dependent variable: 6

Grouping variable: 1-2 4

□ Show appropriate variables only

Further on in the example, we will create the graph by *VARIETY* to illustrate the By Group features. Now, click the ***OK*** button in the variable selection dialog.

Reordering variables for variability plot. For the most informative plot, let's reorder the variables so that *PRODUCTION METHOD* will be the first factor in the list of ***Factors***. Click on that variable in the ***Factors*** list, and then, while pressing the left mouse button, drag it to the top of the list.



Dependent variable: POUNDS
Grouping variable: 4 1-2

Type: Raw data
Identify last 1 or 2 factors in plot with point markers and/or colors: No

PRODUCTION METHOD
SOIL CONDITION
POTSIZE

Variable: PRODUCTION
○ Integer mode
● Unique values
● Unsorted ○ Asc ○ Desc
○ Categories: 10
○ Boundaries: none
○ Codes: none
○ Multiple subsets

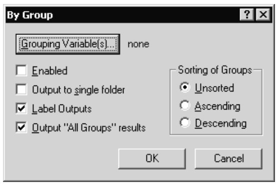
□ Replication
□ Display empty cells

Boxes and summaries (select Factor then set option)
□ Connect means/medians □ Show group means/medians
□ Put boxes around groups □ Display vertical lines between factors

Boxes and summaries (applies to all factors)
□ Show overall mean/median Mean/Median: Median
☑ Put box around organizing factor labels

Finally, also in the ***Variability Plot*** dialog, ensure that *PRODUCTION METHOD* is selected in the ***Factors*** list, and select the ***Put boxes around groups*** check box.

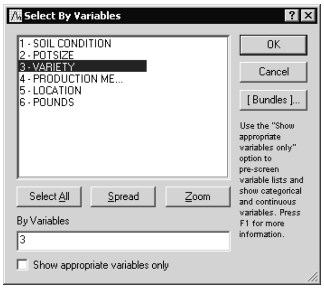
Specifying "by grouping." We want to create the variability plot for *PRODUCTION METHOD, SOIL CONDITION*, and *POTSIZE* for all varieties of tomatoes combined, and broken down by *VARIETY* (one graph per *VARIETY*). Click the ***By Group*** button to display the ***By Group*** dialog.



Grouping Variable(s)... none
□ Enabled
□ Output to single folder
☑ Label Outputs
☑ Output "All Groups" results

Sorting of Groups
● Unsorted
○ Ascending
○ Descending

Click the ***Grouping Variable(s)*** button to display the ***Select By Variables*** dialog, and specify *VARIETY* as the By Group variable.



Select By Variables

1 - SOIL CONDITION
2 - POTSIZE
3 - VARIETY
4 - PRODUCTION ME...
5 - LOCATION
6 - POUNDS

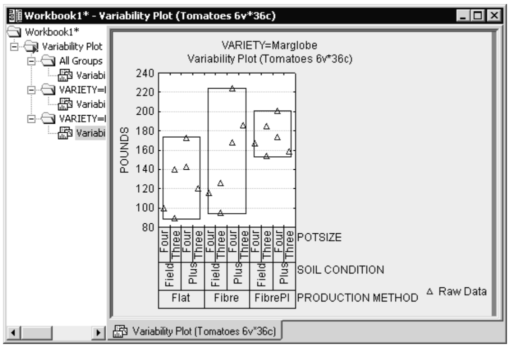
By Variables: 3

□ Show appropriate variables only

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

Note that you can specify more than one By Group variable, in which case all subsequent analyses will be performed broken down by each unique combination of values found in the By Group variables.

Reviewing the variability plots. Now click ***OK*** to close the ***Select By Variables*** dialog, and click ***OK*** to close the ***By Group*** dialog. In the ***Variability Plot*** dialog, click ***OK*** to create the graphs.



Notice how the *Variability Plot* is created 1) for *All Groups*, and 2) for each *Variety* (*Bonny* and *Marglobe*).

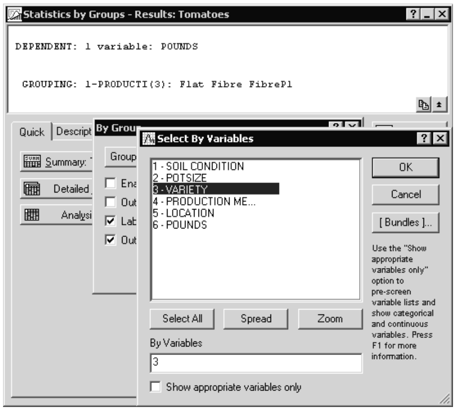
If you review these graphs carefully, you will see that the *Production Method* appears to make little difference (in the observed values for *Pounds*) for *Variety=Bonny*, while for *Variety=Marglobe*, the *FibrePl* method shows the least variability in values, which are generally at the higher end of the distribution of all values for variable *Pounds*.

Descriptive Statistics By Group

Let's next use the descriptive statistics options to further explore this. Select the ***Statistics*** tab. In the ***Base*** group, click ***Basic Statistics*** to display the ***Basic Statistics***

and Tables Startup Panel. Select ***Breakdown & one-way ANOVA***, and click the ***OK*** button to display the ***Statistics by Groups (Breakdown)*** dialog. Click the ***Variables*** button, and in the ***Select the dependent variables and grouping variables*** dialog, specify *Pounds* as the ***Dependent variable*** and *Production Method* as the ***Grouping variable***. Then click ***OK*** to close the variable selection dialog, and click ***OK*** in the ***Statistics by Groups (Breakdown)*** dialog to display the ***Statistics by Groups - Results*** dialog.

We want to compute Statistics by Groups, broken down further by tomato *Variety*. So, click the ***By Group*** button, and in the ***By Group*** dialog, click the ***Grouping Variable(s)*** button. In the ***Select By Variables*** dialog, select *Variety* as the By Group variable.



Statistics by Groups - Results: Tomatoes

```
DEPENDENT: 1 variable: POUNDS

 GROUPING: 1-PRODUCTI(3): Flat Fibre FibrePl
```

Select By Variables

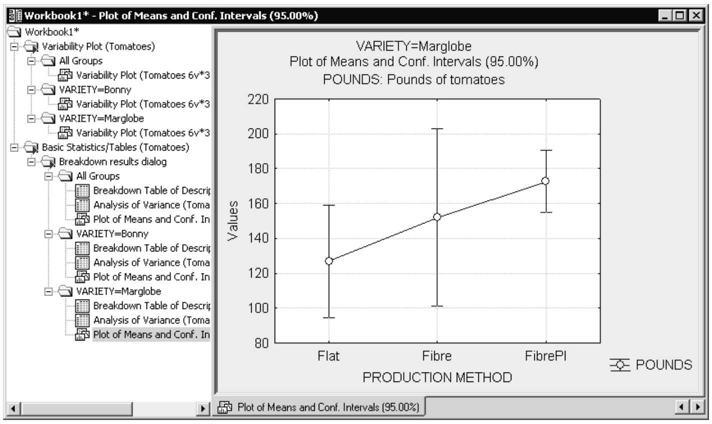
1 - SOIL CONDITION
2 - POTSIZE
3 - VARIETY
4 - PRODUCTION ME...
5 - LOCATION
6 - POUNDS

By Variables: 3

□ Show appropriate variables only

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

Now, click ***OK*** in this dialog and click ***OK*** in the ***By Group*** dialog. In the ***Statistics by Groups - Results*** dialog, click in sequence, 1) the ***Summary*** button, 2) the ***Analysis of Variance*** button, and 3) the ***Interaction plots*** button.


Workbook1* - Plot of Means and Conf. Intervals (95.00%)
Workbook1*
Variability Plot (Tomatoes)
All Groups
Variability Plot (Tomatoes 6v*3
VARIETY=Bonny
Variability Plot (Tomatoes 6v*3
VARIETY=Marglobe
Variability Plot (Tomatoes 6v*3
Basic Statistics/Tables (Tomatoes)
Breakdown results dialog
All Groups
Breakdown Table of Descrip
Analysis of Variance (Toma
Plot of Means and Conf. In
VARIETY=Bonny
Breakdown Table of Descrip
Analysis of Variance (Toma
Plot of Means and Conf. In
VARIETY=Marglobe
Breakdown Table of Descrip
Analysis of Variance (Toma
Plot of Means and Conf. In
VARIETY=Marglobe
Plot of Means and Conf. Intervals (95.00%)
POUNDS: Pounds of tomatoes
220
200
180
160
140
120
100
80
Values
Flat
Fibre
FibrePl
PRODUCTION METHOD
POUNDS
Plot of Means and Conf. Intervals (95.00%)

All results are placed into the respective folder, either the *All Groups* folder or the *Variety=Bonny* or *Variety=Marglobe* folders.

You can now review these results for all groups combined and broken down by *Variety*; as you will see, indeed, *Production Method* appears to have an effect on yield (*Pounds*) for *Variety=Marglobe*, while there is no indication of such an effect for *Variety=Bonny*.

Summary

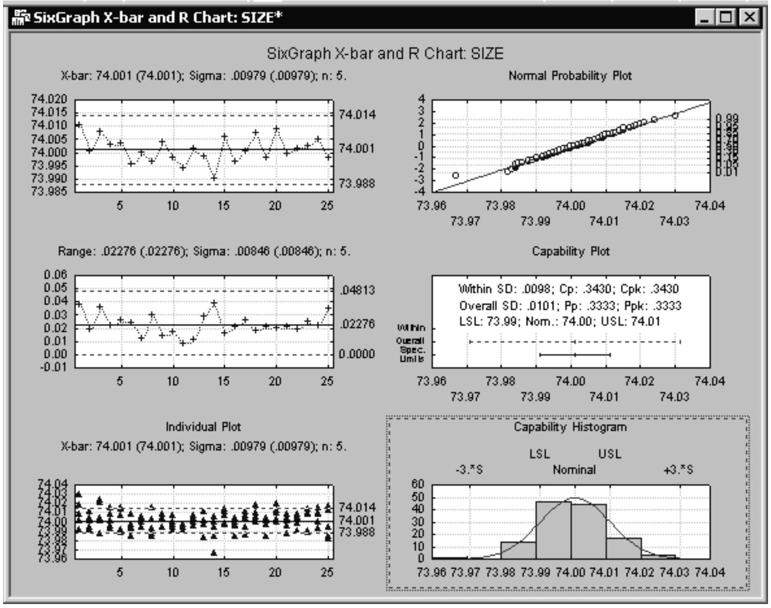
With *STATISTICA*, you can perform ad-hoc by-group analyses from virtually any results dialog, reviewing results for all groups combined or broken down by one or more grouping variable. This very powerful feature for exploratory data analysis can be used to compare groups and verify consistency of results across groups for any analysis.

Before concluding this topic, a few comments about the technical details regarding the implementation of this feature may be useful. When performing by-group analyses, as illustrated in this example, the program will actually rerun the analyses for each group (and all groups), leveraging the *STATISTICA* Visual Basic macro code that is recorded automatically during the interactive analyses, and which can be saved as macros as described elsewhere in this manual (see Chapter 8 - *STATISTICA Visual Basic*). When analyzing very large data problems (e.g., very large unbalanced experimental designs or complex analyses that require iterated computations before results can be displayed), the individual analyses may take up significant amounts of computing time, in particular when there are many unique groups identified in the data (e.g., imagine a complex generalized linear model estimated for each of 100 groups).

Therefore, it is generally a good idea to begin each exploratory analysis by computing simple descriptive statistics, frequency tables, and graphs to understand the structure of the data and identify the number of unique groups (combination of values in the grouping variables) in the data.

Example 5: Summary Results Panels (Quality, Process, Gage–Sixpacks)

Several analyses in *STATISTICA* support summary graphs and reports arranged into a single (graphics) document. In Six Sigma and manufacturing applications, these types of displays are sometimes referred to as “Quality Sixpacks” because they summarize the quality of a single variable with six (or fewer) individual graphs and tables.



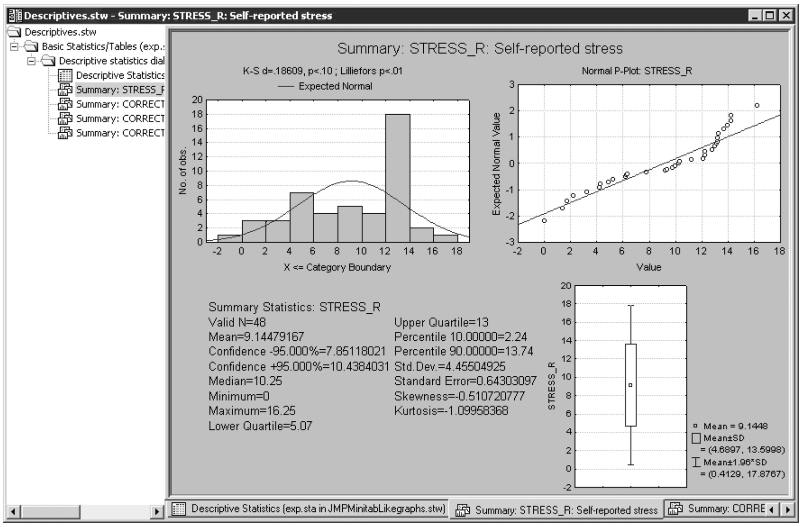
SixGraph X-bar and R Chart: SIZE

X-bar: 74.001 (74.001); Sigma: .00979 (.00979); n: 5.

Range: .02276 (.02276); Sigma: .00846 (.00846); n: 5.

Within SD: .0098; Cp: .3430; Cpk: .3430
Overall SD: .0101; Pp: .3333; Ppk: .3333
LSL: 73.99; Nom.: 74.00; USL: 74.01

STATISTICA incorporates many such displays to summarize basic descriptive statistics, correlations, the results of gage or process capability studies, or other types of data analyses, as shown in the following illustration.


Descriptives.stw - Summary: STRESS_R: Self-reported stress
Descriptives.stw
Basic Statistics/Tables (exp.:
Descriptive statistics dial
Descriptive Statistics
Summary: STRESS_F
Summary: CORRECT
Summary: CORRECT
Summary: CORRECT
Summary: STRESS_R: Self-reported stress
K-S d=.18609, p<.10 ; Lilliefors p<.01
Expected Normal
No. of obs.
X <= Category Boundary
Normal P-Plot: STRESS_R
Expected Normal Value
Value
STRESS_R
Mean = 9.1448
Mean±SD = (4.6897, 13.5998)
Mean±1.96*SD = (0.4129, 17.8767)
Descriptive Statistics (exp.sta in JMPMinitabLikegraphs.stw)
Summary: STRESS_R: Self-reported stress
Summary: CORRE

```
Summary Statistics: STRESS_R
Valid N=48                          Upper Quartile=13
Mean=9.14479167                     Percentile 10.00000=2.24
Confidence -95.000%=7.85118021      Percentile 90.00000=13.74
Confidence +95.000%=10.4384031      Std.Dev.=4.45504925
Median=10.25                        Standard Error=0.64303097
Minimum=0                           Skewness=-0.510720777
Maximum=16.25                       Kurtosis=-1.09958368
Lower Quartile=5.07
```

Process Capability Analysis Consistent with DIN 55319 and ISO 21747

In recent years, European (and other international) manufacturers have developed standards for the computation of process capability indices that will explicitly account for systematic and random process variation over time, as well as non-normal distributions. These indices have, for example, been adopted throughout the auto manufacturing industry and their suppliers, and *STATISTICA* fully supports these standards.

Process capability indices measure the number of times that the observed (normal) distribution of values can fit inside the specification limits for the respective part under consideration. Thus, these indices summarize the quality of a process to produce products or parts that are consistent with design specifications. In short, *DIN (Deutsche Industrie Norm) 55319* and *ISO 21747* describe the rules to apply when choosing among various distribution models and how to account for time-dependent variation in the process.

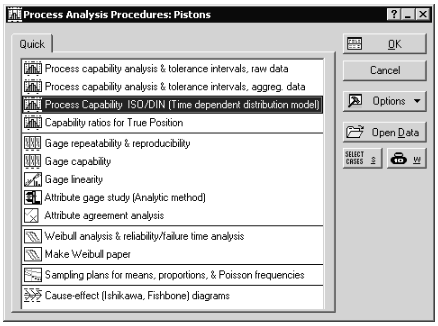
For example, even if a distribution of data points within each sample is *Normal*, if there is systematic or random variation that occurs over time as successive samples are taken, the resultant distribution of values will *not* be *Normal*. Therefore, in many cases the normal distribution-based process capability computations will not be applicable. Also, it is usually of interest to identify any time-dependent variability or trends because they can indicate machine wear or other process problems.

The following example illustrates step-by-step how to compute process capability indices consistent with these international standards, and how to create an efficient single-document summary report.

Select data. This example is based on a data set reported in Montgomery (1985, page 177, 1991, page 234). We'll use the data file *Pistons.sta* that is located in *STATISTICA*'s examples directory. Specifically, we are interested in monitoring the size (diameter) of piston rings for automotive engines. Therefore, constant samples of five observations each have been taken from the ongoing manufacturing process. As is the case in many ongoing manufacturing processes, samples are taken over time, so any variability in the process quality over time will affect the overall variability.

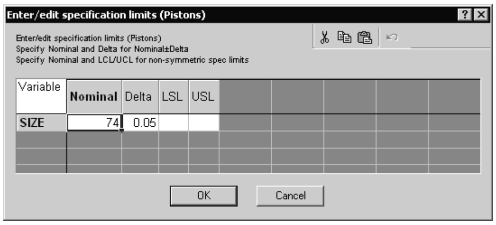
On the ***Home*** tab, click the ***Open*** arrow, and from the drop-down menu, select ***Open Examples*** to display the ***Open a STATISTICA Data File*** dialog; open the *Datasets* folder, and double-click on *Pistons.sta* or select it and click the ***Open*** button.

Specify analysis. Select the ***Statistics*** tab. In the ***Industrial Statistics*** group, click ***Process Analysis***. In ***the Process Analysis Procedures*** Startup Panel, select ***Process Capability ISO/DIN (Time dependent distribution model)***.


Process Analysis Procedures: Pistons
Quick
Process capability analysis & tolerance intervals, raw data
Process capability analysis & tolerance intervals, aggreg. data
Process Capability ISO/DIN (Time dependent distribution model)
Capability ratios for True Position
Gage repeatability & reproducibility
Gage capability
Gage linearity
Attribute gage study (Analytic method)
Attribute agreement analysis
Weibull analysis & reliability/failure time analysis
Make Weibull paper
Sampling plans for means, proportions, & Poisson frequencies
Cause-effect (Ishikawa, Fishbone) diagrams
OK
Cancel
Options
Open Data
SELECT CASES S
W

Click the ***OK*** button in the ***Process Analysis Procedures*** Startup Panel. On the ***Quick*** tab of the ***ISO 21747 - Process Capability Setup*** dialog, click the ***Variables*** button. In the ***Select Variables (and optional grouping variable)*** dialog, select variable *Size* in the ***Variables for the analyses*** list, and *Sample* in the ***by ... (Time/Grouping var.)*** list, and click ***OK***.

In the ***ISO 21747 - Process Capability Setup*** dialog, click the ***Process specs*** button to display the ***Enter/edit specification limits*** dialog, where you can enter the process specification limits. Specification or design limits define the maximum and (or) minimum allowable values for the respective part; in this case, specify the lower and upper spec limits (LSL, USL) as 74 +/- 0.05 (LSL=73.95, USL=74.05). Enter *74* in the ***Nominal*** field, and enter *0.05* in the ***Delta*** field.

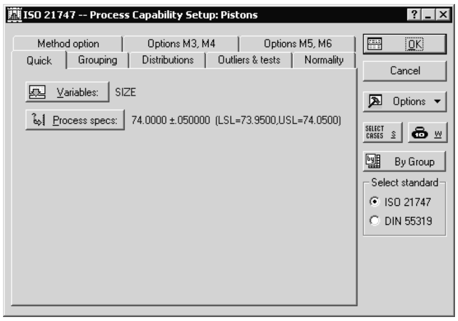


Enter/edit specification limits (Pistons)

Enter/edit specification limits (Pistons)
Specify Nominal and Delta for Nominal±Delta
Specify Nominal and LCL/UCL for non-symmetric spec limits

Variable	Nominal	Delta	LSL	USL
SIZE	74	0.05		

Click ***OK*** to finalize this choice and return to the ***ISO 21747 - Process Capability Setup*** dialog.



ISO 21747 -- Process Capability Setup: Pistons

Variables: SIZE

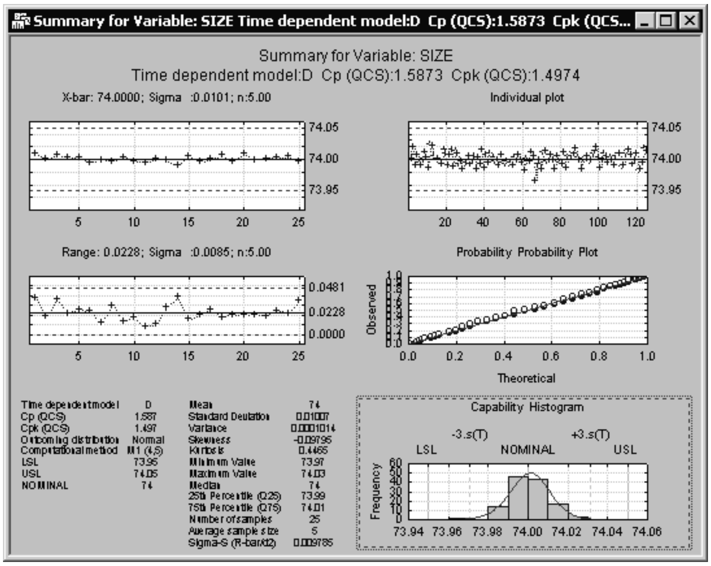
Process specs: 74.0000 ±.050000 (LSL=73.9500,USL=74.0500)

Select standard
● ISO 21747
○ DIN 55319

In this dialog, there are numerous other options available to modify the rules that are applied to select the most appropriate distribution and time-dependent distribution model for the data so that the appropriate process capability indices can be computed. You can click the ? button in the upper-right corner of the dialog or press F1 to display the *STATISTICA Electronic Help* topic containing specific details regarding all options in this dialog. For example, the details regarding the (small) differences in the DIN and ISO specifications are discussed there.

Now click the ***OK*** button in the ***ISO 21747 - Process Capability Setup*** dialog to perform the analyses for variable Size.

Reviewing results. In the ***ISO 21747 - Process Capability Results*** dialog, click the ***Summary*** button to review the analysis summary display.



Summary for Variable: SIZE Time dependent model:D Cp (QCS):1.5873 Cpk (QCS...

Summary for Variable: SIZE
Time dependent model:D Cp (QCS):1.5873 Cpk (QCS):1.4974

Time dependent model	D
Cp (QCS)	1.587
Cpk (QCS)	1.497
Outcoming distribution	Normal
Computational method	M1 (4,5)
LSL	73.95
USL	74.05
NOMINAL	74

Mean	74
Standard Deviation	0.01007
Variance	0.0001014
Skewness	-0.09795
Kurtosis	0.4465
Minimum Value	73.97
Maximum Value	74.03
Median	74
25th Percentile (Q25)	73.99
75th Percentile (Q75)	74.01
Number of samples	25
Average sample size	5
Sigma-S (R-bar/d2)	0.009785

As you can see, all relevant details (as recommended in *ISO 21747* and/or *DIN 55319*) are summarized on a single page (document), which contains all information necessary to judge the process as *capable* or *not capable* (or questionable).

Attribute Gage Analysis

For another example of this type of summary (compound) displays in *STATISTICA*, we will perform an attribute gage analysis.

In general, any measurement system used in manufacturing must be validated to ensure that the respective gages measure the quality characteristic of interest with sufficient accuracy and precision. Often, a gage of particular importance is the one that determines whether a manufactured part is of sufficient quality to be accepted or rejected; in other words, the gage measures a simple accept/reject *attribute*.

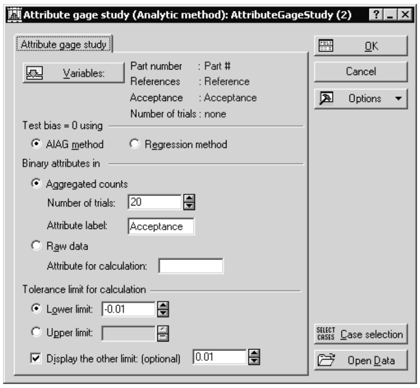
To determine the quality of the gage, a study is periodically performed where the gage (accept/reject decision) is applied to reference parts with known deviations from the desired specifications. This process is described in the respective section of the *STATISTICA Electronic Manual*, as well as the AIAG (Automotive Industry Action Group) *Measurement System Analysis (MSA)* manual (2000).

This example illustrates the analysis described in the *MSA* manual on pages 81-86.

Select data. Open the *AttributeGageStudy.sta* data file*.* This file contains the data, already summarized to acceptance data, of the attribute gage study described in the *MSA* manual, (p. 84)

Specify analysis. Select the ***Statistics*** tab. In the ***Industrial Statistics*** group, click ***Process Analysis***. In the ***Process Analysis Procedures*** Startup Panel, select ***Attribute gage study (Analytic method)***, and click the ***OK*** button.

In the ***Attribute gage study (Analytic method)*** dialog, click the ***Variables*** button. Select *Part#* in the ***Part numbers*** list, *Reference* in the ***Reference values*** list, and *Acceptance* in the ***Acceptance/Response*** list, and then click the ***OK*** button to close this dialog and return to the ***Attribute gage study (Analytic methods)*** dialog. In the ***Tolerance limit for calculation*** group, specify *-0.01* as the ***Lower limit***, select the ***Display the other limit*** check box, and then specify *0.01* as that limit.



Attribute gage study (Analytic method): AttributeGageStudy (2)

Part number : Part #
References : Reference
Acceptance : Acceptance
Number of trials : none

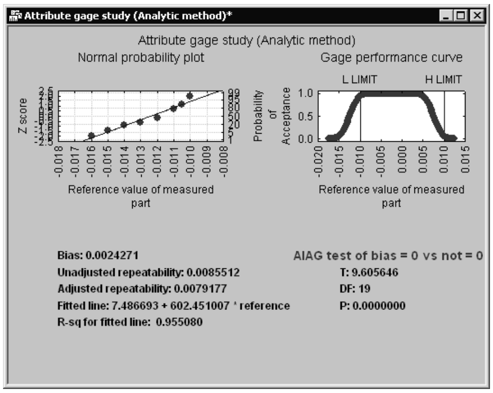
Test bias = 0 using
● AIAG method ○ Regression method

Binary attributes in
● Aggregated counts
Number of trials: 20
Attribute label: Acceptance
○ Raw data
Attribute for calculation:

Tolerance limit for calculation
● Lower limit: -0.01
○ Upper limit:
☑ Display the other limit: (optional) 0.01

We are interested in evaluating the gage performance for a process or type of manufactured part that should be identified as unacceptable (should be rejected), when its real lower limit drops below *-0.01* (expressed here as a deviation from the spec). In the data file, the *Acceptance* probabilities summarize the number of reference parts measurements, from a total of 20 such parts and measurements each, that were declared as unacceptable (i.e., that were rejected).

Reviewing results. Now click ***OK*** in the ***Attribute gage study (Analytic methods***) dialog. In the ***Results*** dialog, click the ***Summary*** button to review the summary results.



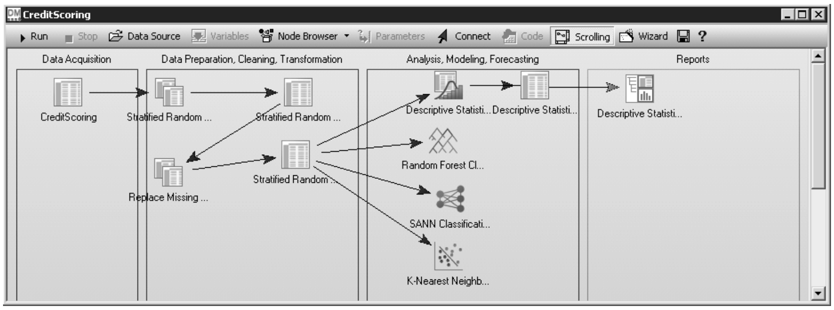
Bias: 0.0024271
Unadjusted repeatability: 0.0085512
Adjusted repeatability: 0.0079177
Fitted line: 7.486693 + 602.451007 * reference
R-sq for fitted line: 0.955080

AIAG test of bias = 0 vs not = 0
T: 9.605646
DF: 19
P: 0.0000000

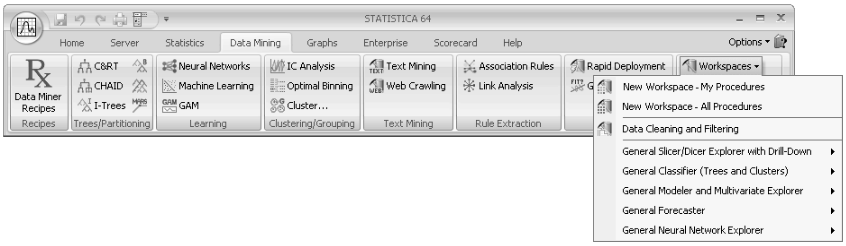
All important results to determine the bias and repeatability (of measurements) of the attribute gage are summarized on a single page. For details on the interpretation of the reported statistics and graphs, refer to the *Electronic Manual.*

Example 6: *STATISTICA Data Miner*

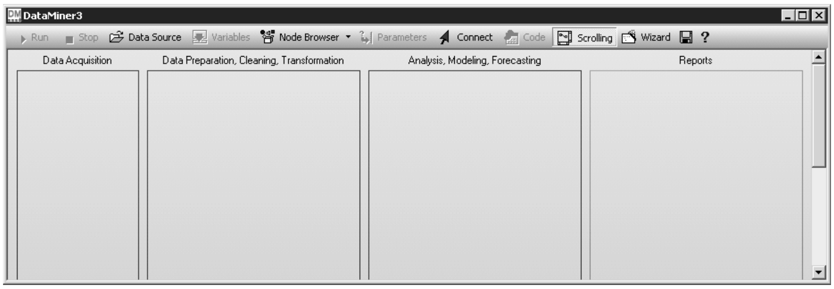
STATISTICA Data Miner (SDM) is a comprehensive system for predictive modeling that offers a wide variety of analytic techniques and model building, validation, and model deployment options. The default, and perhaps the industry standard, type of user interface provided in *SDM* follows the general “interactive data mining workspace” approach that enables users to build models by dragging icons representing steps of data acquisition, data preparation, modeling, and deployment and connect them with arrows. The workspace user interface option in *SDM* represents a powerful alternative to the traditional interactive data analysis user interface, and it can be used not only as a tool for developing and testing predictive data mining modes, but also as a powerful general tool to be used for “visual programming” of analytic workflows for many types of analyses.



To open a new (blank) data mining workspace, select the ***Data Mining*** tab. In the ***Tools*** group, click ***Workspaces*** and from the menu, select either ***New Workspace - My Procedures*** or ***New Workspace - All Procedures***.



A blank data mining workspace will be displayed.



Now, click Data Source on the toolbar to display the ***Select Data Source*** dialog, used to select a data file for analysis. Next, the ***Select dependent variables and predictors*** dialog is displayed; click the Variables button to display the variable selection dialog, used to specify the dependent variables and predictors. Then, click Node Browser to create analytic nodes, and connect them with Connect arrows to specify the desired project workflow.

The following section includes a step-by-step example of *Data Miner Recipes* – an innovative user interface for data mining introduced by StatSoft – which offers a powerful alternative to the workspace-based approach to model building, and can be used by both novices and advanced analysts.

Overview

This example pertains to *STATISTICA Data Miner Recipes*, a StatSoft product that offers a wide selection of methods for predictive data mining.

A general trend in data mining is the increasing emphasis on solutions based on simple analytic processes rather than the creation of ever-more sophisticated general analytic tools. *STATISTICA Data Miner Recipes* (*SDMR*) offers an easy-to-use alternative to the traditional data miner workspace user interface for building predictive data mining models. This approach provides an intuitive graphical interface to enable those with limited data mining experience to execute a “recipe-like” step-by-step analytic process. With these intuitive dialogs, you can perform various data mining tasks such as regression, classification, and clustering. Other recipes can be built quickly as custom solutions.

Completed recipes can be saved and deployed as project files to score new data. The project files can be generated as C/C++ language or PMML script, or sent to *STATISTICA Enterprise.*

The *SDMR* user interface can also be used by advanced analysts to automate and store specific data mining algorithms.

SDMR spans the entire data mining process – from querying external databases to the final deployment of solutions – and, in general, consists of the following steps.

1. Identifies the data from which to learn
   - Connects to ODBC or OLEDB compliant databases
   - Connects to *STATISTICA* data files
2. Cleans data and removes the redundant predictors
   - Flexible and efficient methods for sampling the data (simple, stratified, systematic, etc.)
   - More flexible ways to identify and recode the missing data
   - Identification of outliers
   - Transform the data prior to performing the subsequent steps
   - Identify and eliminate redundant predictors
3. Identifies important predictors from a large pool of predictors that are strongly related to the dependent (outcome or target) variable of interest
   - Feature selection for very large data sets (e.g., thousands of variables)
   - Detection of important interactions among the predictors by using tree-based methods

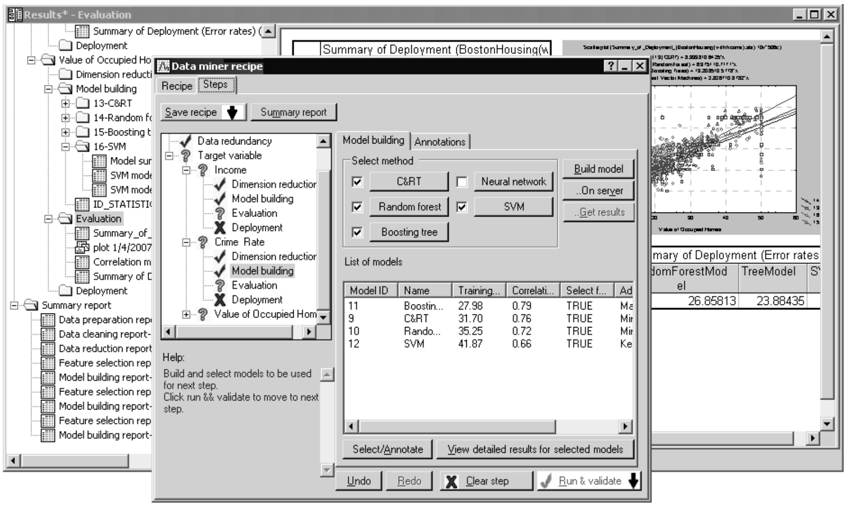
4. Generates a pool of eligible models
    - Leverage the comprehensive selection of cutting edge techniques for predictive data mining available in *SDMR*
    - Offload computationally expensive tasks to *STATISTICA Enterprise Server*, freeing your local computer for other tasks
5. Performs automatic competitive evaluation of models to identify the optimum model with respect to performance and complexity
6. Deploys the model to score new data using the inbuilt efficient deployment engine

STATISTICA Data Miner Recipes provides the solution that maps the steps of the data mining workflow into a results-oriented user interface. From data cleaning to model validation, *SDMR* guides your analysis from start to finish so that you can get actionable results and answers quickly. At the same time, *SDMR* still applies the most comprehensive collection of data mining algorithms in a single package without requiring the user to know the details of those algorithms.

STATISTICA Data Miner Recipes contains the largest collection of data mining methods and algorithms in a single package or library. In most general terms, these algorithms borrow insights and methodologies from various domains such as statistics, engineering, artificial intelligence, cognitive science, etc., to "learn" patterns from data that can be used to make predictions (about insurance or credit risk, process or product quality, equipment failure, medical diagnoses, and so on). The *STATISTICA Electronic Manual* and the on-line *Electronic Statistics Textbook* provide detailed introductions to the various methods and techniques that are usually summarily described as "data mining."

In practice, specific domains and types of data are best analyzed using particular types of methods and algorithms. For example, the data mining techniques that work best for modeling insurance loss data are different from those that work best for predicting emissions from a furnace. However, there *is* a typical workflow – from the definition of the data and analysis problem through sampling, model building, and evaluation – that is applicable to all predictive data mining.

Data Miner Recipes enable those without extensive experience with data mining tools to move very quickly from the definition of a problem to tangible and actionable results.



Select method

☑ C&RT ☐ Neural network
☑ Random forest ☑ SVM
☑ Boosting tree

List of models

Model ID	Name	Training...	Correlati...	Select f...	Ad
11	Boostin...	27.98	0.79	TRUE	Ma
9	C&RT	31.70	0.76	TRUE	Mi
10	Rando...	35.25	0.72	TRUE	Mi
12	SVM	41.87	0.66	TRUE	Ke

Help:
Build and select models to be used for next step.
Click run && validate to move to next step.

In this approach, you simply follow a recipe-like user interface to complete the necessary steps to move to a solution. In fact, most of these steps are entirely automated so that the only required input is to define the data and variables for the analyses, while the program automatically does the rest – determines learning and testing samples, performs feature selection, tries various data mining algorithms and methods, and evaluates results to select the best data mining model. These computations and analyses can be performed with either the desktop *STATISTICA Data Miner* software or, if available, on the *STATISTICA Data Miner Server*.

Data Miner Recipes Project Files

When you save a *Data Miner Recipes* project at any stage of completion, two separate files are created:

- A *Data Miner Recipes* file with the file name extension *.dmrproj*
- A *STATISTICA Workbook* file by the same name, but with the file name extension *.stw*, containing results and detailed information for each step of the recipe

It is important that both files reside in the same file directory. So, if you want to copy a *Data Miner Recipe* project called *MyDataMinerProject* to a new file directory, email it to a colleague, or check it into the *STATISTICA Document Management System*, then both files – *MyDataMinerProject.dmrproj* and *MyDataMinerProject.stw* – must be copied to the new destination.

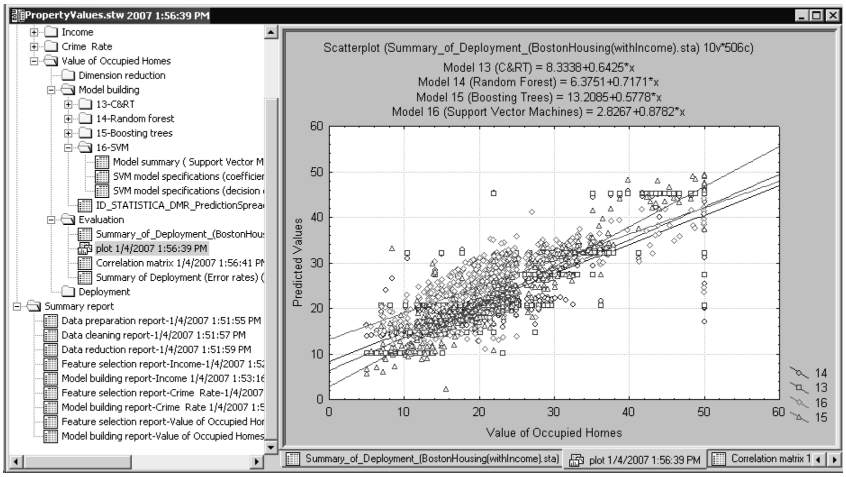
Following are additional details about these two files.

Data Miner Recipes file (.dmrproj). The *Data Miner Recipes* are *XML* (extensible markup language) format files that contain all information regarding users' choices (or choices automatically made by the program), including:

- Data file information (or data connection information)
- Variable selections and variable metadata (e.g., defining continuous and categorical predictors and outcomes)
- Choices about data preprocessing steps (e.g., missing data handling, filtering of duplicate records, transformations, etc.)
- Final variable selections based on the application of feature selection algorithms
- Results from model building and final evaluation and choices of models
- All information necessary to deploy predictive models and to predict new cases (e.g., to score databases, compute component scores, inferred sensor values, predicted risk or failure probabilities, etc.)

Therefore, when deploying *Data Miner Recipes* to the *STATISTICA Enterprise* software to automatically compute predicted values in an enterprise application (automated credit scoring, multivariate control charting and failure analysis, etc.), all information necessary to compute predicted values, classifications, or classification probabilities (e.g., probability of default, loss) is contained inside these XML format files.

Data Miner Recipes Workbook file (.stw). These files contain detailed information describing the results for each step.


PropertyValues.stw 2007 1:56:39 PM
Income
Crime Rate
Value of Occupied Homes
Dimension reduction
Model building
13-C&RT
14-Random forest
15-Boosting trees
16-SVM
Model summary (Support Vector M
SVM model specifications (coefficier
SVM model specifications (decision
ID_STATISTICA_DMR_PredictionSprea
Evaluation
Summary_of_Deployment_(BostonHou
plot 1/4/2007 1:56:39 PM
Correlation matrix 1/4/2007 1:56:41 PI
Summary of Deployment (Error rates) (
Deployment
Summary report
Data preparation report-1/4/2007 1:51:55 PM
Data cleaning report-1/4/2007 1:51:57 PM
Data reduction report-1/4/2007 1:51:59 PM
Feature selection report-Income-1/4/2007 1:5
Model building report-Income 1/4/2007 1:53:1
Feature selection report-Crime Rate-1/4/2007
Model building report-Crime Rate 1/4/2007 1:5
Feature selection report-Value of Occupied Ho
Model building report-Value of Occupied Homes
Predicted Values
0
10
20
30
40
50
60
Value of Occupied Homes
14
13
16
15
Summary_of_Deployment_(BostonHousing(withIncome).sta)
plot 1/4/2007 1:56:39 PM
Correlation matrix 1

Scatterplot (Summary_of_Deployment_(BostonHousing(withIncome).sta) 10v*506c)
Model 13 (C&RT) = 8.3338+0.6425*x
Model 14 (Random Forest) = 6.3751+0.7171*x
Model 15 (Boosting Trees) = 13.2085+0.5778*x
Model 16 (Support Vector Machines) = 2.8267+0.8782*x

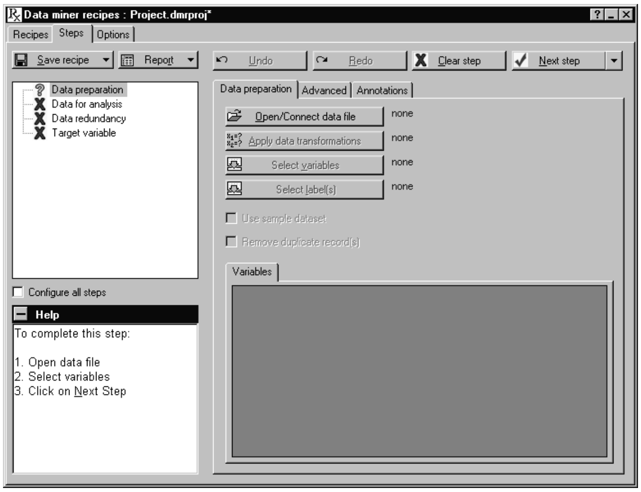
The results stored in this workbook provide complete documentation for the computations and analyses performed as the *Data Miner Recipe* was (or is in the process of being) completed. Therefore, if the data mining analyses are performed in a regulated (e.g., FDA, ISO, etc.) environment, or if data mining is part of an organization's mission critical activities performed under the guidance and in compliance with specific standard operating procedures (SOPs), then it is usually recommended that this file be stored in the *STATISTICA Document Management System* along with the *Data Miner Recipe* project file (*.dmrproj*).

Using *STATISTICA Data Miner Recipes (SDMR)*

This example illustrates how quickly and efficiently data mining projects can be completed using *STATISTICA Data Miner Recipes*, even if the best solution to the (prediction) problem emerges only after (automatically) comparing the efficacy of various advanced data mining algorithms.

In this example, we will explore the use of *SDMR* for credit scoring applications. The example is based on the data file *CreditScoring.sta*, which contains observations on 18 variables for 1,000 past applicants for credit. Each applicant was rated as "good credit" (700 cases) or "bad credit" (300 cases). We want to develop a credit scoring model that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on the values of one or more of the predictor variables. An additional “Train/Test” indicator variable is also included in the data file for validation purposes.

In *STATISTICA*, select the ***Data Mining*** tab. In the ***Recipes*** group, click ***Data Miner Recipes*** to display the ***Data miner recipes*** dialog. On the ***Recipes*** tab, click the ***New*** button to create a new project. The ***Steps*** tab will be selected automatically.



Open/Connect data file: none
Apply data transformations: none
Select variables: none
Select label(s): none

To complete this step:

1. Open data file
2. Select variables
3. Click on Next Step

The step-node panel is located in the upper-left area of the ***Steps*** tab. It contains four major nodes: ***Data preparation***, ***Data for analysis***, ***Data redundancy***, and ***Target variable***.

Nodes (steps). Each node (or step) can exist in one of four states, depending on whether all required options have been specified. Each state is represented by an icon: a red ✗ indicates a wait state, meaning a step cannot be started because it is dependent on a previous step that has not been completed; a yellow ? indicates a ready state, meaning you are ready to start the step because previous steps have been completed; a green ✓ indicates a completed step. Note that you must click the ***Next step*** button to change the yellow ? (ready state) to the green ✓ (completed state). The change will be made only if the step has been successfully completed (i.e., all required information has been specified). Lastly, if you have opened a data set and selected variables, and you do not want to proceed step by step through all the options, you can select the ***Configure all steps*** check box on the ***Steps*** tab. The steps will now be represented by a navy ● icon. You can select any of the steps and modify the options, or you can leave all options at their defaults. Then, click the ***Next step*** arrow, and from the drop-down list, select ***Run to completion***. *STATISTICA Data Miner Recipes* will run the analysis and create the model results.

Options tab. The ***Options*** tab of *STATISTICA Data Miner Recipes* is used to set global options for recipes using very large data files. Options include specifications for sampling and for maximum file size to save in the Project Workbook. Since most of these options are applied to the Data preparation step, they should be set prior to starting work on a new recipe. Modifications to the values on this tab apply only to the current recipe unless you click the ***Save defaults*** button.

Data Preparation

Connecting data. On the ***Data preparation*** tab, click the ***Open/Connect data file*** button. In the ***Select Data Source*** dialog, click the ***Files*** button to browse to and open the *CreditScoring.sta* data file (located in the *Datasets* folder installed with *STATISTICA*). If the data file is already open, it will be listed in the *Open Spreadsheet Documents* folder; double-click it to open it, or select it and click the ***OK*** button.

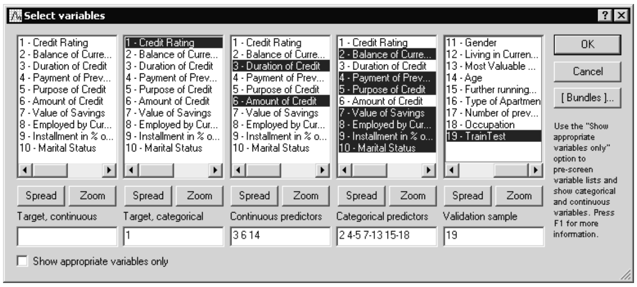
On the ***Data preparation*** tab, click the ***Select variables*** button, and in the ***Select variables*** dialog, select:

Variable *1 - (Credit Rating)* as the ***Target, categorical*** variable,

Variables *3*, *6*, and *14* as ***Input, continuous*** predictors

Variables *2*, *4-5*, *7-13*, and *15-18* as ***Input, categorical*** predictors, and

Variable *19 - TrainTest* as the ***Testing sample*** variable.



Select variables

Target, continuous:

Target, categorical: 1

Continuous predictors: 3 6 14

Categorical predictors: 2 4-5 7-13 15-18

Validation sample: 19

□ Show appropriate variables only

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

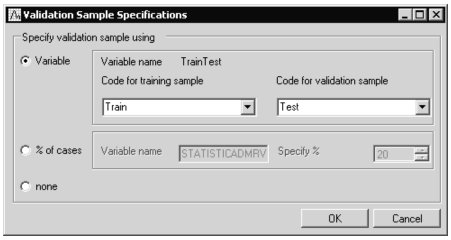
Then, click the ***OK*** button.

Select the ***Advanced*** tab in the ***Data miner recipes*** dialog, and select the ***Use sample data*** check box. Select the ***Stratified random sampling*** option button as the sampling strategy to ensure that each class of the dependent variable *Credit Rating* is represented with approximately equal numbers of cases in train and validation sets. Then click the ***More options*** button to display the ***Stratified sampling*** dialog. Click the ***Strata variables*** button, select *Credit Rating* as the strata variable, and click ***OK*** in this dialog and in the ***Stratified sampling*** dialog.

Click the ***Next step*** button for the ***Data preparation*** step to ensure that this step has been successfully completed (in the step-node panel next to ***Data preparation***, the yellow ? changes to a green ✓).

Data for Analysis

After the ***Data preparation*** step is completed, the ***Data for analysis step*** will be selected automatically. On the ***Data for analysis*** tab, click ***the Select testing sample*** button, and in the ***Testing Sample Specifications*** dialog, select the ***Variable*** option button. Verify that the category (value) *Train* is entered in the ***Code for training sample*** field and *Test* is entered in the ***Code for testing sample*** field.



Validation Sample Specifications

Specify validation sample using

● Variable
Variable name TrainTest
Code for training sample: Train
Code for validation sample: Test

○ % of cases
Variable name: STATISTICADMRV
Specify %: 20

○ none

Then, click the ***OK*** button. The models will be fitted using the training sample and evaluated using the observations in the testing sample. By using observations that did not participate in the model fitting computations, the goodness-of-fit statistics computed for (predicted values derived from) the different data mining models (algorithms) can be used to evaluate the predictive validity of each model and, hence, can be used to compare models and to choose one or more over others.

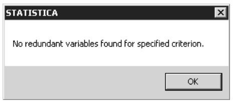
Descriptive statistics. This step will also compute descriptive statistics for all variables selected in the analysis. Descriptive stats provide useful information about ranges and distributions of the data used for the project.

Click the ***Next step*** button to ensure that this step is successfully complete.

Data Redundancy

Now, the ***Data redundancy*** step will be selected. The purpose of this step is to eliminate highly redundant predictors. For example, if the data set contained two measures for weight, one in kilogram the other in pounds, then those two measures would be redundant.

On the ***Data redundancy*** tab, select the ***Correlation coefficient*** option button, and specify the ***Criterion value*** as *0.8*. Click the ***Next step*** button to eliminate the redundant predictors that are highly correlated (r≥0.8). Since there is no redundancy in the data set we are using in this example, a message dialog will be displayed stating this.

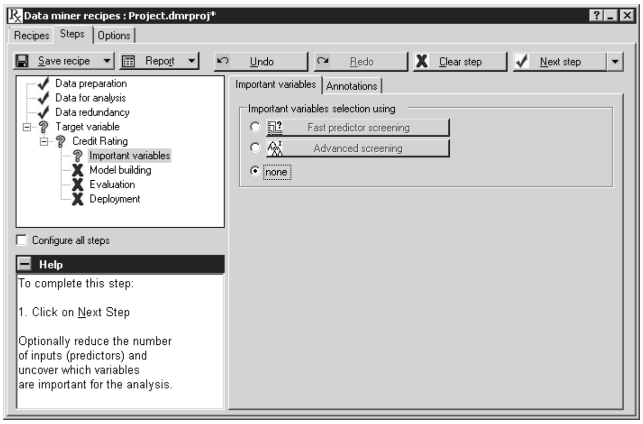


No redundant variables found for specified criterion.

Click the ***OK*** button. The data cleaning and preprocessing for model building is now complete.

Target Variable: Building Predictive Model

Next, we need to build predictive models for the target in this example. In the step-node panel, the ***Target variable*** node has a branching structure with the parent node connecting to four child nodes including ***Important variables***, ***Model building***, ***Evaluation***, and ***Deployment***.



To complete this step:

1. Click on Next Step

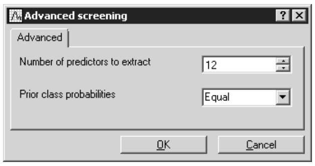
Optionally reduce the number of inputs (predictors) and uncover which variables are important for the analysis.

Important variables selection using: ○ Fast predictor screening ○ Advanced screening ● none

Dimension reduction. The ***Important variables*** node is selected automatically. In this step, the goal is to reduce the dimensionality of the prediction problem, i.e., to select a subset of inputs that is most likely related to the target variable (in this example *Credit Rating*) and, thus, is most likely to yield accurate and useful predictive models. This type of analytic strategy is also sometimes called *feature selection*.

Two strategies are available. When the ***Fast predictor screening*** option button is selected, the program will screen through thousands of inputs and find the ones that are strongly related to the dependent variable of interest. When the ***Advanced screening*** option button is selected, tree methods are used to detect important interactions among the predictors.

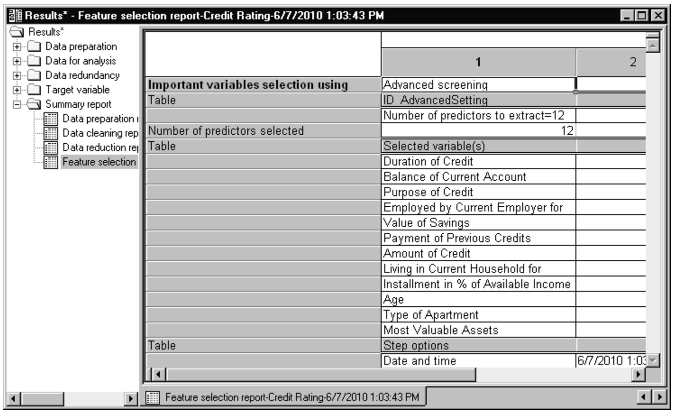
For this example, select the ***Advanced screening*** option button as the feature selection strategy, and then click the ***Advanced screening*** button to display the ***Advanced screening*** dialog. Enter *12* in the ***Number of predictors to extract*** field, and select ***Equal*** in the ***Prior class probabilities*** field.


Advanced screening

Number of predictors to extract: 12

Prior class probabilities: Equal

Click the ***OK*** button in this dialog, and then click the ***Next step*** button to complete this step. To review a summary of the analysis thus far, on the ***Steps*** tab, click the ***Report*** button, and from the drop-down list, select ***Summary report*** to display the *Results* workbook.


Results* - Feature selection report-Credit Rating-6/7/2010 1:03:43 PM

	1	2
Important variables selection using	Advanced screening	
Table	ID AdvancedSetting	
	Number of predictors to extract=12	
Number of predictors selected	12	
Table	Selected variable(s)	
	Duration of Credit	
	Balance of Current Account	
	Purpose of Credit	
	Employed by Current Employer for	
	Value of Savings	
	Payment of Previous Credits	
	Amount of Credit	
	Living in Current Household for	
	Installment in % of Available Income	
	Age	
	Type of Apartment	
	Most Valuable Assets	
Table	Step options	
	Date and time	6/7/2010 1:03

These predictors will be further examined using various cutting-edge data mining and machine learning algorithms available in *SDMR*.

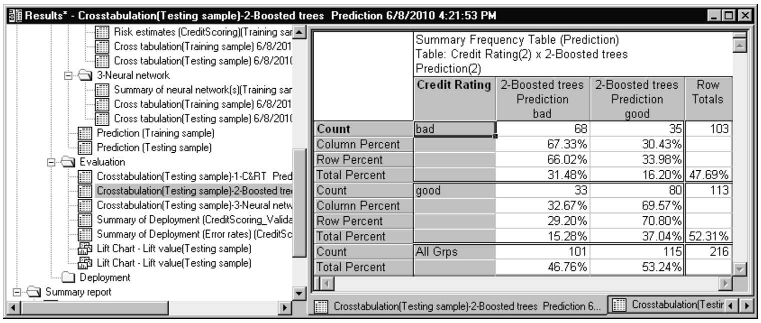
Building models. The ***Data miner recipes*** dialog was minimized so that you could see the *Results* workbook. Click the ***Data miner recipes*** button located on the Analysis Bar (in the lower-left corner of the application) to display the dialog again. Now, the ***Model building*** node is selected. In this step, you can build a variety of models for the selected inputs. On the ***Model building*** tab, ***C&RT***, ***Boosted tree***, and ***Neural network*** are selected by default as the models or algorithms that will automatically be "tried" against the data.

The computations for building predictive models can be performed either locally (on your computer) or on the *STATISTICA Enterprise Server*. However, the latter option is available only if you have a valid *STATISTICA Enterprise Server* account and you are connected to the server installation at your site. For this example, click the ***Build model*** button to perform the computations locally on your computer. This will take a few moments; when finished, click the ***Next step*** button to complete this step.

Evaluating and selecting models. Now, the ***Evaluation*** node is selected. On the ***Evaluation*** tab in the ***Select model(s)*** field, ensure that all models are selected (each check box is selected). Click the ***Evaluate models*** button to perform the competitive evaluation of models for identifying the best performing model in terms of performance in the validation sample.

Notice that the *Boosting Trees* model has the minimum error rate of 31.48%. In other words, 68.52% of the cases in the validation sample are correctly predicted by this model. Note that your results may vary slightly because these advanced data mining methods randomly split the data into subsets during training to produce reliable estimates of the error rates.

The following spreadsheet shows the classification performance of the best model on the validation data set. The columns represent the predicted class frequencies, as predicted by the *Boosting Trees* model, and the rows represent the actual or observed classes in the validation sample.



Results* - Crosstabulation(Testing sample)-2-Boosted trees Prediction 6/8/2010 4:21:53 PM
Risk estimates (CreditScoring)(Training sar
Cross tabulation(Training sample) 6/8/201
Cross tabulation(Testing sample) 6/8/201(
3-Neural network
Summary of neural network(s)(Training sar
Cross tabulation(Training sample) 6/8/201
Cross tabulation(Testing sample) 6/8/201(
Prediction (Training sample)
Prediction (Testing sample)
Evaluation
Crosstabulation(Testing sample)-1-C&RT Pred
Crosstabulation(Testing sample)-2-Boosted tre
Crosstabulation(Testing sample)-3-Neural netw
Summary of Deployment (CreditScoring_Valida
Summary of Deployment (Error rates) (CreditSc
Lift Chart - Lift value(Testing sample)
Lift Chart - Lift value(Testing sample)
Deployment
Summary report
Crosstabulation(Testing sample)-2-Boosted trees Prediction 6...
Crosstabulation(Testir

Summary Frequency Table (Prediction)
Table: Credit Rating(2) x 2-Boosted trees Prediction(2)

	Credit Rating	2-Boosted trees Prediction bad	2-Boosted trees Prediction good	Row Totals
Count	bad	68	35	103
Column Percent		67.33%	30.43%	
Row Percent		66.02%	33.98%	
Total Percent		31.48%	16.20%	47.69%
Count	good	33	80	113
Column Percent		32.67%	69.57%	
Row Percent		29.20%	70.80%	
Total Percent		15.28%	37.04%	52.31%
Count	All Grps	101	115	216
Total Percent		46.76%	53.24%	

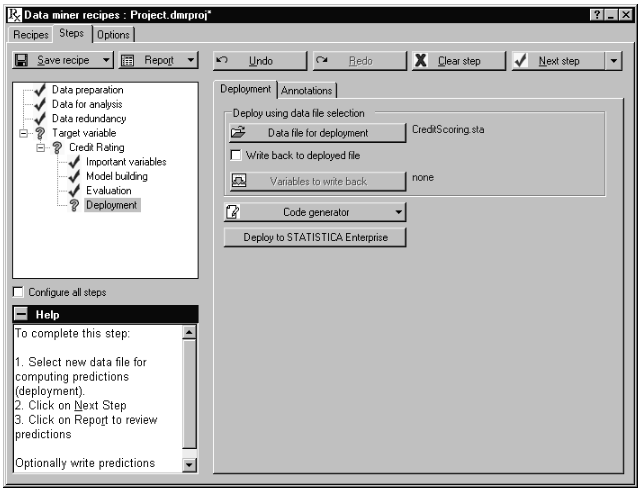
In this matrix, you can see that this model predicted 68 out of 103 “bad credit risks” correctly, but misclassified 35 of them. This information is usually much more informative than the overall misclassification rate, which simply tells us that the overall accuracy is 68.52%.

Display the ***Data miner recipes*** dialog again, and click the ***Next step*** button. A message is displayed with instructions to select only one model for deployment. Click ***OK***, and clear the check boxes adjacent to ***C&RT*** and ***Neural network***. We will deploy the ***Boosting Trees*** model that gave us the best predictive accuracy on the test sample. Now, click the ***Next step*** button again.

Deployment

The final ***Deployment*** step involves using the best model and applying it to new data in order to predict the “good or bad” customers. This step also provides the option for writing back the scoring information (classification probabilities computed by the best model, predicted classification, etc.) to the original input data file or database. This is extremely useful for deploying models on very large data sets to “score” databases.

On the ***Deployment*** tab, click the ***Data file for deployment*** button and open the *CreditScoring.sta* data file (located in the *Datasets* folder installed with *STATISTICA*). For demonstration purposes, we are using the same data file for deployment of the best model.



Data file for deployment: CreditScoring.sta

☐ Write back to deployed file

Variables to write back: none

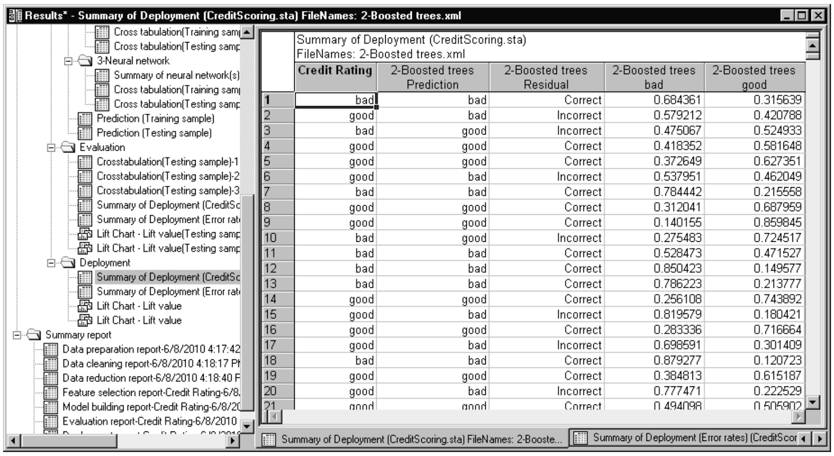
☐ Configure all steps

To complete this step:

1. Select new data file for computing predictions (deployment).
2. Click on Next Step
3. Click on Report to review predictions

Optionally write predictions

Click the ***Next step*** button to score this data file using the best model. The scored file with classifications and prediction probabilities (titled *Summary of Deployment*) is located in the *Deployment* folder in the project workbook as shown below.


Results* - Summary of Deployment (CreditScoring.sta) FileNames: 2-Boosted trees.xml
Cross tabulation(Training samp
Cross tabulation(Testing samp
3-Neural network
Summary of neural network(s)
Cross tabulation(Training samp
Cross tabulation(Testing samp
Prediction (Training sample)
Prediction (Testing sample)
Evaluation
Crosstabulation(Testing sample)-1
Crosstabulation(Testing sample)-2
Crosstabulation(Testing sample)-3
Summary of Deployment (CreditSc
Summary of Deployment (Error rat
Lift Chart - Lift value(Testing samp
Lift Chart - Lift value(Testing samp
Deployment
Summary of Deployment (CreditSc
Summary of Deployment (Error rat
Lift Chart - Lift value
Lift Chart - Lift value
Summary report
Data preparation report-6/8/2010 4:17:42
Data cleaning report-6/8/2010 4:18:17 PI
Data reduction report-6/8/2010 4:18:40 F
Feature selection report-Credit Rating-6/8,
Model building report-Credit Rating-6/8/20
Evaluation report-Credit Rating-6/8/2010
Summary of Deployment (CreditScoring.sta) FileNames: 2-Booste...
Summary of Deployment (Error rates) (CreditScor

Summary of Deployment (CreditScoring.sta)
FileNames: 2-Boosted trees.xml

	Credit Rating	2-Boosted trees Prediction	2-Boosted trees Residual	2-Boosted trees bad	2-Boosted trees good
1	bad	bad	Correct	0.684361	0.315639
2	good	bad	Incorrect	0.579212	0.420788
3	bad	good	Incorrect	0.475067	0.524933
4	good	good	Correct	0.418352	0.581648
5	good	good	Correct	0.372649	0.627351
6	good	bad	Incorrect	0.537951	0.462049
7	bad	bad	Correct	0.784442	0.215558
8	good	good	Correct	0.312041	0.687959
9	good	good	Correct	0.140155	0.859845
10	bad	good	Incorrect	0.275483	0.724517
11	bad	bad	Correct	0.528473	0.471527
12	bad	bad	Correct	0.850423	0.149577
13	bad	bad	Correct	0.786223	0.213777
14	good	good	Correct	0.256108	0.743892
15	good	bad	Incorrect	0.819579	0.180421
16	good	good	Correct	0.283336	0.716664
17	good	bad	Incorrect	0.698591	0.301409
18	bad	bad	Correct	0.879277	0.120723
19	good	good	Correct	0.384813	0.615187
20	good	bad	Incorrect	0.777471	0.222529
21	good	good	Correct	0.494098	0.505902

Summary

The purpose of this example is to demonstrate the efficiency of the data miner workflow implemented in *STATISTICA Data Miner Recipes.* With only a few clicks, the program will take you through the complete analytic process - from the definition of input data and analysis problem, through data cleaning and preparation and model building, all the way to final model selection and deployment.

Even though most of the computational complexities of data mining are resolved automatically in *STATISTICA Data Miner Recipes*, which enables you to move from problem definition to a solution very quickly even if you are a novice, the program will "apply and try" a large number of advanced data mining algorithms and automatically determine which approach is most successful.

Thus, the *STATISTICA Data Miner Recipes* methodology and user interface enables you to leverage the largest collection of data mining algorithms in a single package to solve your problems.

DATA MANAGEMENT

Example 1: Spreadsheet Formulas and Batch Formulas

You can define new variables for *STATISTICA* Spreadsheets in terms of other variables, sometimes referred to as variable transformations. Additionally you can verify data, transform data, and recode data on a single variable (as opposed to a set of transformation formulas, i.e., batch formulas). This is accomplished with spreadsheet formulas.

To access spreadsheet formulas, double-click on a variable header in a *STATISTICA* Spreadsheet to display the ***Variable*** specification dialog. The formula is entered into the ***Long name (label or formula with Functions)*** field (also called the formula editor) located at the bottom of the dialog. When you enter a long variable name in the formula editor that starts with an equal sign, *STATISTICA* recognizes it as a formula and will verify it for formal correctness.

The formula can reference other variables either by name (*MEASURE01*, *TIME*), or by absolute variable number using the *Vx* syntax, where *x* is the absolute variable number. For example, *V3* is variable number 3. *V0* has special meaning, and refers to the current case number.

Spreadsheet formulas are evaluated a case (row) at a time. For each case in the spreadsheet, the formula is evaluated, and references to the other variables are substituted with their values from the current case.

In *STATISTICA*, random access spreadsheet functions enable the formula to access variable values from other cases. A common example of this is the Lag function, which will reference a variable, and lag it forward or backward a certain number of cases.

The following table lists several spreadsheet formulas and their results.

Formula	Result
=contains(v1, "B12C")	Returns 1 if the text “B12C” is found in variable 1. Returns 0 if no match is found.
=(v1+v2+v3)/3	Computes the mean of the first three variables.
=(v0<=10)*1+(v0>10)*2	Recodes cases 1-10 as 1. The other cases are set to 2.
=((v1=1) AND (v2=5))*5	Returns the value of 5 if v1=1 and v2=5, otherwise set to 0.
=student(v4,15)	Returns probability density values of the Student’s *t* distribution based on the values of v4 and 15 degrees of freedom.
=cusum(v3)	Performs a cumulative sum of variable 3.
=v1+v2	Concatenates two text variables: If v1='A' and v2='B', then the result is 'AB'
=vnormal(rnd(1), 50, 3)	Generates random numbers from a Normal distribution (μ = 50, σ = 3)
=DTMonth(DTToday)	Returns number representing month of the parameter, e.g., 3 if it is currently March
=match(v1, 1, 0, 2, 0, v1)	Compares first value to a set of value/result pairs, returning the first result if the corresponding value matches. If no match, then a final default result is used. For example, returns 0 if v1 is 1 or 2, else returns v1.
=trunc((v0-1)/10)	Assigns consecutive integers to the consecutive sets of 10 cases (i.e., cases number 1 through 10 will be assigned 0, cases number 11-20 will be assigned 1, and so on

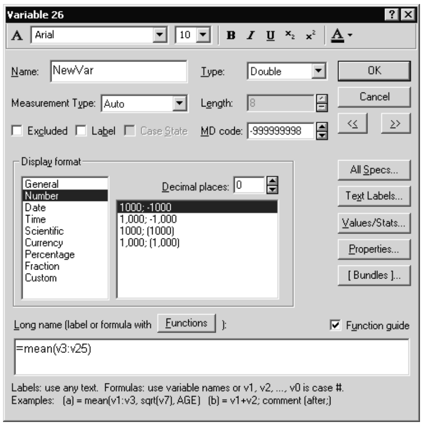
Note that you can click the Functions button in the ***Variable*** specification dialog to display the ***Function Browser*** dialog, which contains the complete list of formulas and operators (=, +, >, and, or, etc.).

Example: Spreadsheet Formula

Open the *Adstudy.sta* data file. We will create a new variable that is the mean of variables *3* through *25* (i.e., *MEASURE01* through *MEASURE23*).

Double-click on the first blank variable header (after variable 25). The ***Add Cases and/or Variables*** dialog will be displayed. Click the ***OK*** button to accept the default, which is to add one variable.

The ***Variable*** specification dialog for the new variable will be displayed. In the ***Display format*** group, select ***Number***. In the ***Long name*** field at the bottom of the dialog, enter*: =mean(v3:v25).*



Variable 26

Arial 10

Name: NewVar

Type: Double

Measurement Type: Auto

Length: 8

□ Excluded □ Label □ Case State

MD code: -999999998

Display format

Decimal places: 0

General
Number
Date
Time
Scientific
Currency
Percentage
Fraction
Custom

1000; -1000
1,000; -1,000
1000; (1000)
1,000; (1,000)

Long name (label or formula with Functions):

☑ Function guide

=mean(v3:v25)

Labels: use any text. Formulas: use variable names or v1, v2, ..., v0 is case #.
Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after;)

Click the ***OK*** button. A dialog will be displayed that informs you whether the formula is formally correct. Click the ***Yes*** button to proceed. The new variable is now filled with the mean of variables *3* through *25* for each case.

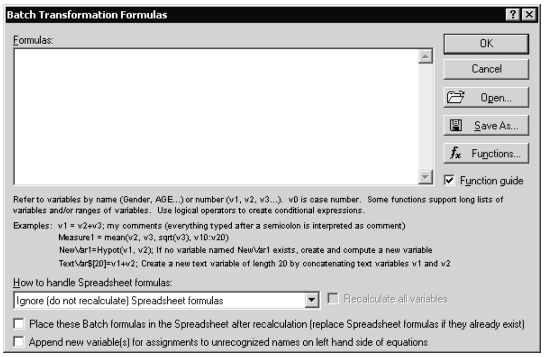
Since you can refer to variables by their names or their numbers, the formula we just created could also be expressed as: *=mean(MEASURE01:MEASURE23).*

Example: Batch Formulas

Spreadsheet formulas are useful for defining a formula for one variable at a time. However, there are many situations in which you need to evaluate several formulas for different variables simultaneously. This can be done with the batch formulas facilities in *STATISTICA*.

Open the *Characteristics.sta* data file. This data file contains information about patients in a study. For this example, we will 1) calculate patient Body Mass Index (BMI) and 2) convert height to centimeters (cm), and add these two variables to the data set.

On the ***Data*** tab, in the ***Transformations*** group, click ***Transforms*** to display the ***Batch Transformation Formulas*** dialog.


Batch Transformation Formulas
?
Formulas:
OK
Cancel
Open...
Save As...
Functions...
Function guide
How to handle Spreadsheet formulas:
Recalculate all variables

☑ Function guide

Refer to variables by name (Gender, AGE...) or number (v1, v2, v3...). v0 is case number. Some functions support long lists of variables and/or ranges of variables. Use logical operators to create conditional expressions.

```
Examples:  v1 = v2+v3; my comments (everything typed after a semicolon is interpreted as comment)
           Measure1 = mean(v2, v3, sqrt(v3), v10:v20)
           NewVar1=Hypot(v1, v2); If no variable named NewVar1 exists, create and compute a new variable
           TextVar$[20]=v1+v2; Create a new text variable of length 20 by concatenating text variables v1 and v2
```

How to handle Spreadsheet formulas: Ignore (do not recalculate) Spreadsheet formulas

□ Recalculate all variables

□ Place these Batch formulas in the Spreadsheet after recalculation (replace Spreadsheet formulas if they already exist)

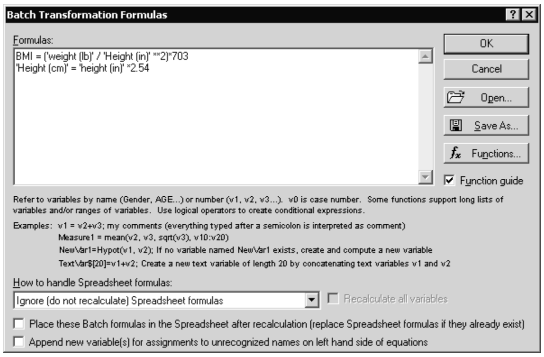
□ Append new variable(s) for assignments to unrecognized names on left hand side of equations

The only differences in syntax between the batch transformation formulas and the spreadsheet formulas is the support for multiple formulas in the batch option, and the fact that because the batch formulas are not attached to any specific variable (in fact they can be freely copied from data file to data file), they cannot start with an equal sign, but must have a target variable (e.g., *v1*=... or *Measure03*=...) so that *STATISTICA* knows to which variable each formula should apply. There is also an option to "distribute" all batch formulas into the respective variables in the spreadsheet and save them with the data file, effectively replacing the spreadsheet formulas (if there are any).

Following are the calculations used to calculate BMI and to convert Height (in) to centimeters, and the formulas to enter in the ***Batch Transformation*** dialog:

Calculation	Batch Transformation Dialog Entry
$BMI = \frac{weight\,(lb)}{height\,(in)^2} * 703$	BMI = ('weight (lb)' / 'Height (in)' **2)*703
$height\,(cm) = height\,(in) * 2.54$	'Height (cm)' = 'height (in)' *2.54

In the ***Formulas*** field, enter the list of transformation formulas to be applied to the active data spreadsheet. Separate each transformation formula by a return (press ENTER on your keyboard).



Formulas:

```
BMI = ('weight (lb)' / 'Height (in)' **2)*703
'Height (cm)' = 'height (in)' *2.54
```

☑ Function guide

Refer to variables by name (Gender, AGE...) or number (v1, v2, v3...). v0 is case number. Some functions support long lists of variables and/or ranges of variables. Use logical operators to create conditional expressions.

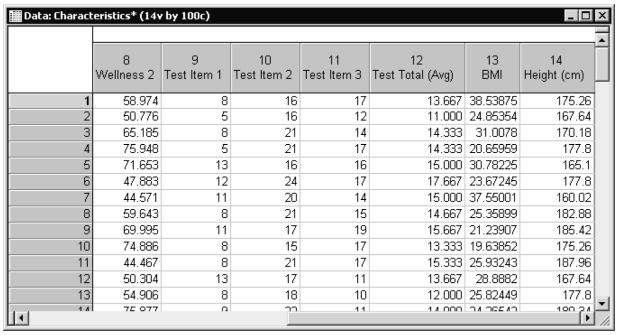
```
Examples:  v1 = v2+v3; my comments (everything typed after a semicolon is interpreted as comment)
           Measure1 = mean(v2, v3, sqrt(v3), v10:v20)
           NewVar1=Hypot(v1, v2); If no variable named NewVar1 exists, create and compute a new variable
           TextVar$[20]=v1+v2; Create a new text variable of length 20 by concatenating text variables v1 and v2
```

How to handle Spreadsheet formulas: Ignore (do not recalculate) Spreadsheet formulas ☐ Recalculate all variables

☐ Place these Batch formulas in the Spreadsheet after recalculation (replace Spreadsheet formulas if they already exist)

☐ Append new variable(s) for assignments to unrecognized names on left hand side of equations

Click the ***OK*** button in the ***Batch Transformation Formulas*** dialog. The ***Add New Variables?*** dialog will be displayed; click the ***Yes*** button to add the two new variables to the *Characteristics.sta* data file. A message will be displayed to inform you whether the expressions you entered in the ***Batch Transformation*** dialog are correct. If they are OK, click ***Yes*** to proceed. *STATISTICA* calculates the formulas and adds the two variables, *BMI* and *Height (cm)*, to the spreadsheet.


Data: Characteristics* (14v by 100c)

	8 Wellness 2	9 Test Item 1	10 Test Item 2	11 Test Item 3	12 Test Total (Avg)	13 BMI	14 Height (cm)
1	58.974	8	16	17	13.667	38.53875	175.26
2	50.776	5	16	12	11.000	24.85354	167.64
3	65.185	8	21	14	14.333	31.0078	170.18
4	75.948	5	21	17	14.333	20.65959	177.8
5	71.653	13	16	16	15.000	30.78225	165.1
6	47.883	12	24	17	17.667	23.67245	177.8
7	44.571	11	20	14	15.000	37.55001	160.02
8	59.643	8	21	15	14.667	25.35899	182.88
9	69.995	11	17	19	15.667	21.23907	185.42
10	74.886	8	15	17	13.333	19.63852	175.26
11	44.467	8	21	17	15.333	25.93243	187.96
12	50.304	13	17	11	13.667	28.8882	167.64
13	54.906	8	18	10	12.000	25.82449	177.8

StatSoft®

The options in the ***Batch Transformation Formulas*** dialog are particularly well suited (optimized) for transforming large data sets. The formulas will be evaluated one by one, in sequence, so that the results of one transformation in the list can serve as the input for the next. Thus, it is possible to create a new variable with one formula and then use that variable in subsequent formulas.

Click the ? button in the upper-right corner of the ***Batch Transformation Formulas*** dialog to display the *STATISTICA Electronic Manual* topic related to these options and links to various other topics containing examples of formulas and syntax rules.

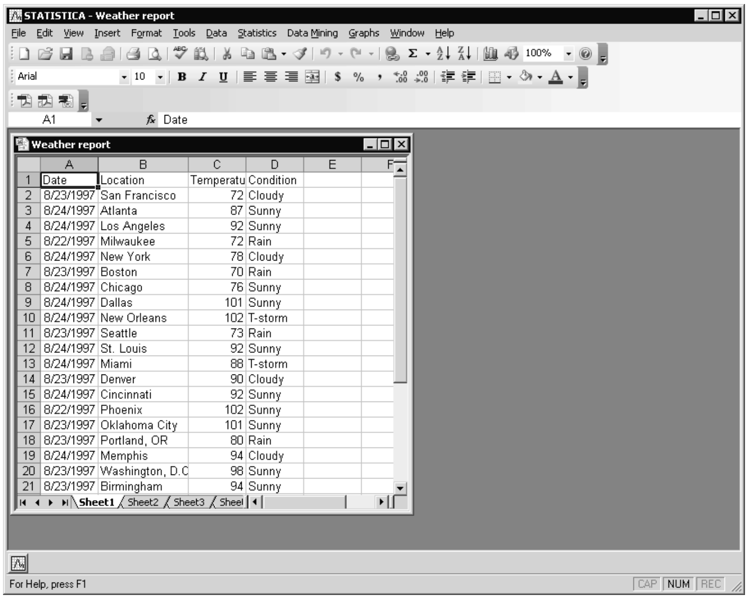
Example 2: Input Data Directly from Excel

In addition to using the traditional *STATISTICA* spreadsheet, you can open Excel files in a *STATISTICA* window and then perform analyses using the Excel file as your data source.

On the *STATISTICA* ***Home*** tab, in the ***File*** group, click the ***Open*** arrow and select ***Open Examples*** from the drop-down menu to display the ***Open a STATISTICA Data File*** dialog.

From the ***Files of type*** drop-down list at the bottom of the dialog, select ***Excel Files (*.xls;*xlsx;*.xlsm)***. Double-click the *Datasets* folder, and then select the *Weather report* data file, which is an Excel file. Click the ***Open*** button, and the ***Opening file*** dialog will be displayed.

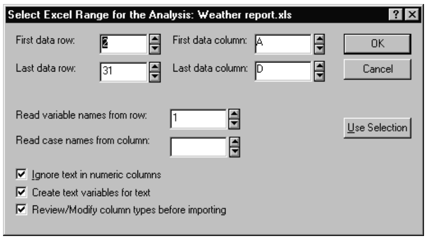
Click the ***Open as an Excel Workbook*** button, and the Excel file will be displayed. Note that when an Excel worksheet is opened in *STATISTICA*, the Excel and *STATISTICA* menus merge, enabling you to access key functionality for both applications.



STATISTICA - Weather report
File Edit View Insert Format Tools Data Statistics Data Mining Graphs Window Help
Arial 10
A1 Date
Weather report
Sheet1 Sheet2 Sheet3 Sheel
For Help, press F1
CAP NUM REC

	A	B	C	D	E	F
1	Date	Location	Temperatu	Condition		
2	8/23/1997	San Francisco	72	Cloudy		
3	8/24/1997	Atlanta	87	Sunny		
4	8/24/1997	Los Angeles	92	Sunny		
5	8/22/1997	Milwaukee	72	Rain		
6	8/24/1997	New York	78	Cloudy		
7	8/23/1997	Boston	70	Rain		
8	8/24/1997	Chicago	76	Sunny		
9	8/24/1997	Dallas	101	Sunny		
10	8/24/1997	New Orleans	102	T-storm		
11	8/23/1997	Seattle	73	Rain		
12	8/24/1997	St. Louis	92	Sunny		
13	8/24/1997	Miami	88	T-storm		
14	8/23/1997	Denver	90	Cloudy		
15	8/24/1997	Cincinnati	92	Sunny		
16	8/22/1997	Phoenix	102	Sunny		
17	8/23/1997	Oklahoma City	101	Sunny		
18	8/23/1997	Portland, OR	80	Rain		
19	8/24/1997	Memphis	94	Cloudy		
20	8/23/1997	Washington, D.C	98	Sunny		
21	8/23/1997	Birmingham	94	Sunny		

From the ***Statistics*** menu, select ***Basic Statistics/Tables***. The ***Select Excel Range for the Analysis*** dialog will be displayed.

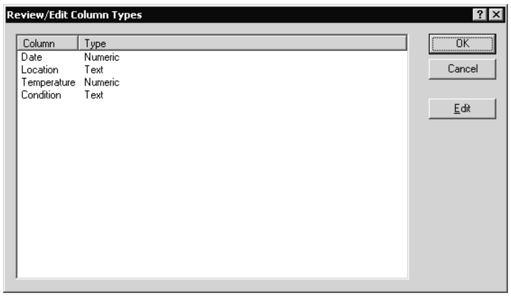


First data row: 2
First data column: A
Last data row: 31
Last data column: D
Read variable names from row: 1
Read case names from column:

☑ Ignore text in numeric columns
☑ Create text variables for text
☑ Review/Modify column types before importing

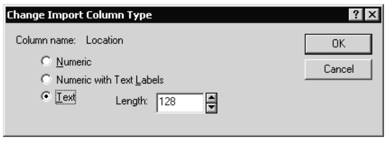
This dialog is displayed whenever you select a command from the ***Statistics***, ***Data Mining***, or ***Graphs*** menu after opening an Excel worksheet in the *STATISTICA* application. Note that *STATISTICA* has determined the logical specifications, but these options can be changed if necessary. When variable names are not included with the Excel worksheet, *STATISTICA* will assign variable names: *Var1*, *Var2*, *Var3*, etc. As with *STATISTICA* spreadsheets, all values in a column will be used for the selected analysis unless case selection conditions are specified.

For this example, click the ***OK*** button in the ***Select Excel Range for the Analysis*** dialog to accept the defaults; the dialog will close, and the ***Review/Edit Column Types*** dialog will be displayed.



Column	Type
Date	Numeric
Location	Text
Temperature	Numeric
Condition	Text

In *STATISTICA*, you can define the data type for the specific columns. Data types include numeric, text, mixed numeric and text, and missing data. Empty cells in an Excel worksheet are always treated as missing data, and when a numeric column contains text values, those values are also treated as missing data. *STATISTICA* provides default data types for all columns based on the first few rows of data (in fact, you can clear the ***Review/Modify column types before importing*** check box in the ***Select Excel Range for the Analysis*** dialog before clicking ***OK*** in that dialog, and the ***Review/Edit Column Types*** dialog will not be displayed). However, you can change the default types if needed: select the name of the column you want to change and click the ***Edit*** button (or double-click on the name of the column you want to change) to display the ***Change Import Column Type*** dialog, where you can specify the type you prefer.



Column name: Location

○ Numeric
○ Numeric with Text Labels
● Text Length: 128

For this example we will accept the defaults, so click the ***Cancel*** button in the ***Change Import Column Type*** dialog, and click the ***OK*** button in the ***Review/Edit Column Types*** dialog. After you click ***OK***, the Startup Panel for the selected analysis or graph will be displayed (in this example, the ***Basic Statistics and Tables*** Startup Panel), and you can proceed with the analysis as usual.

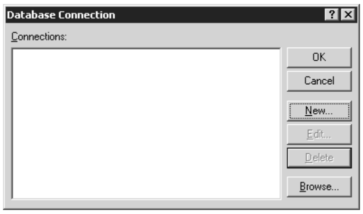
Example 3: Accessing Data Directly from a SQL Server Database

STATISTICA provides access to virtually all databases (including many large system databases such as Oracle, Sybase, etc.) via *STATISTICA Query*, accessible from either the ***Home*** tab (in the ***File*** group, click the ***Open*** arrow to access the ***Open External Data*** submenu) or the ***Data*** tab (in the ***Manage*** group, click ***External Data***). For importing data from a database directly into a *STATISTICA* Spreadsheet so that it can be saved, the tool to use is *STATISTICA Query*.

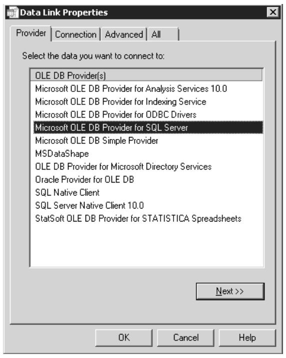
With *STATISTICA Query*, you can easily access data using OLE DB conventions. OLE DB is a database architecture [based on the Component Object Model (COM)] that provides universal data integration over an enterprise's network, from mainframe to desktop, regardless of the data type.

STATISTICA Query supports multiple database tables; specific records (rows of tables) can be selected by entering SQL statements. *STATISTICA Query* automatically builds the SQL statement for you as you select the components of the query via a simple graphical interface and/or intuitive menu options and dialogs. Hence, an extensive knowledge of SQL is not necessary in order for you to create advanced and powerful queries of data in a quick and straightforward manner. Multiple queries based on one or many different databases can also be created to return data to an individual spreadsheet; hence, you can maintain connections to multiple external databases simultaneously.

For this example, create a new database query: select the ***Home*** tab, and in the ***File*** group, click the ***Open*** arrow. From the drop-down list, select ***Open External Data - Create Query***. *STATISTICA Query* will start, and the ***Database Connection*** dialog will be displayed.



From this dialog, you can choose existing database connections or define new ones. For this example, we'll create a new database connection, so click the ***New*** button to display the ***Data Link Properties*** dialog.

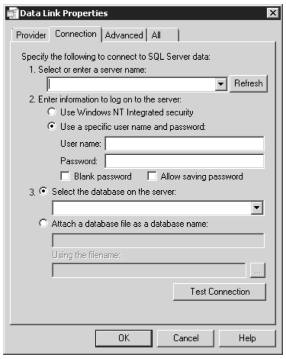


Select the data you want to connect to:

OLE DB Provider(s)
Microsoft OLE DB Provider for Analysis Services 10.0
Microsoft OLE DB Provider for Indexing Service
Microsoft OLE DB Provider for ODBC Drivers
Microsoft OLE DB Provider for SQL Server
Microsoft OLE DB Simple Provider
MSDataShape
OLE DB Provider for Microsoft Directory Services
Oracle Provider for OLE DB
SQL Native Client
SQL Server Native Client 10.0
StatSoft OLE DB Provider for STATISTICA Spreadsheets

You can choose either the OLE DB provider that was supplied by your database vendor, or one of the Microsoft default OLE DB providers that is compatible with your database system.

For this example, we'll use the Northwind sample database installed with Microsoft SQL Server, so select ***Microsoft OLE DB Provider for SQL Server*** and click the ***Next >>*** button. The ***Data Link Properties*** dialog - ***Connection*** tab will be displayed.



Specify the following to connect to SQL Server data:

1. Select or enter a server name:
2. Enter information to log on to the server:
   ○ Use Windows NT Integrated security
   ● Use a specific user name and password:
   User name:
   Password:
   □ Blank password □ Allow saving password
3. ● Select the database on the server:
   ○ Attach a database file as a database name:
   Using the filename:

Select a server from the ***Select or enter a server name*** drop-down list.

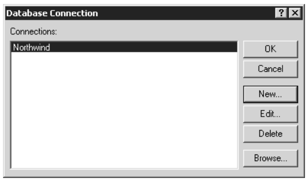
Then, select the log on option button appropriate to your SQL Server Northwind database installation. Select either the ***Use Windows NT Integrated security*** option button, or select the ***Use a specific user name and password*** option button and enter a ***User name*** and ***Password*** in the respective fields.

Next, select *Northwind* from the ***Select the database on the server*** drop-down list.

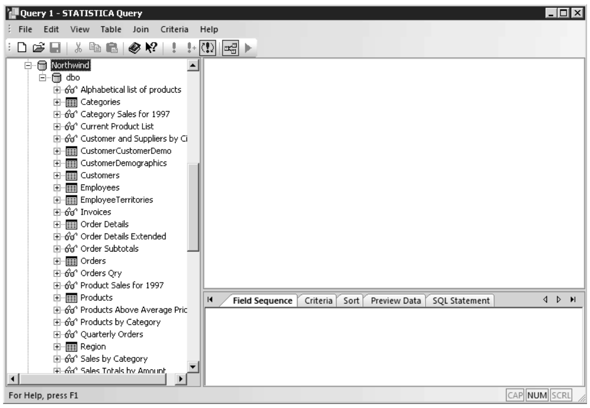
Click the ***Test Connection*** button to attempt a connection to the specified data source. If the connection fails, ensure that the settings are correct. For example, spelling errors and case sensitivity can cause failed connections. If the connection succeeds, click the ***OK*** button in the message dialog.

Click ***OK*** in the ***Data Link Properties*** dialog to display the ***Add a Database Connection*** dialog. Enter *Northwind* in the ***Name*** edit box, and click ***OK***.

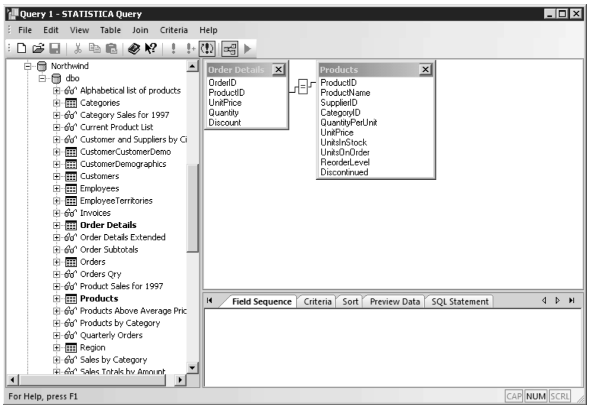
The ***Database Connection*** dialog will be displayed again, with the new *Northwind* connection defined.



Select this connection, and click ***OK***. The *STATISTICA Query* window will be displayed, with all the database tables in the tree view on the left.

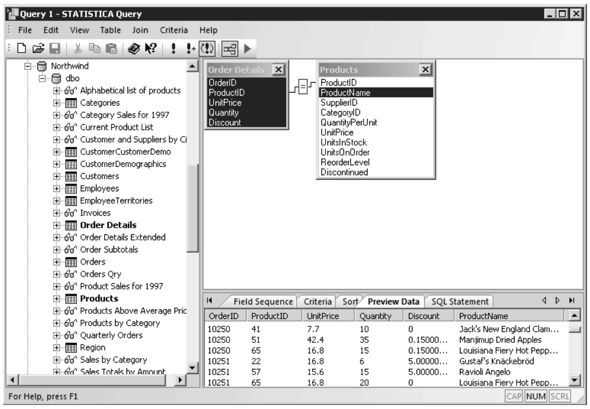


Right-click on the *Order Details* table, and from the shortcut menu, select ***Add*** to add the table to the table view pane (the upper-right pane in the *STATISTICA Query* window). Then, right-click on the *Products* table, and add it to the table view pane. Since both tables contain the *ProductID* field, *STATISTICA Query* automatically joins the two tables on this key.



To select the fields to include in the query, right click in the *Order Details* table in the table view pane, and from the shortcut menu, select ***Select All Fields***. In the *Product* table, select the *ProductName* field.

Click the ***Preview Data*** tab in the lower-right pane to display a preview of the query.

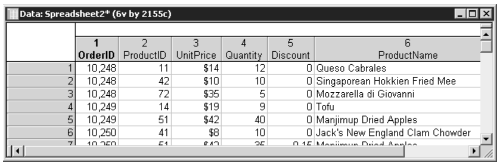


OrderID	ProductID	UnitPrice	Quantity	Discount	ProductName
10250	41	7.7	10	0	Jack's New England Clam...
10250	51	42.4	35	0.15000...	Manjimup Dried Apples
10250	65	16.8	15	0.15000...	Louisiana Fiery Hot Pepp...
10251	22	16.8	6	5.00000...	Gustaf's Knäckebröd
10251	57	15.6	15	5.00000...	Ravioli Angelo
10251	65	16.8	20	0	Louisiana Fiery Hot Pepp...

Click the ***SQL Statement*** tab to display the SQL Statement generated by the query.

To return the data to a *STATISTICA* Spreadsheet, click the green arrow on the *STATISTICA Query* toolbar. The ***Returning External Data to Spreadsheet*** dialog will be displayed, where you can control whether the query will be placed into a new or current spreadsheet and adjust other query parameters. Select the ***New Spreadsheet*** option button, and click the ***Run Now*** button to run the query. If the

Connect to OLE DB Provider dialog is displayed, click the ***OK*** button. After a few moments, the data is returned to the *STATISTICA* Spreadsheet.



Data: Spreadsheet2* (6v by 2155c)

	1 OrderID	2 ProductID	3 UnitPrice	4 Quantity	5 Discount	6 ProductName
1	10,248	11	$14	12	0	Queso Cabrales
2	10,248	42	$10	10	0	Singaporean Hokkien Fried Mee
3	10,248	72	$35	5	0	Mozzarella di Giovanni
4	10,249	14	$19	9	0	Tofu
5	10,249	51	$42	40	0	Manjimup Dried Apples
6	10,250	41	$8	10	0	Jack's New England Clam Chowder
7	10,250	51	$42	35	0.15	Manjimup Dried Apples

Now the data can be analyzed with any of the *STATISTICA* tools. Note that the spreadsheet retains the database connection, and you can re-run the query at any time: select the ***Data*** tab, and in the ***Manage*** group click ***External Data***. Select ***Refresh Data*** from the drop-down menu. You can also press F5 on your keyboard when the spreadsheet is open.

Example 4: Data Preparation – Cleaning and Filtering

Summary of Options for Data Filtering/Recoding

In practice, most of the time required to complete a data analysis or data mining project is spent on the preparation of data. Sometimes as much as 90% of all time and effort required to complete a project is related to the proper cleaning and preparation of the data.

When building prediction models using data mining tools, or even when just computing simple descriptive statistics (averages, frequency distributions), results of analyses can be very misleading if, for example, large numbers of duplicate records are included (e.g., the same part numbers are recorded multiple times) or the data include outliers or miscoded values (outside the valid data ranges) or excessive numbers of missing (blank) data.

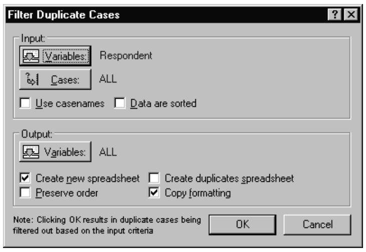
On the ***Data*** tab, in the ***Transformations*** group, click ***Filter/Recode*** to display a drop-down menu containing commands to address such data quality issues quickly and effectively so that meaningful and valid data analyses or data mining projects can be completed in less time.

Filter Duplicate Cases

Use this option when you suspect that your data file may contain duplicate records (e.g., duplicate/identical customer records).

For example, suppose that in an analysis of customer records, to identify typical customer demographics ("profiles"), you want to count each customer only once; however, your customer database is organized by transactions, so each customer may appear multiple times. In this case, you can use the ***Filter Duplicate Cases*** options to create a data file for the analyses containing only unique records (i.e., where each customer ID is listed only once).

Duplicate information example. Open the *Duplicates.sta* data file. From the ***Filter/Recode*** menu, select ***Filter Duplicate Cases*** to display the ***Filter Duplicate Cases*** dialog. In the ***Input*** group box, the ***Variables*** option is used to specify the basis of distinction for duplicates. Click the ***Variables*** button, and in the variable selection dialog, select *Respondent* so that all respondents will be checked for duplicates. Click ***OK*** in the variable selection dialog to return to the ***Filter Duplicate Cases*** dialog.



Filter Duplicate Cases

Input:

Variables: Respondent

Cases: ALL

□ Use casenames □ Data are sorted

Output:

Variables: ALL

☑ Create new spreadsheet □ Create duplicates spreadsheet

□ Preserve order ☑ Copy formatting

Note: Clicking OK results in duplicate cases being filtered out based on the input criteria

In the ***Input*** group box, click the ***Cases*** button to display the ***Spreadsheet Case Selection Conditions*** dialog, which contains options to select only specified observations or cases for the de-duping operations. In this example, we will filter all the cases, so click the ***Cancel*** button in the ***Spreadsheet Case Selection Conditions*** dialog.

The ***Use casenames*** check box is cleared by default; we will leave this option as is for this example. When this check box is selected, case names are used as one of the bases for distinction, i.e., *STATISTICA* will treat as duplicates any cases that have the same case name (provided the cases match on any other specified variables as well). When the check box is cleared, duplicate case names are ignored.

Clear the ***Data are sorted*** check box (because the current data file has not been sorted – when you have an extremely large data file, it is more efficient to sort the data first).

In the ***Output*** group box, verify that all variables are selected (*ALL* will be adjacent to the ***Variables*** button). This option is used to select the variables in the input spreadsheet that will be included in the output (filtered) spreadsheet; the default is *ALL*.

Verify that the ***Create new spreadsheet*** check box is selected (the default), and select the ***Create duplicates spreadsheet*** check box. Leave the last two options at their defaults: the ***Preserve order*** check box is cleared [the new spreadsheets will be sorted by the variable(s) that were selected as the basis of distinction, in this example, *Respondent*], and the ***Copy formatting*** check box is selected. Click ***OK***.

Two new spreadsheets will be generated. One of the spreadsheets is 10v by 51c (10 variables by 51 cases) and contains the respondents from the original spreadsheet excluding the duplications. The other spreadsheet is 10v by 9c and contains the duplicate respondents that were extracted from the original spreadsheet.

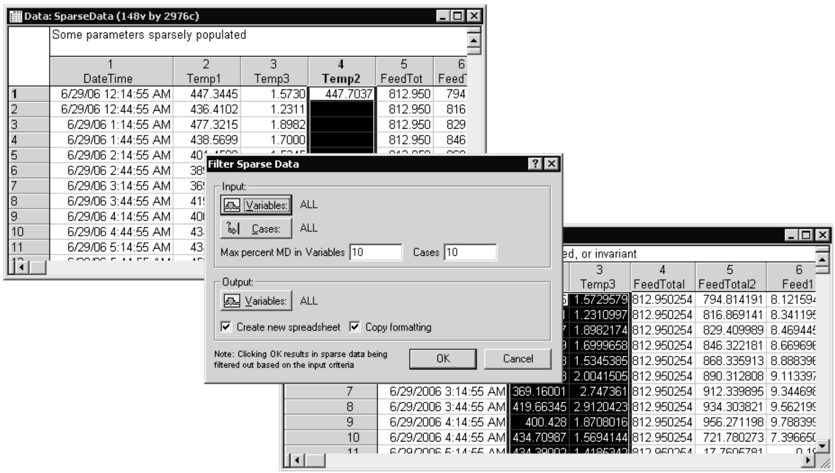
Look at the original spreadsheet, *Duplicates.sta*, and notice that some of the variable headers – *Respondent*, *State*, and *Colors* – are formatted differently. Then look at the two new spreadsheets; the variable headers for *Respondent*, *State*, and *Colors* have the same formatting in all three spreadsheets. *STATISTICA* uses sub-setting to create the new spreadsheets and ensures that variable properties of the parent spreadsheet are maintained in the child spreadsheets.

Now, close the two new spreadsheets, but leave the *Duplicates.sta* spreadsheet open. Notice that it is 10v by 60c. From the ***Filter/Recode*** menu, select ***Filter Duplicate Cases*** to display the ***Filter Duplicate Cases*** dialog again. In the ***Input*** group box, click the ***Variables*** button, and in the variable selection dialog, select *Respondent* and click ***OK***. In the ***Input*** group box, clear the ***Data are sorted*** check box. In the ***Output*** group box, clear the ***Create new spreadsheet*** check box. Click ***OK***. The dialog closes and, instead of creating a new spreadsheet with the duplicates excluded, the *Duplicates.sta* spreadsheet is modified. All duplicate cases are removed from it; it now has 10v by 51c.

Note that the filter duplicate cases functionality does not use case sensitivity (upper-case, lower-case letters) for a comparison of uniqueness, i.e., if you have two respondents – C. Barrett and C. BARRETT – the second respondent will be excluded.

Filter Sparse Data

It is not uncommon that some variables (parameters, or data fields) available for (for example) predictive modeling have very few valid data. For example, in a customer database self-reported (by customers) *Income* may be recorded; however, very few customers actually volunteered their current incomes, so most of the data (in that field of the database) is blank (or missing). In manufacturing data, a data field may exist to record a specific parameter, but the sensor might be faulty for an extended period of time, recording mostly missing (invalid) data.



Data: SparseData (148v by 2976c)

Some parameters sparsely populated

	1 DateTime	2 Temp1	3 Temp3	4 Temp2	5 FeedTot	6 Feed
1	6/29/06 12:14:55 AM	447.3445	1.5730	447.7037	812.950	794
2	6/29/06 12:44:55 AM	436.4102	1.2311		812.950	816
3	6/29/06 1:14:55 AM	477.3215	1.8982		812.950	829
4	6/29/06 1:44:55 AM	438.5699	1.7000		812.950	846
5	6/29/06 2:14:55 AM	401				
6	6/29/06 2:44:55 AM	38				
7	6/29/06 3:14:55 AM	36				
8	6/29/06 3:44:55 AM	41				
9	6/29/06 4:14:55 AM	40				
10	6/29/06 4:44:55 AM	43				
11	6/29/06 5:14:55 AM	43				

Filter Sparse Data

Input:

Variables: ALL

Cases: ALL

Max percent MD in Variables 10 Cases 10

Output:

Variables: ALL

☑ Create new spreadsheet ☑ Copy formatting

Note: Clicking OK results in sparse data being filtered out based on the input criteria

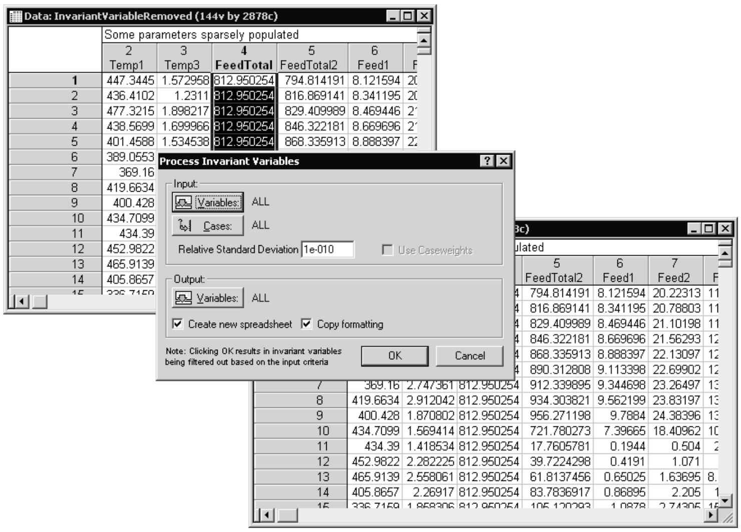
ed, or invariant

			3 Temp3	4 FeedTotal	5 FeedTotal2	6 Feed1
			1.5729579	812.950254	794.814191	8.12159
			1.2310997	812.950254	816.869141	8.34119
			1.8982174	812.950254	829.409989	8.46944
			1.6999658	812.950254	846.322181	8.66969
			1.5345385	812.950254	868.335913	8.88839
			2.0041505	812.950254	890.312808	9.11339
7	6/29/2006 3:14:55 AM	369.16001	2.747361	812.950254	912.339895	9.34469
8	6/29/2006 3:44:55 AM	419.66345	2.9120423	812.950254	934.303821	9.56219
9	6/29/2006 4:14:55 AM	400.428	1.8708016	812.950254	956.271198	9.78839
10	6/29/2006 4:44:55 AM	434.70987	1.5694144	812.950254	721.780273	7.39665

Including such “sparsely populated” (with data) variables in an analysis may lead to erroneous results, or prevent you from building predictive models altogether (depending on how the missing data are handled later in the analyses). Therefore, you may want to identify such sparse variables ahead of time using the ***Filter Sparse Data*** options (accessible from the ***Filter/Recode*** menu located on the ***Data*** tab in the ***Transformations*** group), and eliminate them from subsequent consideration.

Process Invariant Variables

A similar (to the sparse-data case) data quality issue that often occurs, in particular in industrial manufacturing (process) data, is that some variables (parameters) that are recorded and included in the analyses are invariant, i.e., all values are the same.



Data: InvariantVariableRemoved (144v by 2878c)

Some parameters sparsely populated

	2 Temp1	3 Temp3	4 FeedTotal	5 FeedTotal2	6 Feed1
1	447.3445	1.572958	812.950254	794.814191	8.121594
2	436.4102	1.2311	812.950254	816.869141	8.341195
3	477.3215	1.898217	812.950254	829.409989	8.469446
4	438.5699	1.699966	812.950254	846.322181	8.669696
5	401.4588	1.534538	812.950254	868.335913	8.888397
6	389.0553				
7	369.16				
8	419.6634				
9	400.428				
10	434.7099				
11	434.39				
12	452.9822				
13	465.9139				
14	405.8657				

Process Invariant Variables

Input:

Variables: ALL

Cases: ALL

Relative Standard Deviation 1e-010

□ Use Caseweights

Output:

Variables: ALL

☑ Create new spreadsheet ☑ Copy formatting

Note: Clicking OK results in invariant variables being filtered out based on the input criteria

				5 FeedTotal2	6 Feed1	7 Feed2
1				794.814191	8.121594	20.22313
2				816.869141	8.341195	20.78803
3				829.409989	8.469446	21.10198
4				846.322181	8.669696	21.56293
5				868.335913	8.888397	22.13097
6				890.312808	9.113398	22.69902
7	369.16	2.747361	812.950254	912.339895	9.344698	23.26497
8	419.6634	2.912042	812.950254	934.303821	9.562199	23.83197
9	400.428	1.870802	812.950254	956.271198	9.7884	24.38396
10	434.7099	1.569414	812.950254	721.780273	7.39665	18.40962
11	434.39	1.418534	812.950254	17.7605781	0.1944	0.504
12	452.9822	2.282225	812.950254	39.7224298	0.4191	1.071
13	465.9139	2.558061	812.950254	61.8137456	0.65025	1.63695
14	405.8657	2.26917	812.950254	83.7836917	0.86895	2.205

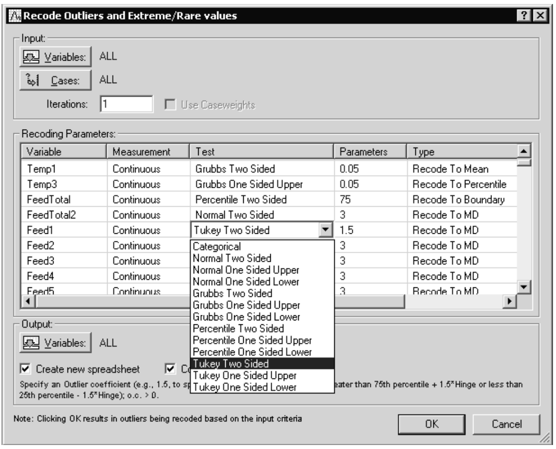
Such variables are not useful for predictive modeling, and the ***Process Invariant Variables*** options (accessible from the ***Filter/Recode*** menu located on the ***Data*** tab in the ***Transformations*** group) enable you to identify those variables automatically, and exclude them from further analyses.

Recode Outliers

Extreme data values or outliers can greatly affect various analyses and cause poor accuracy of prediction (data mining) models. There is no formal definition of what constitutes an “outlier” or “extreme value,” and *STATISTICA’*s graphical tools may provide the best way to review data to identify such unusual observations (e.g., you could create box plots of the key variables to identify extreme observations and brush or flag them in the data).

To automatically process lists of variables to identify and remove outliers, the ***Recode Outliers*** options (accessible from the ***Filter/Recode*** menu located on the

Data tab in the ***Transformations*** group) provide several tests for outliers (approaches for identifying extreme values).



Recode Outliers and Extreme/Rare values

Input:

Variables: ALL

Cases: ALL

Iterations: 1 ☐ Use Caseweights

Recoding Parameters:

Variable	Measurement	Test	Parameters	Type
Temp1	Continuous	Grubbs Two Sided	0.05	Recode To Mean
Temp3	Continuous	Grubbs One Sided Upper	0.05	Recode To Percentile
FeedTotal	Continuous	Percentile Two Sided	75	Recode To Boundary
FeedTotal2	Continuous	Normal Two Sided	3	Recode To MD
Feed1	Continuous	Tukey Two Sided	1.5	Recode To MD
Feed2	Continuous		3	Recode To MD
Feed3	Continuous		3	Recode To MD
Feed4	Continuous		3	Recode To MD
Feed5	Continuous		3	Recode To MD

Output:

Variables: ALL

☑ Create new spreadsheet ☑ C

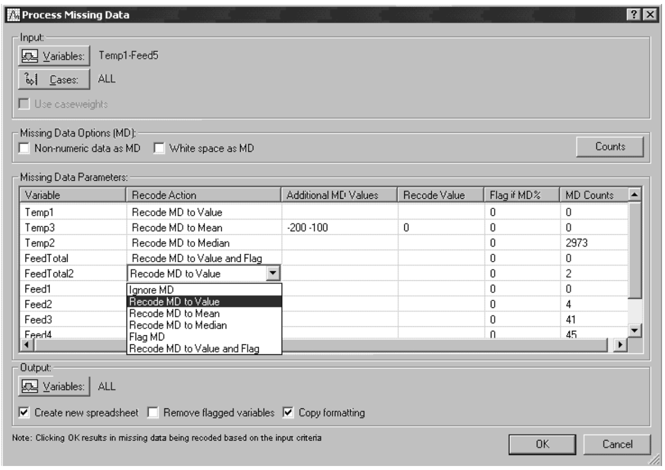
Specify an Outlier coefficient (e.g., 1.5, to sp ... eater than 75th percentile + 1.5*Hinge or less than 25th percentile - 1.5*Hinge); o.c. > 0.

Note: Clicking OK results in outliers being recoded based on the input criteria

Outliers can be recoded to missing data or to valid data values (e.g., to the respective percentile boundary values, etc.).

Process Missing Data

Missing data or invalid data values must obviously be dealt with in a manner that is consistent with the goals of the analyses. In some cases, missing or invalid data may themselves provide useful information about a process or variable of interest. For example, in marketing research, it is common that respondents will refuse to provide detailed personal information regarding their health, financial status (e.g., savings), etc., and such refusal itself may be correlated with other significant variables of interest (e.g., refusal to answer questions related to income may itself be a good indicator of high income, if indeed wealthier individuals in the survey tended not to answer those questions).



Process Missing Data

Input:

Variables: Temp1-Feed5

Cases: ALL

□ Use caseweights

Missing Data Options (MD):

□ Non-numeric data as MD □ White space as MD

Missing Data Parameters:

Variable	Recode Action	Additional MD Values	Recode Value	Flag if MD%	MD Counts
Temp1	Recode MD to Value			0	0
Temp3	Recode MD to Mean	-200 -100	0	0	0
Temp2	Recode MD to Median			0	2973
FeedTotal	Recode MD to Value and Flag			0	0
FeedTotal2	Recode MD to Value			0	2
Feed1				0	0
Feed2				0	4
Feed3				0	41
Feed4				0	45

Output:

Variables: ALL

☑ Create new spreadsheet □ Remove flagged variables ☑ Copy formatting

Note: Clicking OK results in missing data being recoded based on the input criteria

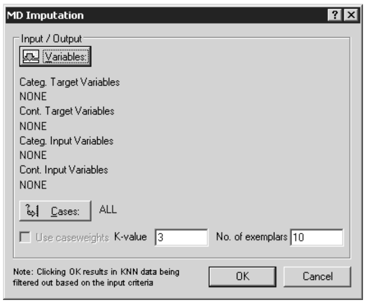
The ***Process Missing Data*** options (accessible from the ***Filter/Recode*** menu located on the ***Data*** tab in the ***Transformations*** group) enable you to recode missing data flexibly, define multiple missing data values or codes for a single variable (which can then be recoded to the variable missing data code), or just to flag variables that have more than a certain percentage of missing data.

Imputation of Missing Data (*k*-Nearest Neighbor)

It is often not clear how best to recode missing data, and in fact, sometimes by recoding missing data for a particular variable to a specific value (e.g., the mean), the final results may be biased. For example, suppose in a survey all respondents who refuse to report their income tend to be in the higher income bracket. In that case, assigning the mean-income to those individuals (i.e., recoding missing data for variable *Income* to the mean income for the whole sample) may yield highly misleading results.

STATISTICA includes a very efficient method (applicable to very large data sets and databases) for replacing missing data with valid data values that are consistent with the other observations in the sample. Details regarding the *k-nearest neighbor* method and algorithm are provided in the *Electronic Help* for the *Machine Learning* module of *STATISTICA Data Miner*.

In short, using the ***MD Imputation*** options (accessible from the ***Filter/Recode*** menu), in a first pass through the data, the *k*-nearest neighbor algorithm will select a (smaller) sample from all available data. In the second pass through the data file, when missing data are encountered, they are replaced with valid (observed) values found in similar observations in the smaller sample (with respect to all other variables that were selected). So to continue this example, if indeed higher-income respondents are less likely to report this fact, but *do* report other indicators of high-income (e.g., ownership of a luxury car, more square footage of their home, etc.), then the *k*-nearest neighbor algorithm will accurately assign those individuals (who failed to report their income) to the high-income bracket.



Categ. Target Variables
NONE
Cont. Target Variables
NONE
Categ. Input Variables
NONE
Cont. Input Variables
NONE

Cases: ALL

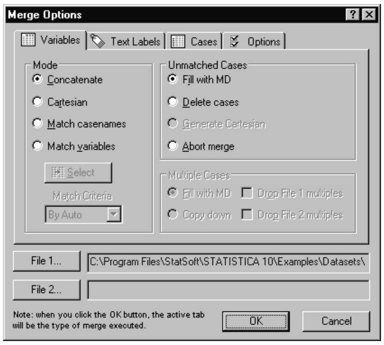
□ Use caseweights K-value 3 No. of exemplars 10

Note: Clicking OK results in KNN data being filtered out based on the input criteria

The *k*-nearest neighbor algorithm is fast and efficient, and provides an effective method for replacing missing data in the input file with “reasonable guesses” based on similar data points in the sample. This approach does not make any particular assumptions about the nature of the relationships between variables (i.e., require that a “model” be estimated for each variable to predict missing data values), but simply uses the observed data as the model.

Merge Data Files

The *STATISTICA* ***Merge Options*** dialog enables you to merge two data files either by the variables or by the cases so that you can centralize all of the observations to one table. Select the ***Data*** tab, and in the ***Manage*** group, click ***Merge*** to display the ***Merge Options*** dialog.



Mode
● Concatenate
○ Cartesian
○ Match casenames
○ Match variables
Match Criteria: By Auto

Unmatched Cases
● Fill with MD
○ Delete cases
○ Generate Cartesian
○ Abort merge

Multiple Cases
● Fill with MD ☐ Drop File 1 multiples
○ Copy down ☐ Drop File 2 multiples

File 1... C:\Program Files\StatSoft\STATISTICA 10\Examples\Datasets\
File 2...

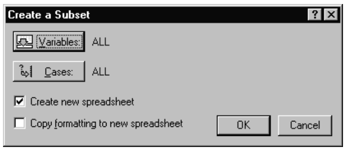
Note: when you click the OK button, the active tab will be the type of merge executed.

Click the Help ? button in the upper-right corner of the dialog to access Help topics describing all the options in this dialog.

Creating Subsets

If you have a large spreadsheet, you may want to create a new spreadsheet containing a specified subset of the current spreadsheet. For example, open *Boston2.sta*. This data set contains over a thousand cases. We want to extract housing tracts with low median prices.

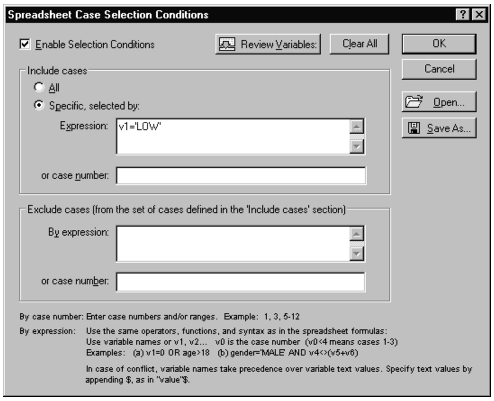
Select the ***Data*** tab, and in the ***Manage*** group, click ***Subset*** to display the ***Create a Subset*** dialog.



Variables: ALL
Cases: ALL
☑ Create new spreadsheet
☐ Copy formatting to new spreadsheet

Click the ***Cases*** button to display the ***Spreadsheet Case Selection Conditions*** dialog, which contains options to create conditions to define the selection of cases to be considered for the sample.

Select the ***Enable Selection Conditions*** check box to activate the options, and then select the ***Specific, selected by*** option button in the ***Include cases*** group box to specify which cases to include in the analysis. Type *v1='LOW'* in the ***Expression*** text box.



☑ Enable Selection Conditions

Include cases

○ All

● Specific, selected by:

Expression: v1='LOW'

or case number:

Exclude cases (from the set of cases defined in the 'Include cases' section)

By expression:

or case number:

By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12

By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)

In case of conflict, variable names take precedence over variable text values. Specify text values by appending $, as in "value"$.

Click the ***OK*** button to set the selection conditions and return to the ***Create a Subset*** dialog, and click the ***OK*** button in this dialog to create the new spreadsheet.

The resultant spreadsheet contains 334 cases (instead of the original 1,012 cases) and all 15 variables from the original spreadsheet. For the *PRICE* variable, all cases have a value of *LOW*.

Example 5: Using *STATISTICA ETL* (*Extract, Transform, and Load*)

The *STATISTICA ETL (Extract, Transform, and Load)* module provides unique capabilities for processing and merging data, in particular, process data that are difficult to manage using standard database tools. *ETL* automates the process of validating and aligning multiple diverse data sources into a single source suitable for ad-hoc or automated analyses.

ETL offers two options for aligning data: *Time-indexed*, which aggregates data from multiple data sources based on a date/time stamp variable and aligns data by minute, hour, day, week, month, quarter, or year; and *ID-based*, which aggregates data from multiple data sources based on an identifier variable and an optional time variable, and optionally aligns data by *N* equal intervals or *N* user-specified intervals.

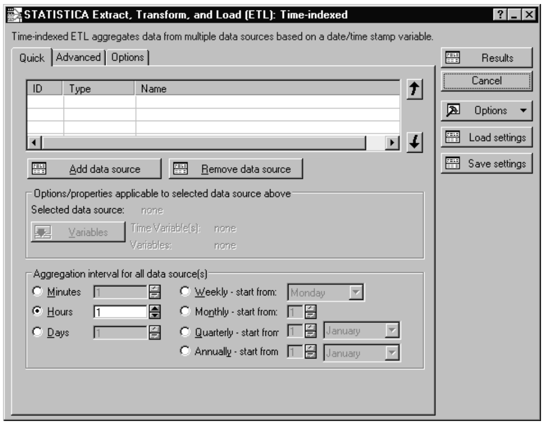
This example illustrates how the *ETL* module handles stock-related data sets with different time intervals. Stocks are bought and sold at varying prices throughout each day. Microsoft (ticker MSFT) and Oracle (ticker ORCL) are software companies that trade on the NASDAQ electronic stock exchange. In this example, we will compare data sets containing historical stock prices with different date/time stamps. The first set contains daily Microsoft price quotes from NASDAQ, while the second set contains weekly Oracle price quotes from another source.

Open *MicrosoftPrices.sta* and *OraclePrices.sta*: on the ***Home*** tab in the ***File*** group, click the ***Open*** arrow. From the drop-down menu, select ***Open Examples*** to display the ***Open a STATISTICA Data File*** dialog. Double-click on the *Datasets* folder, select *MicrosoftPrices.sta* and *OraclePrices.sta*, and click the ***Open*** button.

Both data files contain the following columns (variables): *DATE* - the day on which a trade takes place; *OPEN* - opening price for the day, first trade of the day; *HIGH* - the highest price of the day; *LOW* - the lowest price of the day; *CLOSE* - closing price for the day, last trade of the day; and *VOLUME* - the daily number of traded shares of a security.

However, they have different date ranges: Microsoft - 10/22/2007-01/04/2008; Oracle - 10/18/2007-12/28/2007. In order to compare the data, the ranges will need to be aligned.

Select the ***Data*** tab. In the ***Manage*** group, click ***External Data***, and select ***Time-indexed Process Data*** from the ***Extract, Transform, and Load (ETL)*** submenu. The ***STATISTICA Extract, Transform, and Load (ETL): Time-indexed*** Startup Panel is displayed.



STATISTICA Extract, Transform, and Load (ETL): Time-indexed

Time-indexed ETL aggregates data from multiple data sources based on a date/time stamp variable.

ID	Type	Name

Options/properties applicable to selected data source above

Selected data source: none

Time Variable(s): none

Variables: none

Aggregation interval for all data source(s)

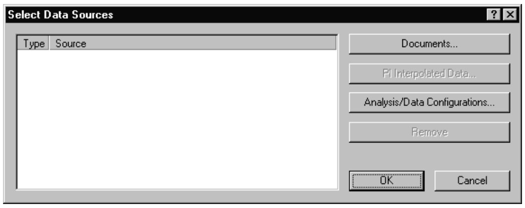
○ Minutes 1 ○ Weekly - start from: Monday

● Hours 1 ○ Monthly - start from: 1

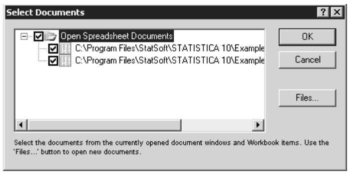
○ Days 1 ○ Quarterly - start from 1 January

○ Annually - start from 1 January

Click the ***Add data source*** button to display the ***Select Data Sources*** dialog.



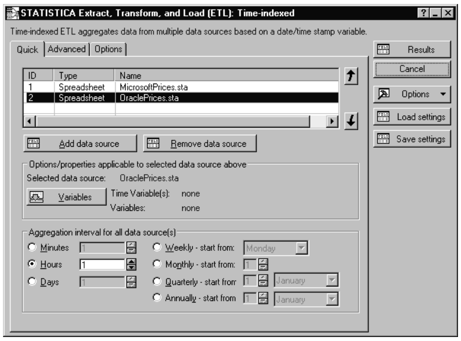
Click the ***Documents*** button to display the ***Select Documents*** dialog. Select the ***Open Spreadsheets Documents*** check box to select both data files (*MicrosoftPrices.sta* and *OraclePrices.sta*).



- ☑ Open Spreadsheet Documents
  - ☑ C:\Program Files\StatSoft\STATISTICA 10\Example
  - ☑ C:\Program Files\StatSoft\STATISTICA 10\Example

Select the documents from the currently opened document windows and Workbook items. Use the 'Files...' button to open new documents.

Click the ***OK*** button in the ***Select Documents*** dialog, and then click the ***OK*** button in the ***Select Data Sources*** dialog. The ***STATISTICA Extract, Transform, and Load (ETL): Time-indexed*** Startup Panel will appear as shown below:



Time-indexed ETL aggregates data from multiple data sources based on a date/time stamp variable.

ID	Type	Name
1	Spreadsheet	MicrosoftPrices.sta
2	Spreadsheet	OraclePrices.sta

Options/properties applicable to selected data source above

Selected data source: OraclePrices.sta

Time Variable(s): none

Variables: none

Aggregation interval for all data source(s)

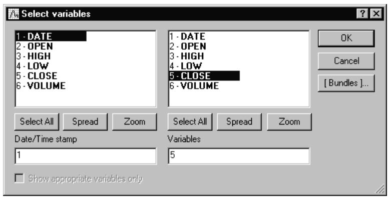
○ Minutes 1 ○ Weekly - start from: Monday

● Hours 1 ○ Monthly - start from: 1

○ Days 1 ○ Quarterly - start from 1 January

○ Annually - start from 1 January

Select *MicrosoftPrices.sta* in the file list at the top of the dialog, and click the ***Variables*** button to display the ***Select variables*** dialog. Select *DATE* from the ***Date/Time stamp*** list, and select *CLOSE* from the ***Variables*** list.



Date/Time stamp	Variables
1 - DATE	1 - DATE
2 - OPEN	2 - OPEN
3 - HIGH	3 - HIGH
4 - LOW	4 - LOW
5 - CLOSE	5 - CLOSE
6 - VOLUME	6 - VOLUME

Date/Time stamp: 1

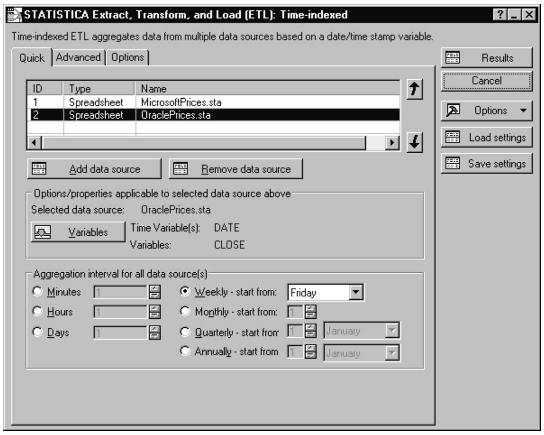
Variables: 5

□ Show appropriate variables only

Click the ***OK*** button to close this dialog and return to the ***STATISTICA Extract, Transform, and Load (ETL): Time-indexed*** Startup Panel.

Now select *OraclePrices.sta* in the file list. Click the ***Variables*** button, and select variable *1* from the ***Date/Time stamp*** list and variable *5* from the ***Variables*** list, and then click the ***OK*** button.

In the ***Aggregation interval for all data source(s)*** group box, select the ***Weekly*** option button, and change the ***start from*** field to *Friday*.



STATISTICA Extract, Transform, and Load (ETL): Time-indexed

Time-indexed ETL aggregates data from multiple data sources based on a date/time stamp variable.

ID	Type	Name
1	Spreadsheet	MicrosoftPrices.sta
2	Spreadsheet	OraclePrices.sta

Options/properties applicable to selected data source above

Selected data source: OraclePrices.sta

Time Variable(s): DATE

Variables: CLOSE

Aggregation interval for all data source(s)

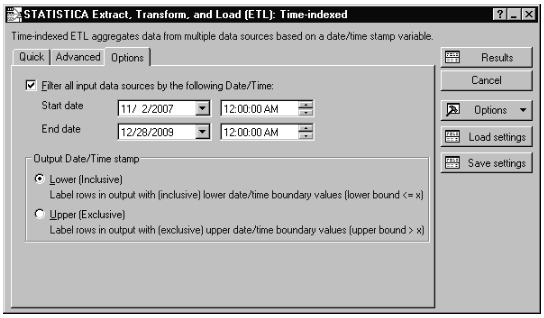
○ Minutes 1 ● Weekly - start from: Friday

○ Hours 1 ○ Monthly - start from: 1

○ Days 1 ○ Quarterly - start from 1 January

○ Annually - start from 1 January

For additional date/time options, select the ***Options*** tab. Select the ***Filter all input data sources by the following Date/Time*** check box. To limit the data that is returned from both of the selected data files, enter *11/2/2007* in the ***Start date*** field and *12/28/2007* in the ***End date*** field. This will return eight weeks of data (Friday to Friday).



STATISTICA Extract, Transform, and Load (ETL): Time-indexed

Time-indexed ETL aggregates data from multiple data sources based on a date/time stamp variable.

☑ Filter all input data sources by the following Date/Time:

Start date 11/ 2/2007 12:00:00 AM

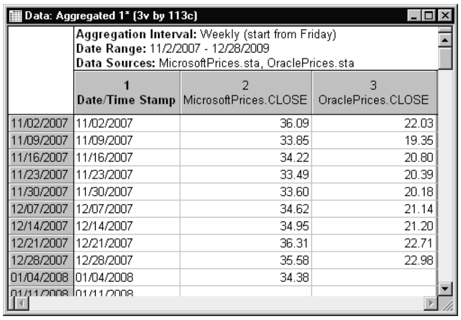
End date 12/28/2009 12:00:00 AM

Output Date/Time stamp

● Lower (Inclusive)
Label rows in output with (inclusive) lower date/time boundary values (lower bound <= x)

○ Upper (Exclusive)
Label rows in output with (exclusive) upper date/time boundary values (upper bound > x)

Now, click the ***Results*** button to merge the data into a spreadsheet.



Data: Aggregated 1* (3v by 113c)

Aggregation Interval: Weekly (start from Friday)
Date Range: 11/2/2007 - 12/28/2009
Data Sources: MicrosoftPrices.sta, OraclePrices.sta

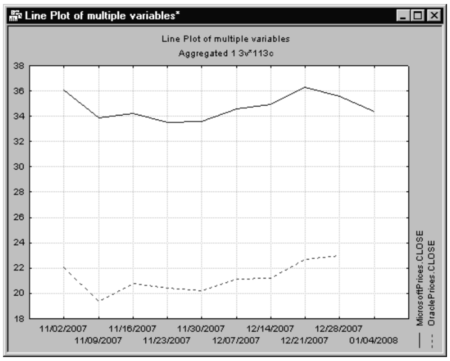
	1 Date/Time Stamp	2 MicrosoftPrices.CLOSE	3 OraclePrices.CLOSE
11/02/2007	11/02/2007	36.09	22.03
11/09/2007	11/09/2007	33.85	19.35
11/16/2007	11/16/2007	34.22	20.80
11/23/2007	11/23/2007	33.49	20.39
11/30/2007	11/30/2007	33.60	20.18
12/07/2007	12/07/2007	34.62	21.14
12/14/2007	12/14/2007	34.95	21.20
12/21/2007	12/21/2007	36.31	22.71
12/28/2007	12/28/2007	35.58	22.98
01/04/2008	01/04/2008	34.38	
01/11/2008	01/11/2008		

The two data files are now aligned weekly by date for the range 11/2/2007 to 12/28/2007. The daily closing Microsoft prices are aggregated as means, while the weekly closing Oracle prices are unchanged.

The *Results* spreadsheet displays date/time stamps as cases names so that they can be used for graphing the aggregated and aligned data.

Select the ***Graphs*** tab. In the ***More*** group, click ***2D*** and ***select Line Plots (Variables)*** to display the ***2D Line Plots - Variables*** dialog.

Click the ***Variables*** button, and in the variable selection dialog, select variables *2* and *3*. Then, click the ***OK*** button. In the ***2D Lineplots - Variables*** dialog, select ***Multiple*** for the ***Graph type***, and click the ***OK*** button. The following image shows the resultant graph plotting Microsoft and Oracle prices.

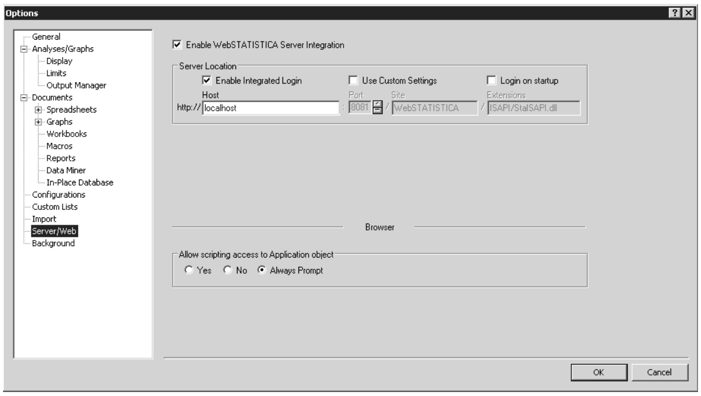


ENTERPRISE INSTALLATIONS

Example 1: *STATISTICA Enterprise Server* – Download/Offload Analyses from/to Servers

STATISTICA Enterprise Server extends the capabilities of the *STATISTICA* platform, turning several standalone workstations into a powerful, enterprise-wide collaborative-intelligence system. One of the key features of *STATISTICA Enterprise Server*'s client-server architecture is that it enables you to utilize server-side resources to run multiple, possibly time-consuming, or repetitive statistical analyses ("offload" tasks to the server) while at the same time freeing the local system for other tasks that require immediate attention. This can be achieved using either a Web browser ("thin" client) or desktop version of *STATISTICA* ("thick" client, *STATISTICA Enterprise Server* client). While the former allows access to *STATISTICA Enterprise Server* using only a browser, the latter requires *STATISTICA* installation on your computer. *STATISTICA Enterprise Server*'s tight integration with the *STATISTICA* application provides common user experience and workflow for both client and server-side operations, a generally more feature-rich and responsive user interface, and all the additional components and tools of desktop *STATISTICA*.

Offloading an analysis (or a custom script) to *STATISTICA Enterprise Server*. First, ensure that *STATISTICA Enterprise Server* integration is enabled. Select the ***Home*** tab, and in the ***Tools*** group click ***Options*** to display the ***Options*** dialog. In the tree view, select ***Server/Web***. Select the ***Enable STATISTICA Enterprise Server Integration*** check box. The only required parameter is *STATISTICA Enterprise Server's* network path (and connection settings, if they are different from the default). Ask your network administrator for these values. It is possible to ***Enable Integrated Login*** if it is supported and enabled on the server; otherwise you will need to enter your user name and password when logging in to *STATISTICA Enterprise Server*.



Options

☑ Enable WebSTATISTICA Server Integration

Server Location

☑ Enable Integrated Login ☐ Use Custom Settings ☐ Login on startup

Host: http:// localhost : Port 8081 / Site WebSTATISTICA / Extensions ISAPI/StatSAPI.dll

Allow scripting access to Application object: ○ Yes ○ No ● Always Prompt

After specifying the options on this tab, click the ***OK*** button.

The ***Server*** tab has now been added to the ribbon bar. In the ***User*** group, click ***Log In***, and enter your user name and password if requested. Upon successfully establishing a connection, the options on the ***Server*** tab will become available.

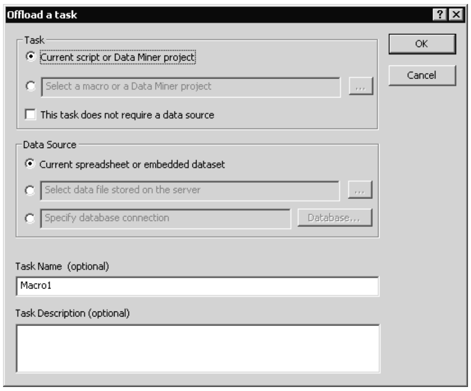
The ***Open***, ***Save***, and ***Save As*** commands in the ***File*** group are used to upload a currently open file to the server or download a file and open it locally. There are also explicit options in the ***Transfer*** group to ***Download File*** to and ***Upload File*** from specific folders on the server and the client.

Note: As real-world examples of time- or resource-consuming analyses are usually based on large data sets and/or involve iterative algorithms represented by *STATISTICA* components that are not included in all configurations of *STATISTICA*, we are deliberately going to use an example that does not require much time to complete. But even in a situation where a single analysis is quick and not resource-intensive, you might need to run a fairly complicated, time-consuming sequence of tasks, possibly scheduled at certain time intervals. In this case, the *STATISTICA Enterprise Server* scheduling facilities could be used once you have created and uploaded a custom script that represents the required tasks (for example, by combining the macros recorded during a *STATISTICA* session).

Now, record a sample analysis macro; for example, complete the steps described in *Example 2: ANOVA* (page 34).

After completing the example, in the ***ANOVA Results*** dialog, click the ***Options*** button (located at the bottom of the dialog), and from the drop-down list, select ***Create Macro***. In the ***New Macro*** dialog, accept all defaults, and click ***OK***. Test the generated macro by running it (press F5) to ensure that it produces results as expected. Click on the macro code window to ensure it has the focus.

Then, on the ***Server*** tab in the ***Tasks*** group, click ***Offload*** to display the ***Offload a task*** dialog.



Task

● Current script or Data Miner project

○ Select a macro or a Data Miner project

□ This task does not require a data source

Data Source

● Current spreadsheet or embedded dataset

○ Select data file stored on the server

○ Specify database connection

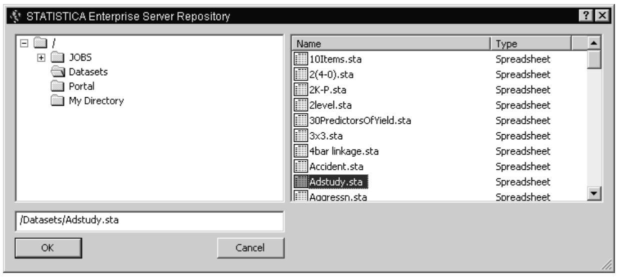
Task Name (optional): Macro1

Task Description (optional):

We need to select a task to offload (a script or a *Data Miner* project) and, optionally, a data set on which the task will operate (the data set could be an optional component since *Data Miner* projects may have their data sets embedded and macros might explicitly load data sets or not require them at all).

Since there is an open active data set (*Adstudy.sta*) and an open *STATISTICA* Macro (our sample analysis), the default settings of the options in the ***Offload a task*** dialog specify to use them for offloading. Instead, this example will demonstrate how to reference a task and a server-side data set. This option is useful since it gives you the advantage of central server-side storage, which is especially beneficial in the case of large data sets (possibly dynamically updated) that are used by multiple users.

To reference a server-side data set, in the ***Data Source*** group box, select the ***Select data file stored on the server*** option button to display the ***STATISTICA Enterprise Server Repository*** dialog.

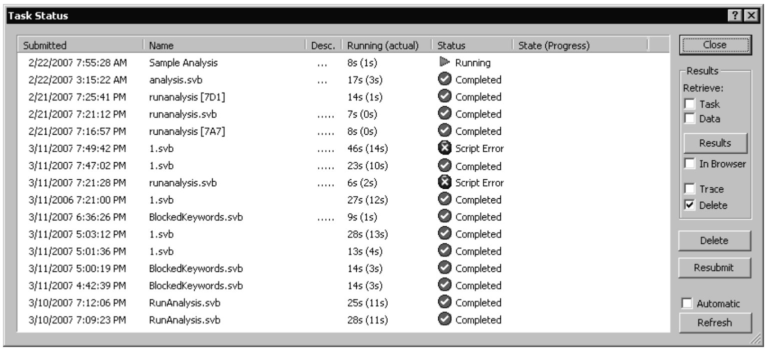


Name	Type
10Items.sta	Spreadsheet
2(4-0).sta	Spreadsheet
2K-P.sta	Spreadsheet
2level.sta	Spreadsheet
30PredictorsOfYield.sta	Spreadsheet
3x3.sta	Spreadsheet
4bar linkage.sta	Spreadsheet
Accident.sta	Spreadsheet
Adstudy.sta	Spreadsheet
Aggressn.sta	Spreadsheet

/Datasets/Adstudy.sta

The directory structure in the tree view of the dialog represents the *STATISTICA Enterprise Server* Repository (possibly abridged according to your particular permissions). Click on the *Datasets* folder in the left pane, and select *Adstudy.sta* in the right pane (or you can enter the path in the edit box at the bottom of the dialog).

Click ***OK*** in the ***STATISTICA Enterprise Repository*** dialog and in the ***Offload a task*** dialog. *STATISTICA* will submit the task to the server, uploading files if needed. Now you can switch to other activities, while periodically monitoring the status of offloaded tasks by clicking ***Status*** in the ***Tasks*** group on the ***Server*** tab to display the ***Task Status*** dialog. The following illustration shows a ***Task Status*** dialog containing several offloaded tasks.



Submitted	Name	Desc.	Running (actual)	Status	State (Progress)
2/22/2007 7:55:28 AM	Sample Analysis	...	8s (1s)	▶ Running	
2/22/2007 3:15:22 AM	analysis.svb	...	17s (3s)	✓ Completed	
2/21/2007 7:25:41 PM	runanalysis [7D1]		14s (1s)	✓ Completed	
2/21/2007 7:21:12 PM	runanalysis.svb	.....	7s (0s)	✓ Completed	
2/21/2007 7:16:57 PM	runanalysis [7A7]	.....	8s (0s)	✓ Completed	
3/11/2007 7:49:42 PM	1.svb	.....	46s (14s)	✗ Script Error	
3/11/2007 7:47:02 PM	1.svb	.....	23s (10s)	✓ Completed	
3/11/2007 7:21:28 PM	runanalysis.svb	.....	6s (2s)	✗ Script Error	
3/11/2006 7:21:00 PM	1.svb		27s (12s)	✓ Completed	
3/11/2007 6:36:26 PM	BlockedKeywords.svb	.....	9s (1s)	✓ Completed	
3/11/2007 5:03:12 PM	1.svb		28s (13s)	✓ Completed	
3/11/2007 5:01:36 PM	1.svb		13s (4s)	✓ Completed	
3/11/2007 5:00:19 PM	BlockedKeywords.svb		14s (3s)	✓ Completed	
3/11/2007 4:42:39 PM	BlockedKeywords.svb		14s (3s)	✓ Completed	
3/10/2007 7:12:06 PM	RunAnalysis.svb		25s (11s)	✓ Completed	
3/10/2007 7:09:23 PM	RunAnalysis.svb		28s (11s)	✓ Completed	

☐ Task
☐ Data
☐ In Browser
☐ Trace
☑ Delete
☐ Automatic

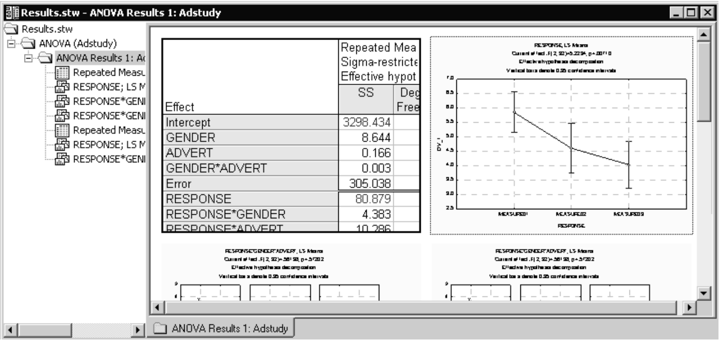
The task list status can be updated manually by clicking the ***Refresh*** button or automatically by selecting the ***Automatic*** check box in the lower-right portion of the ***Task Status*** dialog. Tasks go through ***Pending*** and ***Running*** states to either ***Completed*** or ***Script Error***.

If your task fails, double-click on the task entry to view additional information about the failure. When the error is fixed (e.g., SVB script or *Data Miner* workspace is updated), select the failed task and click the ***Resubmit*** button.

Once the task completes successfully, you can retrieve the results. Note that since the results are located on the server, they are available from any *STATISTICA* client workstation as long as you are logged in under the same credentials. The ***Results*** group box contains a ***Task*** check box and a ***Data*** check box to retrieve the task source and the data set (if applicable) back to the client. When the ***In Browser*** check box is selected, the results will be opened in the browser, switching to a thin client. This option is useful if the results are expected to be significant in size; e.g., if the analysis generates many data sets and/or graphs, you can search through them in the browser and select only the specific results you want to retrieve to your desktop. ***Trace report*** provides a diagnostic report of task execution.

To save disk space on the server, it is a good practice to delete task results that are no longer needed. A message will be displayed every time results are requested asking if the results should be deleted after retrieval (unless the ***Delete task after retrieval*** check box is cleared). Click ***OK*** to delete the results.

Once our task completes, we retrieve the results and close the ***Task Status*** dialog. Results are equivalent whether run locally or on the server.



	Repeated Mea Sigma-restricte Effective hypot	
Effect	SS	Deg Free
Intercept	3298.434	
GENDER	8.644	
ADVERT	0.166	
GENDER*ADVERT	0.003	
Error	305.038	
RESPONSE	80.879	
RESPONSE*GENDER	4.383	
RESPONSE*ADVERT	10.286	

Example 2: Using *STATISTICA* in Regulated Environments

In a regulated environment, analyses conducted for GxP (Good Manufacturing Practices, Good Clinical Practices, Good Laboratory Practices) applications are ones that impact consumer safety such as in clinical trials, manufacturing, and quality control. When a business conducts analyses for a GxP application, regulatory bodies recommend that the company be able to prove that the results of the validated analysis system (e.g., *STATISTICA*) are under control. *STATISTICA*, through its audit trail and spreadsheet/report locking features, offers the tools you need to meet this regulatory requirement.

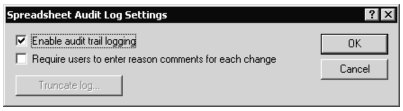
In order to meet traceability requirements for GxP applications, there are at least three concerns: 1) control of the input data being submitted to the analysis (i.e., knowing who made what change, at what time, for what reason; and the old values and new values), 2) control of the results tables and graphs (e.g., demonstrate that they were not altered in any way after they were created), and 3) traceability between the version of the input spreadsheet and the results outputs. *STATISTICA* provides this information through its Spreadsheet Audit Trails and GxP Reports functionality.

See also *STATISTICA Document Management System* in the *Electronic Help* for more details about versioning/history of *STATISTICA* documents.

Control of Input Data

Enable Audit Trail Logging

Open a *STATISTICA* Spreadsheet. Select the ***Tools*** tab, click ***Audit Trail***, and select ***Settings*** from the drop-down menu. The ***Spreadsheet Audit Log Settings*** dialog will be displayed. Select the ***Enable audit trail logging*** check box to enable audit trail logging for the current spreadsheet.



Spreadsheet Audit Log Settings

☑ Enable audit trail logging

☐ Require users to enter reason comments for each change

Note that when spreadsheet audit trail logging is enabled, the spreadsheet is automatically set to direct mode, i.e., changes made to the spreadsheet will be immediately written to disk. Thus, when audit trail logging is enabled, changes to the data file cannot be undone.

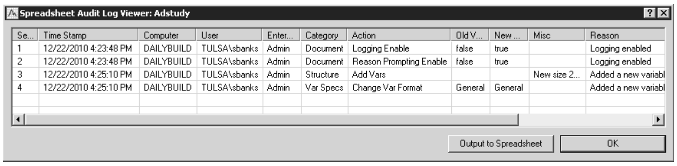
Select the ***Require users to enter reason comments for each change*** check box to require users to explain each change made to the spreadsheet.

The ***Truncate log*** button is available only if audit trail logging has previously been specified, and there is a current ***Spreadsheet Audit Log Viewer*** attached to the spreadsheet. Clicking this button will truncate the spreadsheet log and delete all existing entries. You will be prompted to confirm this action before the current entries are deleted. Once the log is truncated, the truncate action will be recorded in the newly truncated log file.

Click ***OK*** in the ***Spreadsheet Audit Log Settings*** dialog, and audit trail logging will be enabled; in fact, the Enter reason for change dialog will be displayed immediately in order to enter the reason for enabling the logging function. Enter a comment, and click ***OK***.

Now, right-click in the header of the last variable in the spreadsheet, and select ***Add Variables*** from the shortcut menu. In the ***Add Variables*** dialog, we will accept all defaults, so click ***OK***. The ***Enter reason for change*** dialog will be displayed; you must enter a comment and click ***OK*** before the change will be made. When audit trail logging is enabled, every change made to the spreadsheet will be documented, and when the ***Require users to enter reason comments for each change*** check box is selected, user comments also will be stored and displayed in the ***Spreadsheet Audit Log Viewer***.

Next, on the ***Tools*** tab, click ***Audit Trail*** and select ***View Log*** to display the ***Spreadsheet Audit Log Viewer***.



Spreadsheet Audit Log Viewer: Adstudy

Se...	Time Stamp	Computer	User	Enter...	Category	Action	Old V...	New ...	Misc	Reason
1	12/22/2010 4:23:48 PM	DAILYBUILD	TULSA\sbanks	Admin	Document	Logging Enable	false	true		Logging enabled
2	12/22/2010 4:23:48 PM	DAILYBUILD	TULSA\sbanks	Admin	Document	Reason Prompting Enable	false	true		Logging enabled
3	12/22/2010 4:25:10 PM	DAILYBUILD	TULSA\sbanks	Admin	Structure	Add Vars			New size 2...	Added a new variabl
4	12/22/2010 4:25:10 PM	DAILYBUILD	TULSA\sbanks	Admin	Var Specs	Change Var Format	General	General		Added a new variabl

The log viewer displays a grid of information regarding the audited actions including the sequence number, time of change, the computer used to make the change, user information, the nature of the change, and the reason for the change. Column widths in the log grid can be increased and decreased using standard Windows techniques. The Spreadsheet Audit Trails are saved and embedded into each respective spreadsheet.

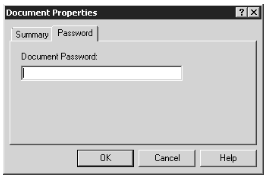
Password encryption vs. locking. A spreadsheet can be password encrypted so that it cannot be opened without the correct password. Only users who know the password can open the spreadsheet. Once a password encrypted spreadsheet is opened, it can be modified.

Alternatively, locking a spreadsheet makes portions of the spreadsheet read-only, enabling you to prevent changes to some or all aspects of the spreadsheet. The spreadsheet can be opened by anyone, but locked portions cannot be altered.

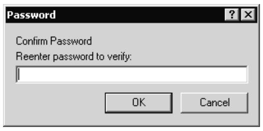
Both the password encryption options and spreadsheet locking facilities can be used simultaneously.

Password Encrypt a Spreadsheet

Open a *STATISTICA* Spreadsheet. Click the Start button in the upper-left corner of the ribbon bar, and from the drop-down menu select ***Properties*** to display the ***Document Properties*** dialog. Select the ***Password*** tab.



Enter a password in the ***Document Password*** field, and click the ***OK*** button. The ***Password*** dialog will be displayed, where you reenter the password to confirm it; passwords are context-sensitive.



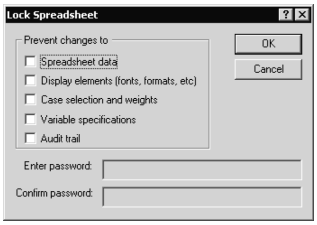
Click the ***OK*** button in the ***Password*** dialog, and close the data file. A dialog is displayed where you can choose to save the changes; click the ***Yes*** button so that the password will be encrypted. The next time anyone attempts to open this spreadsheet, the ***Password*** dialog will be displayed, and the correct password must be entered before the spreadsheet will open.

Lock a Spreadsheet

In order to meet compliance requirements, it is necessary to have control of the reliability of input data. Using the spreadsheet locking options, you can prevent changes to all spreadsheet features, from the appearance of the data (i.e., display elements, variable specifications) to the actual data and any case selection conditions or weights that are defined for the spreadsheet. Of course, sometimes changes have to be made (e.g., when data are incorrectly entered). The *STATISTICA* Spreadsheet Audit Trail facility, when enabled, will record each change made to the spreadsheet.

With *STATISTICA Enterprise* products, only users with System Administrator permissions can modify Spreadsheet Audit Trail settings. For more information, see the *Electronic Help* for *STATISTICA Enterprise* facilities.

With a spreadsheet open, select the ***Tools*** tab. Click ***Locking*** to display the ***Lock Spreadsheet*** dialog.



Here, you can specify which aspects of the spreadsheet that you want to lock. When users try to change a locked feature, a message will be displayed, informing them that the spreadsheet is locked.

Select the ***Spreadsheet data*** check box to prevent changes to the actual data contained in the spreadsheet. Users will be unable to change the data values and the missing data code. They will also be unable to perform any data management operations that affect the spreadsheet (e.g., change the data type or the length for text variables). If this check box is cleared, users will be able to edit the data (e.g., by updating queries and Spreadsheet Formulas or by simply typing in new values).

Select the ***Display elements (fonts, formats, etc.)*** check box to prohibit the modification of fonts and formats used in the spreadsheet. Options for changing the font size, color, type, and style (i.e., bold, underline) will be dimmed. Additionally, the options for applying spreadsheet layouts (accessible by selecting the ***Format*** tab and clicking ***Layouts*** in the ***Spreadsheet*** group) will be unavailable.

Select the ***Case selection and weights*** check box to prevent users from changing case selection conditions and case weights for the locked spreadsheet. Users will not be able to toggle the use of selection conditions or change the currently defined selection conditions. Most options on the ***Selection*** tab of the ***Spreadsheet Case Selection Conditions*** dialog will be dimmed; however, options on the other tabs of that dialog (e.g., creating subsamples, applying formats to selection conditions) are still available. Options on the ***Case Weights*** dialog will be unavailable.

Select the ***Variable specifications*** check box to prevent changes to the variable specifications (e.g., measurement type, missing data code, display format, long variable name). Users will still be able to view the individual variable specification dialog (accessible by double-clicking the variable header) and the ***Variable Specifications Editor***; however, options for changing these specifications will be dimmed.

Select the ***Audit trail*** check box to prevent changes to the audit trail settings. Users will be unable to modify the audit trail settings.

Enter a password to use when locking and unlocking the spreadsheet, confirm the password (which is context sensitive), and click ***OK***. Although a password is not required, it is strongly recommended. If a password is not entered and confirmed, any user can unlock spreadsheet features by simply clearing the selected check boxes. Note that if locks have already been defined, you must enter the correct password before locks can be changed or modified.

Now try making changes in the spreadsheet; a message will be displayed informing you that the operation cannot be completed because the spreadsheet is locked.

Controlling Results and Traceability

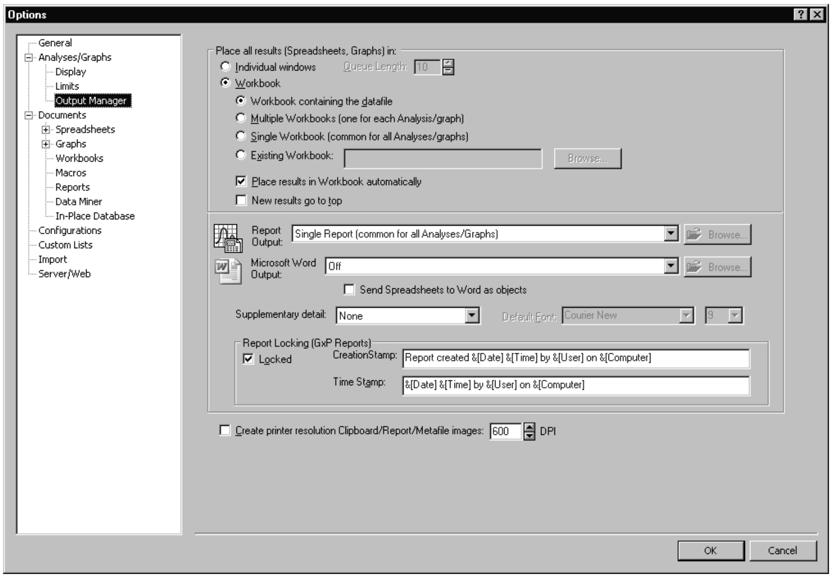
To meet compliance requirements, another step is to ensure that reported results are under control. *STATISTICA* provides options for creating GxP reports. In GxP mode, all results are sent to a report window, and the window is locked. All options for removing results (***Cut***, ***Extract - Original***, ***Clear***, etc.) and adding results (***Paste***, ***Insert***) are disabled. *STATISTICA* can also include a creation date in all reports as well as a time stamp for all results that are added from results dialogs. The appearance and content of the creation date and time stamp are completely configurable and can include user and computer information in addition to the time and date. Thus, in GxP mode, you can know when the results were created and by whom. You can also be certain that results have not been removed.

An additional feature of GxP mode is a traceability option. When running in GxP mode, *STATISTICA* automatically verifies whether spreadsheet audit trails are enabled. If they are, *STATISTICA* includes the spreadsheet name and version number in the report. Sometimes version numbers are not available, for example, if audit trails are not enabled or the results are created from an In-place Database connection. When that is the case, *STATISTICA* will provide an explanation for why a version number is not available.

Create a GxP Report

Select the ***Home*** tab. In the ***Tools*** group, click ***Options*** to display the ***Options*** dialog. In the tree view, select ***Output Manager***, located under ***Analyses/Graphs***. From the ***Report Output*** drop-down list, select either ***Send to Multiple Reports (one for each Analysis/Graph)*** or ***Single Report (common for all Analyses/ Graphs)***.

Select the ***Locked*** check box to make the ***Report Locking (GxP Reports)*** options become available and to ensure that documents cannot be removed from the report. Options pertaining to reports such as ***Cut***, ***Paste***, ***Delete***, ***Extract***, etc., will be disabled.



Place all results (Spreadsheets, Graphs) in:

○ Individual windows   Queue Length: 10

● Workbook

● Workbook containing the datafile

○ Multiple Workbooks (one for each Analysis/graph)

○ Single Workbook (common for all Analyses/graphs)

○ Existing Workbook:

☑ Place results in Workbook automatically

☐ New results go to top

Report Output: Single Report (common for all Analyses/Graphs)

Microsoft Word Output: Off

☐ Send Spreadsheets to Word as objects

Supplementary detail: None   Default Font: Courier New   9

Report Locking (GxP Reports)

☑ Locked   CreationStamp: Report created &[Date] &[Time] by &[User] on &[Computer]

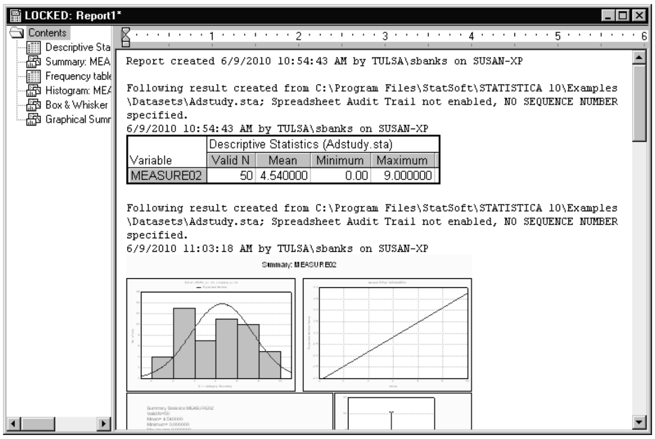
Time Stamp: &[Date] &[Time] by &[User] on &[Computer]

☐ Create printer resolution Clipboard/Report/Metafile images: 600 DPI

To include a creation stamp at the top of the file, you can accept the default format in the ***CreationStamp*** field, or enter your own. The following codes can be used in this field: *&[Date], &[Time]*, *&[User]*, and *&[Computer]*. Any text you enter will be displayed as is.

To include a time stamp above each object as it is added to the report, you can accept the default format in the ***Time Stamp*** field, or enter your own. The following codes can be used in this field: *&[Date]*, *&[Time]*, *&[User]*, and &[Computer].

Click ***OK*** in the ***Options*** dialog, and now perform any analysis; e.g., use ***Basic Statistics*** to create a quick ***Descriptive Statistics*** summary spreadsheet. When you click the ***Summary*** button, the results will be sent to a locked report that lists the creator, date, time, etc., of the analysis.



```
Report created 6/9/2010 10:54:43 AM by TULSA\sbanks on SUSAN-XP

Following result created from C:\Program Files\StatSoft\STATISTICA 10\Examples
\Datasets\Adstudy.sta; Spreadsheet Audit Trail not enabled, NO SEQUENCE NUMBER
specified.
6/9/2010 10:54:43 AM by TULSA\sbanks on SUSAN-XP
```

	Descriptive Statistics (Adstudy.sta)			
Variable	Valid N	Mean	Minimum	Maximum
MEASURE02	50	4.540000	0.00	9.000000

```
Following result created from C:\Program Files\StatSoft\STATISTICA 10\Examples
\Datasets\Adstudy.sta; Spreadsheet Audit Trail not enabled, NO SEQUENCE NUMBER
specified.
6/9/2010 11:03:18 AM by TULSA\sbanks on SUSAN-XP
```

Example 3: *STATISTICA Enterprise*

STATISTICA Enterprise products extend the functionality of *STATISTICA* applications by offering collaborative work, central administration, system level customization, and other features necessary when using *STATISTICA* applications as part of the enterprise-level computer systems.

STATISTICA Enterprise Manager is a component of the *STATISTICA Enterprise* system that enables users to configure various aspects of the *Enterprise* system including user administration, system view organization, database connection maintenance, data configurations, and analysis configurations.

For this example, we will:

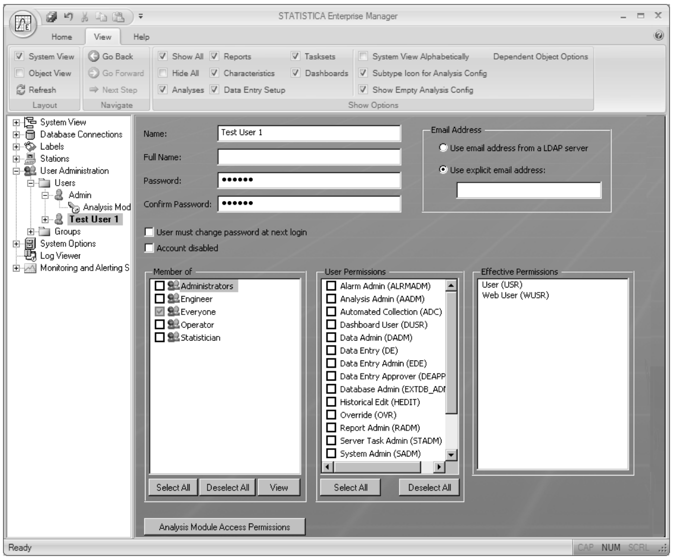
1. Create a new user
2. Create a new group
    a. Assign permissions to the group
    b. Add the user (see No. 1) to the group
3. Create a system view node
4. Create a new database connection
5. Create a data configuration
6. Create an analysis configuration
7. Run the analysis configuration

System View vs. Object View

Before starting this example, one thing should be noted. In *STATISTICA Enterprise Manager*, on the ***View*** tab, you can select either ***System View*** or ***Object View***. In ***System View***, objects, e.g., data configurations and analysis configurations, are shown as child nodes. In ***Object View***, objects are shown as child nodes within their respective categories. For this example, ***System View*** should be selected.

1. Create a New User

Launch the *Enterprise Manager*, and log in as a user who is part of the default Administrator group. In the tree view (the left pane), click the plus sign ⊞ next to the *User Administration* node to expand it, and then select the *Users* folder. In the properties page (the right pane), click the ***New User*** button to display the options to create a new user. In the ***Name*** field, enter *Test User 1*, and define a password and confirm the password.



Name: Test User 1

Full Name:

Password: ••••••

Confirm Password: ••••••

Email Address

○ Use email address from a LDAP server

● Use explicit email address:

☐ User must change password at next login

☐ Account disabled

Member of

☐ Administrators
☐ Engineer
☑ Everyone
☐ Operator
☐ Statistician

User Permissions

☐ Alarm Admin (ALRMADM)
☐ Analysis Admin (AADM)
☐ Automated Collection (ADC)
☐ Dashboard User (DUSR)
☐ Data Admin (DADM)
☐ Data Entry (DE)
☐ Data Entry Admin (EDE)
☐ Data Entry Approver (DEAPP
☐ Database Admin (EXTDB_ADI
☐ Historical Edit (HEDIT)
☐ Override (OVR)
☐ Report Admin (RADM)
☐ Server Task Admin (STADM)
☐ System Admin (SADM)

Effective Permissions

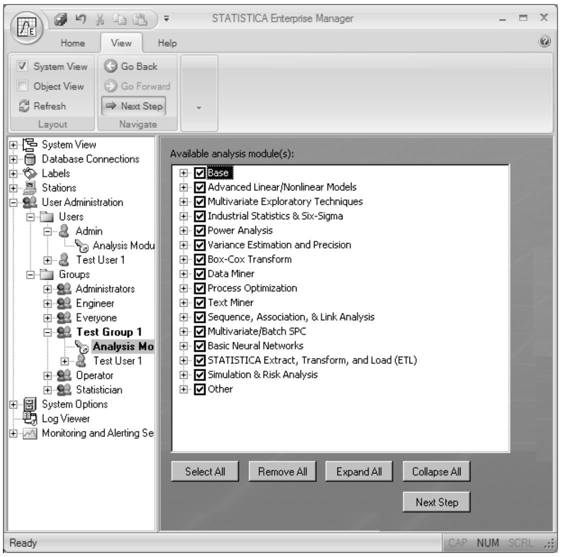
User (USR)
Web User (WUSR)

Then, click the ***Commit Changes*** button located at the top of the application on the Quick Access toolbar to save the changes. A message will be displayed that informs you that the user doesn't have permission to login. Click the ***Yes*** button to continue.

We will now create a group, give the group permissions, and assign the new user to that group to allow the user to have permission to log on to the *Enterprise Manager*. With this method, any permission changes will only need to be applied to the group instead of the individual users, making maintenance of users in *STATISTICA Enterprise* easier.

2. Create a New Group

In the *User Administration* node, select the *Groups* folder, and in the properties page, click the ***New Group*** button to display the options to create a new group. In the ***Name*** field, enter *Test Group 1*. In the ***Group Members*** frame, select the check box adjacent to *Test User 1*. This will add the previously created user to the group. In the ***Group Permissions*** frame, select the check boxes adjacent to *Analysis Admin (AADM)* and *Web User (WUSR)*. In the tree view, click the plus sign ⊞ adjacent to the *Test Group 1* node to expand it, and select *Analysis modules*. In the properties page, click the ***Select All*** button to select all of the modules in the ***Available analysis module(s)*** list.



Available analysis module(s):

- ☑ Base
- ☑ Advanced Linear/Nonlinear Models
- ☑ Multivariate Exploratory Techniques
- ☑ Industrial Statistics & Six-Sigma
- ☑ Power Analysis
- ☑ Variance Estimation and Precision
- ☑ Box-Cox Transform
- ☑ Data Miner
- ☑ Process Optimization
- ☑ Text Miner
- ☑ Sequence, Association, & Link Analysis
- ☑ Multivariate/Batch SPC
- ☑ Basic Neural Networks
- ☑ STATISTICA Extract, Transform, and Load (ETL)
- ☑ Simulation & Risk Analysis
- ☑ Other

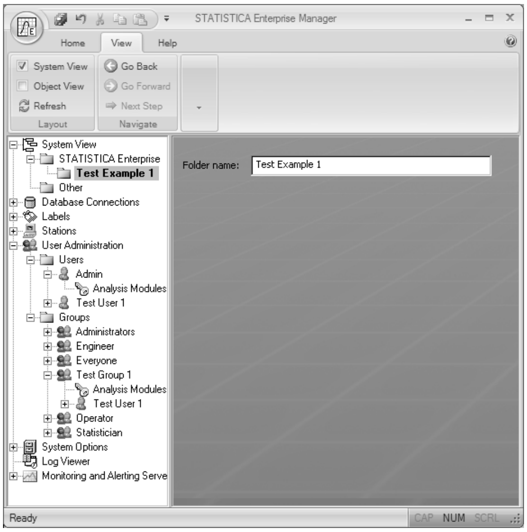
This will give users of this group permission to log on to both Web and desktop *STATISTICA* and run all of the available analyses and reports.

Click the ***Commit Changes*** button to save the changes.

We have now created the necessary user and group security to run analyses and reports. When creating the data, analysis, and report configurations in the next steps, we will assign this group to those objects to allow only users within the group to run them.

3. Create a System View Node

Now we will create a *System View* node to hold this example's data, analyses, and report configuration. In the tree view, click the plus sign ⊞ adjacent to the *System View* node to expand it. Right-click on the *STATISTICA Enterprise* folder, and from the shortcut menu, select ***New Folder***. In the ***Folder name*** text box in the properties page, enter *Test Example 1* as the new folder's name.

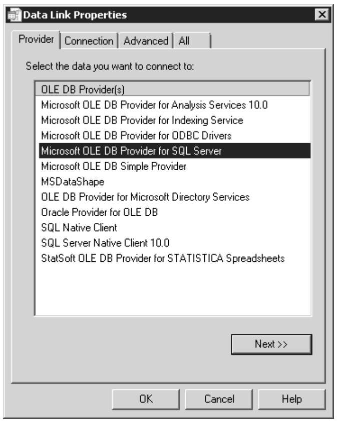


Folder name: Test Example 1

Click ***Commit Changes*** to save the change. This folder will now be used to house the data, analyses, and report configurations.

4. Create a New Database Connection

Right-click on the *Database Connections* node in the tree view, and from the shortcut menu, select ***New Database Connection*** to display the ***Data Link Properties*** dialog.

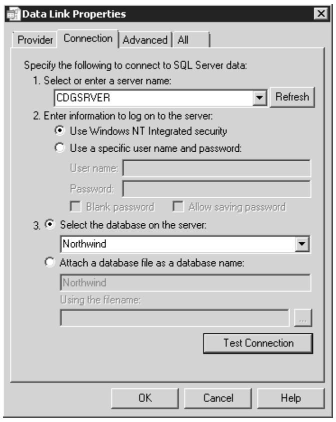

Data Link Properties
Provider
Connection
Advanced
All
Select the data you want to connect to:
OLE DB Provider(s)
Microsoft OLE DB Provider for Analysis Services 10.0
Microsoft OLE DB Provider for Indexing Service
Microsoft OLE DB Provider for ODBC Drivers
Microsoft OLE DB Provider for SQL Server
Microsoft OLE DB Simple Provider
MSDataShape
OLE DB Provider for Microsoft Directory Services
Oracle Provider for OLE DB
SQL Native Client
SQL Server Native Client 10.0
StatSoft OLE DB Provider for STATISTICA Spreadsheets
Next >>
OK
Cancel
Help

For this example, we'll use the Northwind sample database installed with Microsoft SQL Server, so select ***Microsoft OLE DB Provider for SQL Server***, and click the ***Next >>*** button. The ***Data Link Properties*** dialog - ***Connection*** tab will be displayed.

Select a server from the ***Select or enter a server name*** drop-down list.

Then, select the log on option button appropriate to your SQL Server Northwind database installation. Select either the ***Use Windows NT Integrated security*** option button, or select the ***Use a specific user name and password*** option button and enter a ***User name*** and ***Password*** in the respective fields.

Next, select *Northwind* from the ***Select the database on the server*** drop-down list.



Data Link Properties

Provider | Connection | Advanced | All

Specify the following to connect to SQL Server data:

1. Select or enter a server name:

CDGSRVER [Refresh]

2. Enter information to log on to the server:

● Use Windows NT Integrated security

○ Use a specific user name and password:

User name:

Password:

□ Blank password □ Allow saving password

3. ● Select the database on the server:

Northwind

○ Attach a database file as a database name:

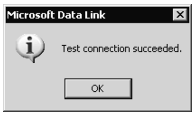
Northwind

Using the filename:

[Test Connection]

[OK] [Cancel] [Help]

Click the ***Test Connection*** button to attempt a connection to the specified data source. A prompt will be displayed that acknowledges that the *Test connection succeeded.* If it didn't succeed, check your access permissions to the file and ensure that the settings are correct. For example, spelling errors and case sensitivity can cause failed connections.

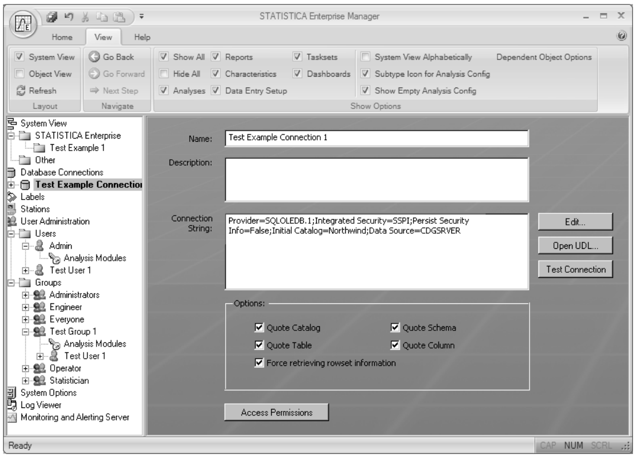


Microsoft Data Link

Test connection succeeded.

[OK]

Click ***OK*** in the prompt, and click ***OK*** in the ***Data Link Properties*** dialog. In the resulting properties page, enter *Test Example Connection 1* in the ***Name*** field.



☑ Show All ☑ Reports ☑ Tasksets ☐ System View Alphabetically
☐ Hide All ☑ Characteristics ☑ Dashboards ☑ Subtype Icon for Analysis Config
☑ Analyses ☑ Data Entry Setup ☑ Show Empty Analysis Config

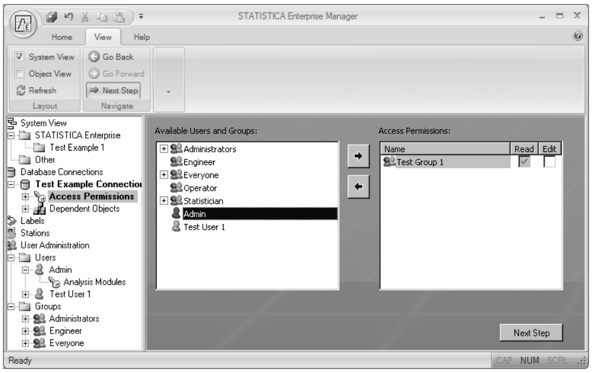
Name: Test Example Connection 1

Description:

Connection String: Provider=SQLOLEDB.1;Integrated Security=SSPI;Persist Security Info=False;Initial Catalog=Northwind;Data Source=CDGSRVER

Options:
☑ Quote Catalog ☑ Quote Schema
☑ Quote Table ☑ Quote Column
☑ Force retrieving rowset information

Then, click the ***Access Permissions*** button. From the list of ***Available Users and Groups***, select ***Test Group1***, and then click the top arrow button ➡ to move ***Test Group 1*** to the ***Access Permissions*** list.



Access Permissions:

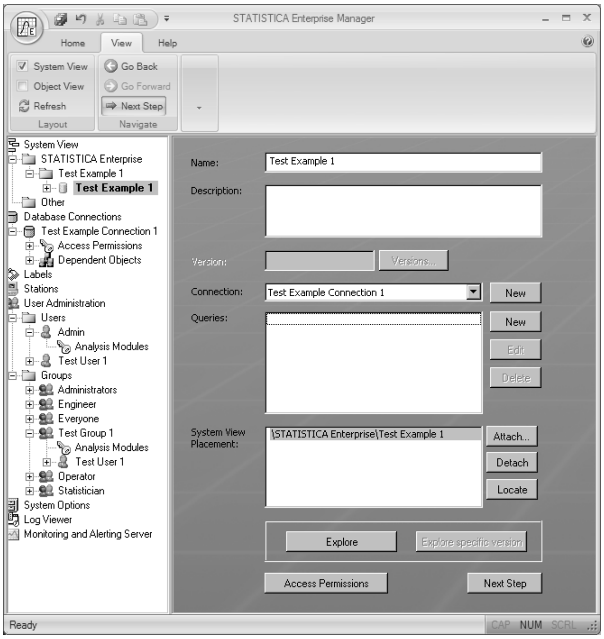
Name	Read	Edit
Test Group 1	☑	☐

Now, click the ***Commit Changes*** button.

With the database connection created to the Northwind database, we will now create a data configuration to extract data from the database.

5. Create a Data Configuration

Right-click on the *Test Example 1* folder in the tree view, and from the shortcut menu, select ***New Data Configuration***. In the properties page, enter *Test Example 1* in the ***Name*** field. Click the arrow next to the ***Connection*** field, and from the drop-down list, select ***Test Example Connection 1***.



☑ System View
□ Object View

Name: Test Example 1

Description:

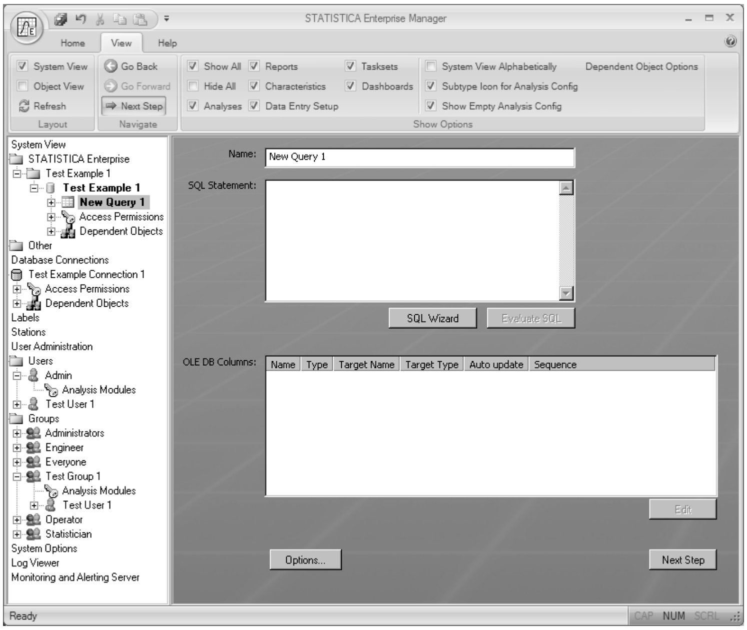
Version:

Connection: Test Example Connection 1

Queries:

System View Placement: \STATISTICA Enterprise\Test Example 1

Click the ***Next Step*** button to display the new query options.



☑ System View
□ Object View

☑ Show All
□ Hide All
☑ Analyses
☑ Reports
☑ Characteristics
☑ Data Entry Setup
☑ Tasksets
☑ Dashboards
□ System View Alphabetically
☑ Subtype Icon for Analysis Config
☑ Show Empty Analysis Config

Name: New Query 1

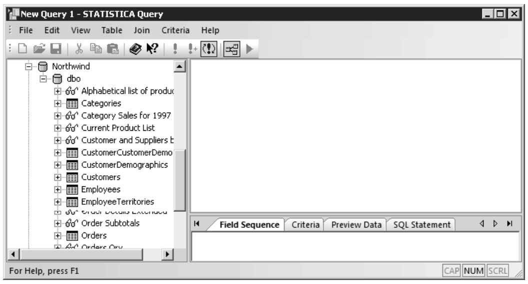
SQL Statement:

OLE DB Columns:

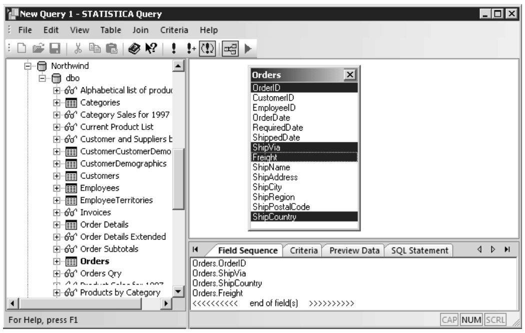
Name	Type	Target Name	Target Type	Auto update	Sequence

StatSoft®

Click the ***SQL Wizard*** button to display the ***New Query*** dialog, which will open in *STATISTICA*.

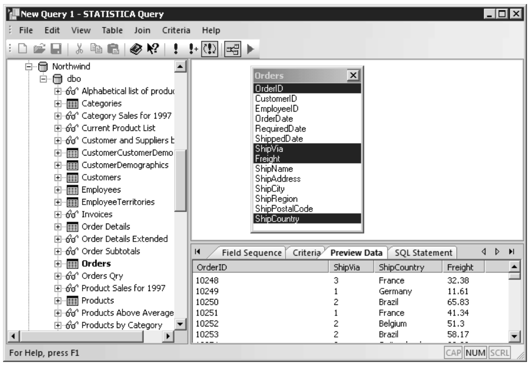


Drag the *Orders* table from the left pane into the editor viewer (the upper-right pane), and then select, in the following order, the *OrderID*, *Ship Via*, *ShipCountry*, and *Freight* fields.



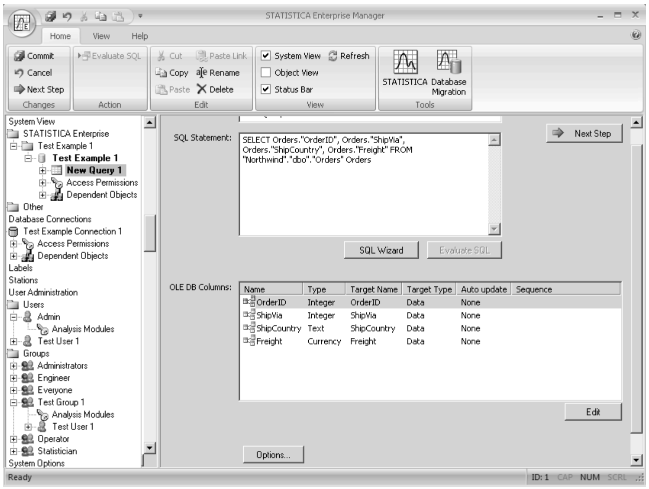
```
Orders.OrderID
Orders.ShipVia
Orders.ShipCountry
Orders.Freight
<<<<<<<<<<    end of field(s)    >>>>>>>>>>
```

Select the ***Preview Data*** tab in the query properties view (lower-right pane) and click the Refresh ! toolbar button (the red exclamation mark). This will test the query to ensure that values are being retrieved from the defined query.



OrderID	ShipVia	ShipCountry	Freight
10248	3	France	32.38
10249	1	Germany	11.61
10250	2	Brazil	65.83
10251	1	France	41.34
10252	2	Belgium	51.3
10253	2	Brazil	58.17

Click the Return Data to *STATISTICA* ▶ toolbar button (green arrow) to submit this query back to the data configuration.



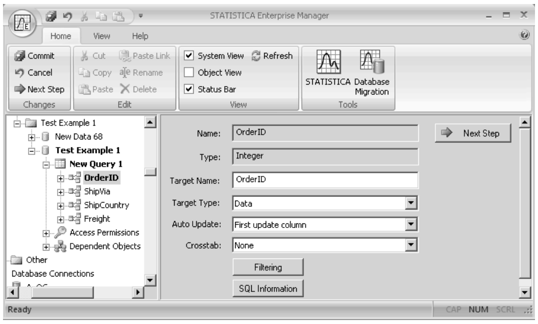
SQL Statement:

```
SELECT Orders."OrderID", Orders."ShipVia",
Orders."ShipCountry", Orders."Freight" FROM
"Northwind"."dbo"."Orders" Orders
```

OLE DB Columns:

Name	Type	Target Name	Target Type	Auto update	Sequence
OrderID	Integer	OrderID	Data	None	
ShipVia	Integer	ShipVia	Data	None	
ShipCountry	Text	ShipCountry	Data	None	
Freight	Currency	Freight	Data	None	

Click the *OrderID* row to highlight it, and then click the ***Edit*** button to display options to edit the *OrderID* column. Click the ***Auto Update*** arrow, and from the drop-down list, select ***First update column***. This enables you to detect changes in the *OrderID* column. In addition, the column is sorted.



Name: OrderID

Type: Integer

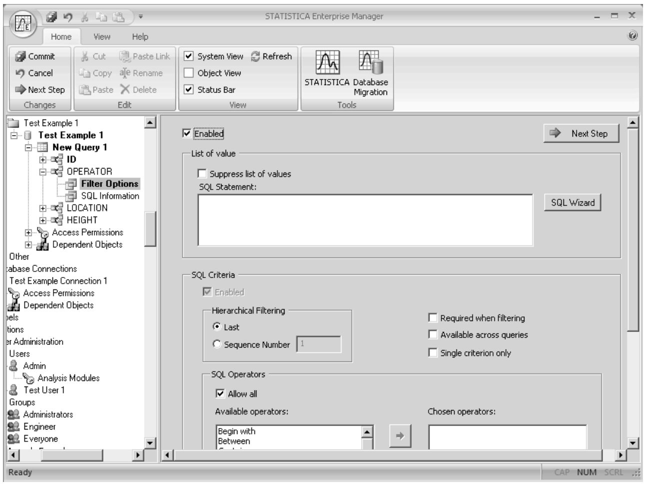
Target Name: OrderID

Target Type: Data

Auto Update: First update column

Crosstab: None

Click the ***Next Step*** button to edit the *ShipVia* column. Click the ***Filtering*** button to display the filtering options, and select the ***Enabled*** check box to allow filtering on the *ShipVia* column.



☑ Enabled

List of value

□ Suppress list of values

SQL Statement:

SQL Criteria

☑ Enabled

Hierarchical Filtering

● Last

○ Sequence Number 1

□ Required when filtering

□ Available across queries

□ Single criterion only

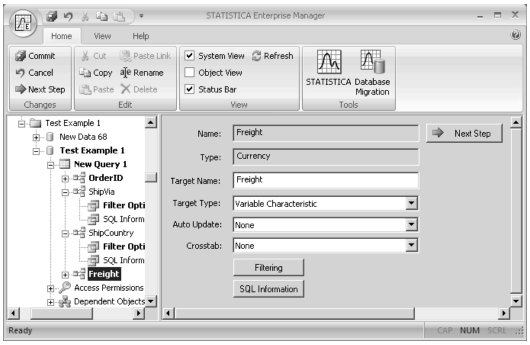
SQL Operators

☑ Allow all

Available operators: Begin with, Between

Chosen operators:

Click the ***Next Step*** button to return to *ShipVia* column editing options, and then click the ***Next Step*** button to edit the *ShipCountry* column. Click the ***Filtering*** button to display the filtering options, and select the ***Enabled*** check box to allow filtering on the *ShipCountry* column. Click the ***Next Step*** button to return to the *ShipCountry* column editing options, and then click the ***Next Step*** button to edit the *Freight* column. Click the ***Target Type*** arrow, and from the drop-down list, select ***Variable Characteristic***. This option will make this column available to perform packaged SPC analyses (this is the column containing the data to be analyzed).



Name: Freight

Type: Currency

Target Name: Freight

Target Type: Variable Characteristic

Auto Update: None

Crosstab: None

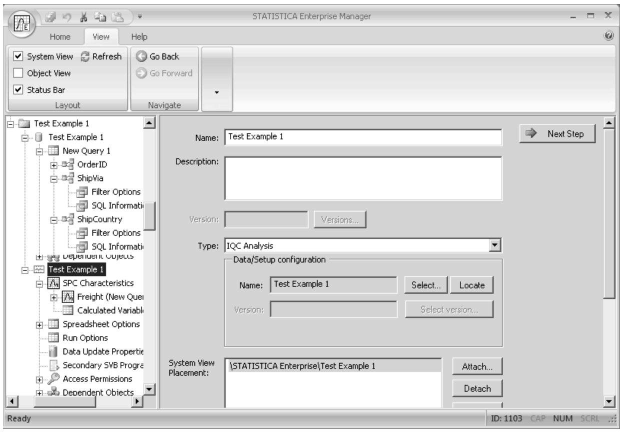
Next, click the ***Next Step*** button to display the ***Access Permissions*** options for this object. From the list of ***Available Users and Groups***, select ***Test Group 1***, and then click the top arrow button ➔ to move ***Test Group 1*** to the ***Access Permissions*** list. Now this data configuration will be executable (but not editable) by the users of *Test Group 1*.

Click the ***Commit Changes*** button to commit this new data configuration to *STATISTICA Enterprise Manager*.

6. Create an Analysis Configuration

Now that a data configuration has been defined to extract data from the Northwind database, an analysis configuration to analyze the data needs to be created.

In the tree view, right-click on the *Test Example 1* folder, and from the shortcut menu, select ***New Analysis Configuration*** to display the ***Select a Data Configuration*** dialog. Select the *Test Example 1* object, and click the ***OK*** button. If a dialog is displayed with the statement: *When selected, this option will replace the permissions of this Analysis with those of the selected Data*, click ***OK***.



Name: Test Example 1

Description:

Version:

Type: IQC Analysis

Data/Setup configuration

Name: Test Example 1

Version:

System View Placement: \STATISTICA Enterprise\Test Example 1

Click the ***Next Step*** button to continue creating the analysis configuration (leaving the default name the same as the data configuration for expediency only). Click the ***Next Step*** button once again to continue editing the analysis configuration.

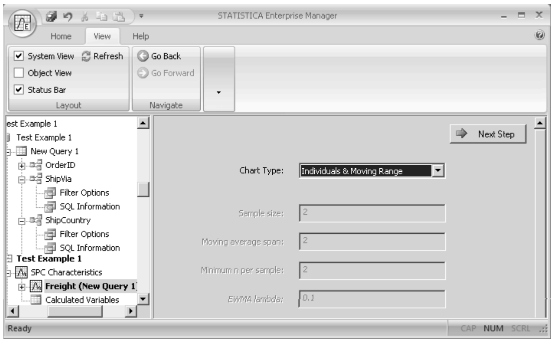


Chart Type: Individuals & Moving Range

Sample size: 2

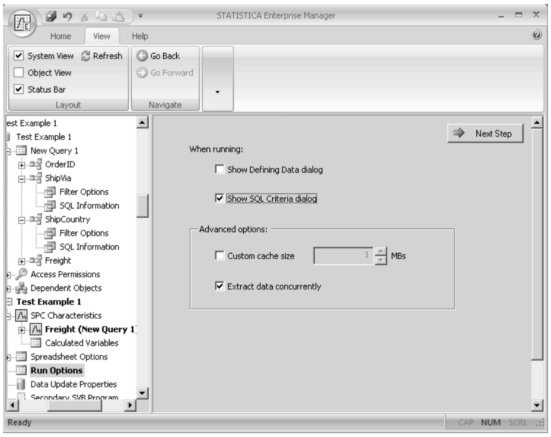
Moving average span: 2

Minimum n per sample: 2

EWMA lambda: 0.1

In the properties page for the *SPC Characteristics - Freight* column, change the ***Chart Type*** to ***Individuals & Moving Range*** (as shown in the above illustration).

No other SPC options need to be configured, so select the *Run options* node in the tree view, and select the ***Show SQL Criteria dialog*** check box in the properties page.



When running:

□ Show Defining Data dialog

☑ Show SQL Criteria dialog

Advanced options:

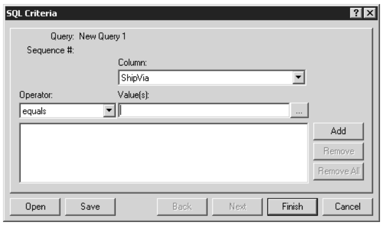
□ Custom cache size 1 MBs

☑ Extract data concurrently

This option will specify that *STATISTICA* prompt for filtering on those columns that have ***Filter*** options in the data configuration (if, when defining the ***Filter*** options, they were set to ***Required when filtering***, this step would not be required as it would always force a filtering prompt when running – in this example it was not required to force filtering). Click the ***Commit Changes*** button to save this analysis configuration to *STATISTICA Enterprise*.

7. Run the Analysis Configuration

Close the *Enterprise Manager*, and log on to *STATISTICA* as the *Test User 1* user created in Step 1. Select the ***Enterprise*** tab, and in the ***Enterprise*** group, click ***Run Analysis/Report*** to display the ***Run Analysis or Report*** dialog (this dialog may be displayed automatically depending on your configuration). Select the *Test Example 1* analysis, and click the ***OK*** button; the ***SQL Criteria*** dialog will be displayed.



Query: New Query 1

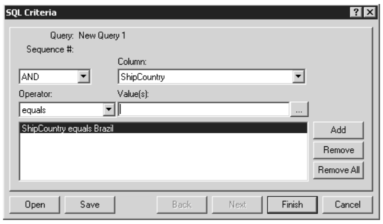
Sequence #:

Column: ShipVia

Operator: equals

Value(s):

Click the ***Column*** arrow, and select ***ShipCountry*** from the drop-down list. Click the browse button ... to display the ***Value of ShipCountry*** dialog, which contains the list of available *ShipCountry* values. Select ***Brazil*** and click the ***OK*** button.


SQL Criteria

Query: New Query 1

Sequence #:

AND

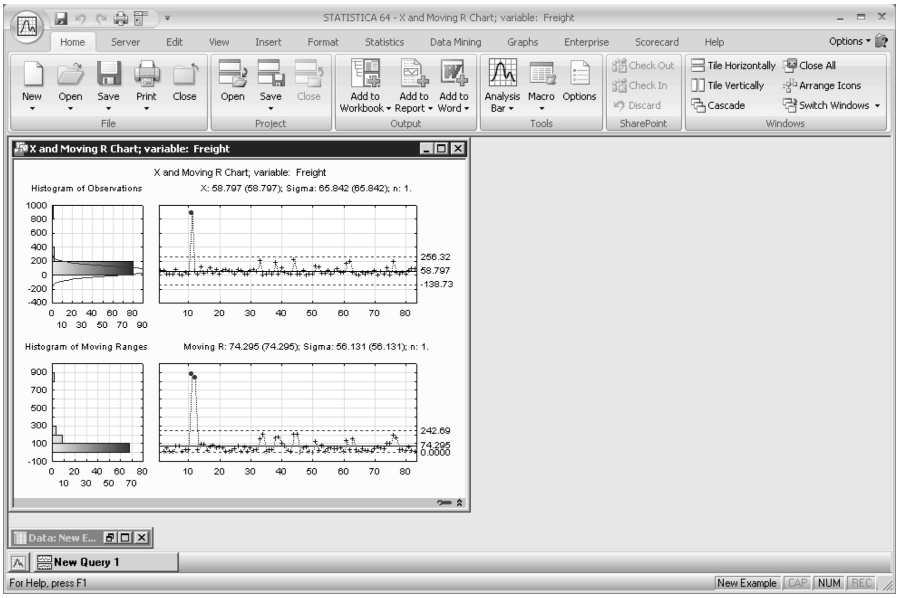
Column: ShipCountry

Operator: equals

Value(s):

ShipCountry equals Brazil

Click the ***Finish*** button to complete the filtering step, extract the data, and perform a packaged analysis on the *Freight* column.


STATISTICA 64 - X and Moving R Chart; variable: Freight

X and Moving R Chart; variable: Freight

X: 58.797 (58.797); Sigma: 65.842 (65.842); n: 1.

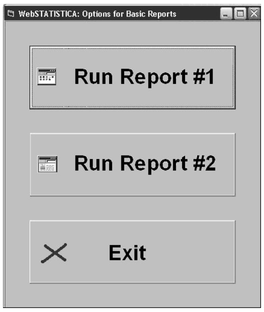
Moving R: 74.295 (74.295); Sigma: 56.131 (56.131); n: 1.

Custom User Interfaces

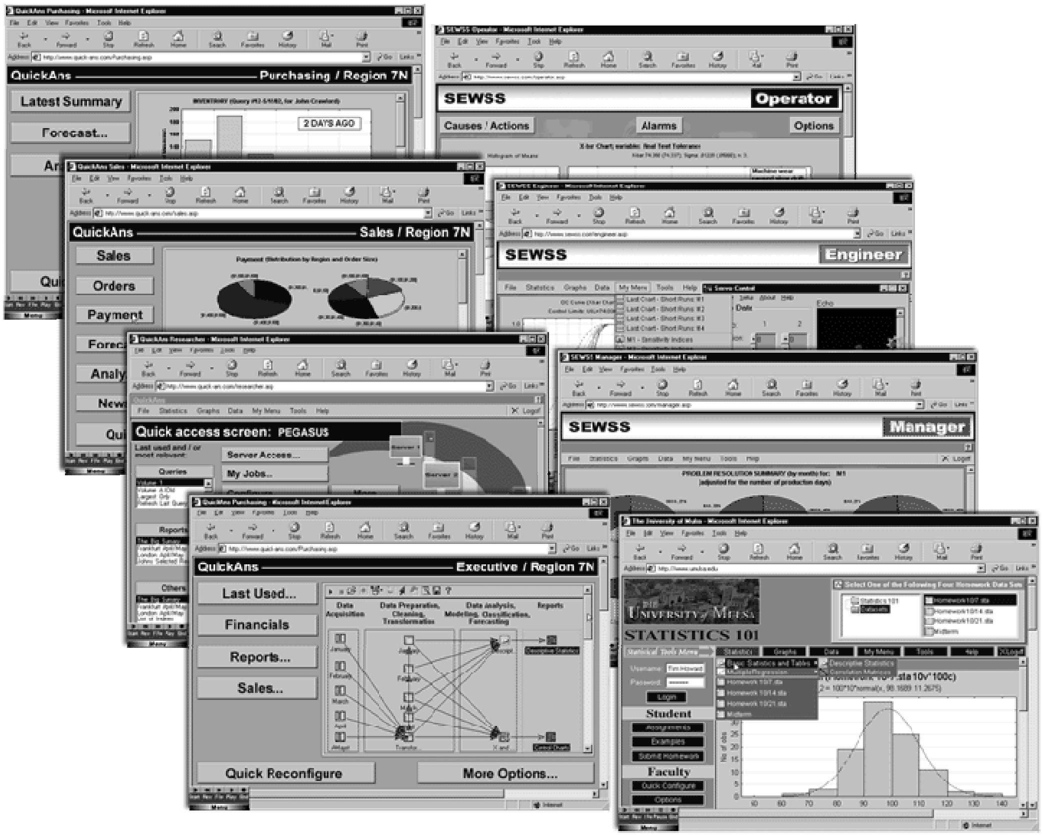
Note that this simple example illustrates how to enable and run an analysis configuration using the standard *STATISTICA* user interface and output components.

However, one of the major strengths of *STATISTICA Enterprise* is the ease of creating custom user interfaces (e.g., for different categories of users depending on their roles in the organization, expertise, or data access privileges).

You can easily create a customized user interface at any degree of complexity, from highly simplified ones, e.g., one that contains only three options:



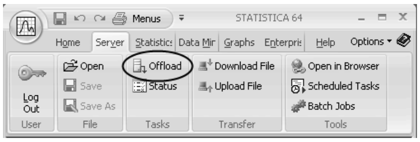
to very elaborate user interfaces of virtually unlimited flexibility:



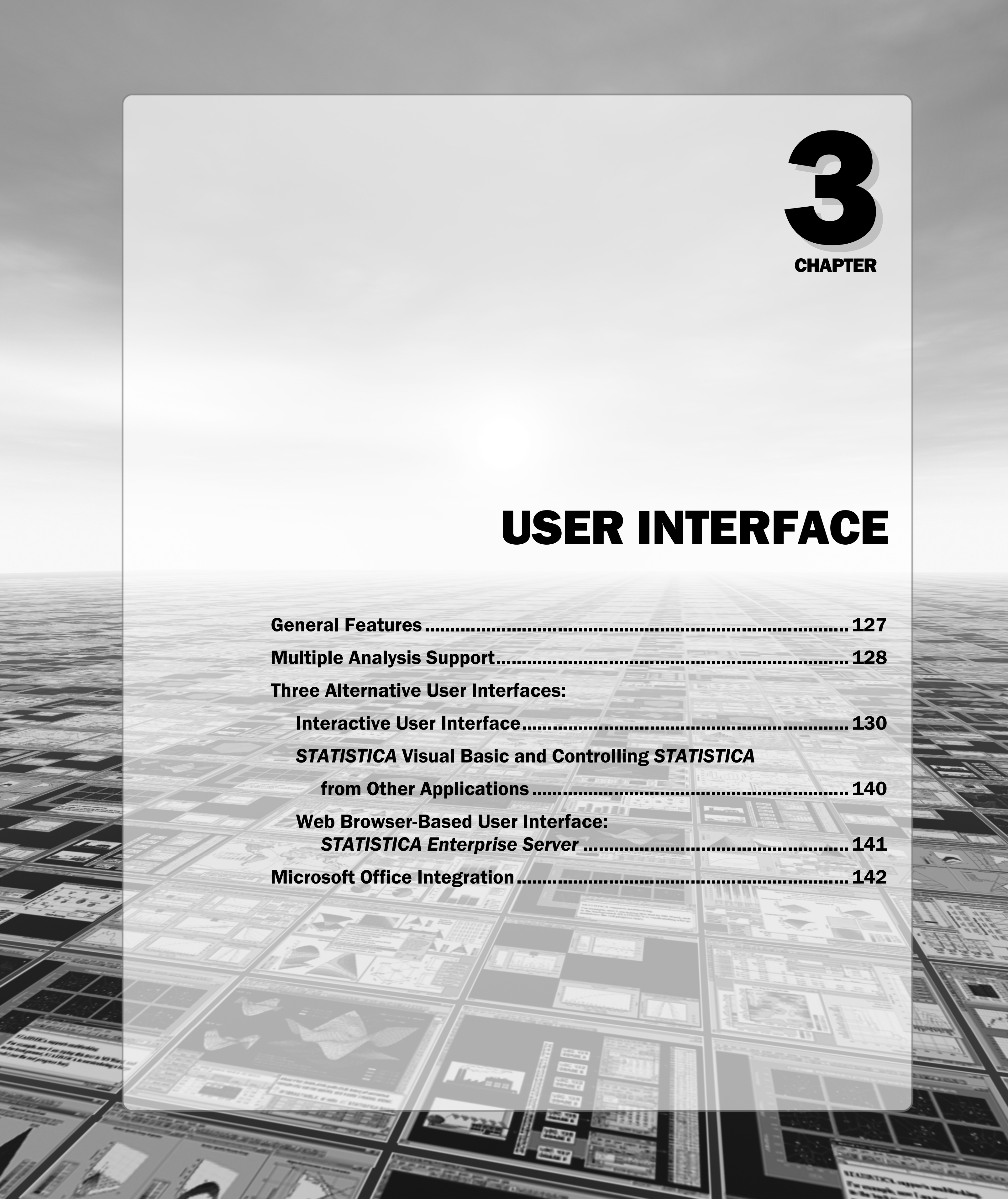
Please refer to the *STATISTICA Enterprise* documentation (*Electronic Manual*) for more details and examples.

The *STATISTICA Enterprise Server* Option

STATISTICA Enterprise Server provides all of the functionality described in this example and also enables offloading tasks to the server and remote access via a browser interface.



See Appendix B – *STATISTICA Enterprise Server*, page 263, for more information.



3

CHAPTER

USER INTERFACE

3

CHAPTER

USER INTERFACE

GENERAL FEATURES

Customized Operation

The *STATISTICA* system can be controlled in several ways. The following sections summarize the features of the main alternative user interfaces of *STATISTICA*:

1. Interactive interface (see page 130)
2. *STATISTICA* Visual Basic (see page 140)
3. Web browser-based interfaces (see page 141)
4. Microsoft Office Integration (see page 142)

However, note that:

- Many aspects of these user interfaces do not exclude each other; thus, depending on your specific applications and preferences, you can combine them;
- The customizable Quick Access Toolbar and classic menus can be used to integrate the alternative user interfaces and, for example, to provide quick access to macro (Visual Basic) programs or commonly used files; and
- Almost all features of these alternative user interfaces can be customized (leading to a different appearance and behavior of *STATISTICA*); it is generally recommended that you customize your system in order to take full advantage of *STATISTICA*'s potential to meet your preferences and

optimal requirements of the tasks that you need to accomplish (see *Customization of the Interactive User Interface* on page 213).

Alternative Access to the Same Facilities - Custom Styles of Work

Even without any customization, the default settings of *STATISTICA* offer alternative user interface means and solutions to achieve the same results. This "alternative access" principle present in every aspect of its user interface enables *STATISTICA* to support different styles of work. For example, most of the commonly used tools can be accessed alternatively:

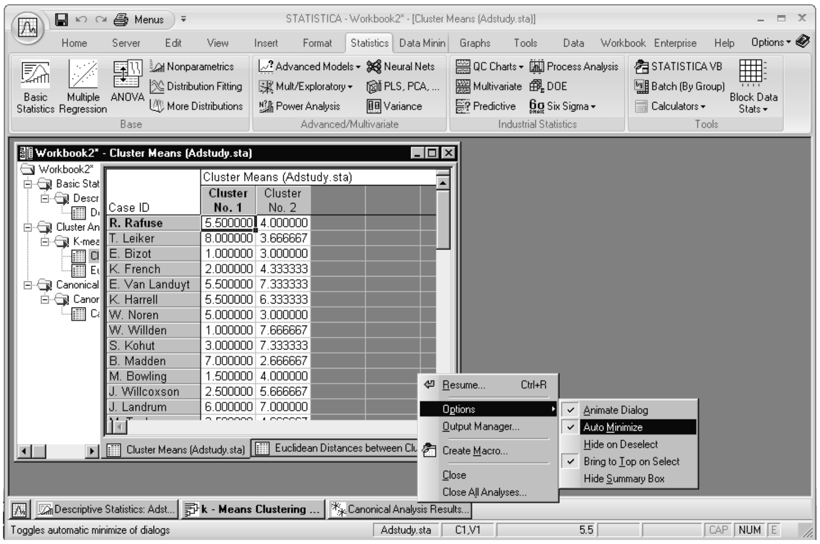
- From the ribbon bar or the classic menus
- Via keyboard shortcuts
- By using the clickable fields on the status bar
- Via the custom Quick Access toolbar (user-defined toolbar with buttons and special controls, which can include macros and commands)
- From the shortcut menus associated with specific objects (cells, workbook icons, parts of graphs), which are displayed by right-clicking on the item.

It is suggested that you explore the alternative user interface facilities of *STATISTICA* before becoming attached to one style or another.

MULTIPLE ANALYSIS SUPPORT

As mentioned before, you can have several instances of *STATISTICA* open at the same time. Each of them can run the same or different types of analyses (traditionally called modules), such as *Basic Statistics*, *Multiple Regression*, *ANOVA,* etc. Moreover, in one *STATISTICA* instance, multiple analyses can be open simultaneously. They can be of the same or a different kind (e.g., five *Multiple Regressions* and two *ANOVAs*), and each of them can be performed on the same or a different input data file (multiple input data files can be opened simultaneously).

Individual "analyses" – functional units of your work. In order to facilitate taking advantage of this multitasking functionality, your work with *STATISTICA* is organized into functional units called "analyses" that are represented with buttons on the analysis bar at the bottom of the application window (above the status bar, see the following illustration, where *Descriptive Statistics*, *Cluster Analysis*, and *Canonical Analysis* are running simultaneously). Consecutive buttons are added as you start new analyses. A variety of options are provided to control (and/or permanently configure) this aspect of *STATISTICA*.



Cluster Means (Adstudy.sta)

Case ID	Cluster No. 1	Cluster No. 2
R. Rafuse	5.500000	4.000000
T. Leiker	8.000000	3.666667
E. Bizot	1.000000	3.000000
K. French	2.000000	4.333333
E. Van Landuyt	5.500000	7.333333
K. Harrell	5.500000	6.333333
W. Noren	5.000000	3.000000
W. Willden	1.000000	7.666667
S. Kohut	3.000000	7.333333
B. Madden	7.000000	2.666667
M. Bowling	1.500000	4.000000
J. Willcoxson	2.500000	5.666667
J. Landrum	6.000000	7.000000

Toggles automatic minimize of dialogs

By default, when you select specific output from a results dialog, the output (a spreadsheet or a graph) is displayed and the dialog is automatically minimized into its respective analysis button at the bottom of the screen. Click that button (or press CTRL+R) to display the dialog again and resume the analysis.

A selection of options pertaining to analysis management are available on the shortcut menu (accessed by right-clicking on an analysis button on the analysis bar) related to each respective analysis button (as shown above).

A useful hint for those with large screens. If you have a large screen, you can turn off the default minimization of the analysis dialogs and take advantage of the fact that most of these dialogs are small and, thus, can remain on the workspace without interfering with the viewing of analysis results. You can adjust this option either for a particular analysis (clear the ***Auto Minimize*** command on the analysis button shortcut menu, shown in the previous image), or globally for the entire program [select ***Analyses/Graphs*** in the tree pane of the ***Options*** dialog (accessible by selecting the ***Tools*** tab and clicking ***Options***), and clear the ***Auto minimize dialogs when displaying output*** check box].

When you run multiple analyses and the *STATISTICA* workspace becomes cluttered, you can hide all windows related to specific analyses (or close them altogether via the analysis button shortcut menu command ***Close All Analyses***). You can also open new *STATISTICA* instances, which offers another simple way to organize and manage your work.

INTERACTIVE USER INTERFACE

Overview

Main components of the interactive user interface of *STATISTICA*. Although the interactive user interface of *STATISTICA* is not the only one available (see Chapter 7 – *Customizing STATISTICA,* page 213, and Chapter 8 – *STATISTICA Visual Basic*, page 219), in most cases it is the easiest and most commonly used. Many components of this user interface can be seen in the *STATISTICA* application window.

First, similar to most software programs, tabs, menu bars and various toolbars are displayed at the top of the window. These are customizable and displayed in the most appropriate manner for your tasks.

At the bottom of the window, the analysis bar (containing minimized analysis/graph dialogs) and the status bar are displayed. Additionally, shortcut menus are available when you right-click in appropriate places.

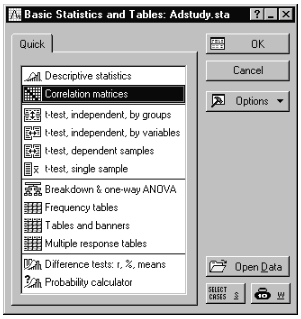
Data files can be displayed in spreadsheets, workbooks, reports, or individual windows. Results spreadsheets or graphs can be displayed in workbooks, reports, or individual windows. Note that additional documents (such as Word or Bitmap images) can also be displayed in spreadsheets, workbooks, or reports. Finally, *STATISTICA* Visual Basic code is displayed in macro windows.

Normally you would not simultaneously see all of these facilities and tools at one time. You always have the ability to make the user interface of *STATISTICA* as simple or complex as your particular needs and comfort level demand (see page 213). These various tools and facilities are described in detail in the *Electronic Manual* (*STATISTICA* Help).

Modules. While *STATISTICA* offers a variety of statistical and graphical procedures, each procedure can be performed in the same instance of *STATISTICA*. This means that, for example, it is possible to calculate residual statistics using options in the *Multiple Regression* module, then immediately use that output in the *Factor Analysis* or another exploratory module without first starting another instance of *STATISTICA*. For more information on using results as input data, see *Can I Use the Results of My Analysis to Perform Another Analysis?* in the *Electronic Manual*.

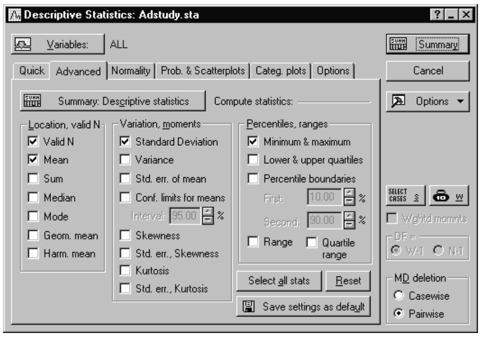
The Flow of Interactive Analysis

Startup Panel. When a statistical procedure is selected from the ***Statistics***, ***Data Mining***, or ***Graphs*** tabs, its respective Startup Panel is displayed (as shown below; ***Basic Statistics*** was selected from the ***Statistics*** tab - ***Base*** group to display the ***Basic Statistics and Tables*** Startup Panel).



Each Startup Panel contains a list of the types of analyses available in that particular module. Clicking anywhere outside the panel automatically minimizes it as a button on the analysis bar. If your system includes a high-resolution screen, you can change this default and keep the consecutive dialogs (in each analysis sequence) displayed on the workspace.

Analysis specification and output selection (results) dialogs. When the desired analysis is selected in the Startup Panel, the analysis specification dialog is displayed, in which you select the variables to be analyzed and other options and features of the task to be performed. Often, these dialogs contain several tabs that group the options, analyses, and/or results in logical categories to make it easier to locate specific features.



Descriptive Statistics: Adstudy.sta

Variables: ALL

Location, valid N
☑ Valid N
☑ Mean
□ Sum
□ Median
□ Mode
□ Geom. mean
□ Harm. mean

Variation, moments
☑ Standard Deviation
□ Variance
□ Std. err. of mean
□ Conf. limits for means
Interval: 95.00 %
□ Skewness
□ Std. err., Skewness
□ Kurtosis
□ Std. err., Kurtosis

Percentiles, ranges
☑ Minimum & maximum
□ Lower & upper quartiles
□ Percentile boundaries
First: 10.00 %
Second: 90.00 %
□ Range □ Quartile range

□ Wghtd momnts
DF = ● W-1 ○ N-1

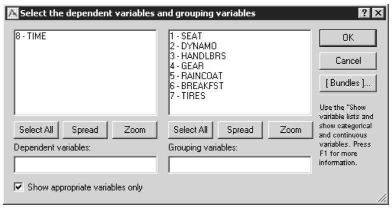
MD deletion
○ Casewise
● Pairwise

In some simple analyses (such as ***Descriptive Statistics***, shown in the illustration above), the analysis specification dialog also serves as an output selection dialog where you can specify the type and format of the output (e.g., specific spreadsheets or graphs). Most analyses, however, have a separate analysis specification dialog and results dialog.

Spreadsheet facilities for scenario (what-if) analyses and customized appearance. *STATISTICA* provides you with the capability to append supplementary information about variable measurement types and case states to your spreadsheets. This "metadata" can be used to create a more comprehensive description of your data set, facilitate "what-if" types of exploratory analyses, and customize the appearance of cases in graphs.

Case states and brushing. You can assign case states to cases in order to customize the appearance of points in graphical displays, thus making it very easy to identify influential and interesting points. A wide selection of symbols and colors is available to customize the appearance of selected points. Not only can case states be assigned in the spreadsheet before a graph is created, they can also be assigned interactively in the graph via the ***Brushing*** facilities (accessible by clicking the ***Brushing*** button in the ***Customize Graph*** group on the ***Edit*** tab when a graph is displayed). The case states assigned in the graph propagate back to the spreadsheet. The ability to assign case states in either the spreadsheet or graph further facilitates the exploratory visual analysis of data.

Measurement types and automatic variable pre-screening. The modeling or measurement type of a variable can be explicitly defined in order to indicate what analyses and graphs are appropriate for such a variable. These measurement types will map directly to subsequent analyses and graphs, identifying appropriate variables in each case (e.g., variables of type categorical will be present within the list of categorical predictors available in a *Factorial ANOVA*).



Select the dependent variables and grouping variables

8 - TIME

1 - SEAT
2 - DYNAMO
3 - HANDLBRS
4 - GEAR
5 - RAINCOAT
6 - BREAKFST
7 - TIRES

Dependent variables:

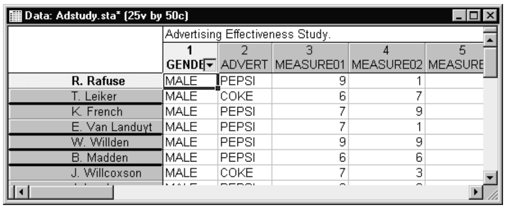
Grouping variables:

Use the "Show variable lists and show categorical and continuous variables. Press F1 for more information.

☑ Show appropriate variables only

In all variable selection dialogs (such as the one shown above), the ***Show appropriate variables only*** option is provided, which enables you to pre-screen or filter variables according to their ***Measurement Type*** (specified in the ***Variable*** specification dialog, accessible by double-clicking on a variable header in a spreadsheet); if that type is ***Auto***, then the ***Automatic variable pre-screening and classification*** options (located in the ***Analysis/Graph*** options pane of the ***Options*** dialog, accessible by selecting the ***Tools*** tab and clicking ***Options***) determine how *STATISTICA* will automatically determine the ***Measurement Type***.

Auto filtering (cloaking variables and cases). Filtering (accessible by selecting the ***Data*** tab and clicking ***Auto Filter*** in the ***Transformations*** group) is a quick and easy way to display a specific portion of the data in your spreadsheet without sorting the data or creating a subset. When a variable is filtered, only the values that meet the specified criteria are displayed in the spreadsheet. Cases that do not meet the criteria are hidden from sight but not removed from the spreadsheet (e.g., in the spreadsheet shown below, only the cases for *GENDER = MALE* are displayed).



Data: Adstudy.sta* (25v by 50c)

	Advertising Effectiveness Study.				
	1 GENDE	2 ADVERT	3 MEASURE01	4 MEASURE02	5 MEASURE
R. Rafuse	MALE	PEPSI	9	1	
T. Leiker	MALE	COKE	6	7	
K. French	MALE	PEPSI	7	9	
E. Van Landuyt	MALE	PEPSI	7	1	
W. Willden	MALE	PEPSI	9	9	
B. Madden	MALE	PEPSI	6	6	
J. Willcoxson	MALE	COKE	7	3	

Although hidden, they are still available for statistical and graphical analyses.

Output. As described in more detail in Chapter 4 – *Five Channels for Output From Analyses* (page 147) and as illustrated in *Example 1: Correlations* (page 11) and *Example 2: ANOVA* (page 34), the consecutive output spreadsheets and graphs are displayed in workbooks by default. These workbooks can be saved and later reopened, making it easy to return to specific results as needed.

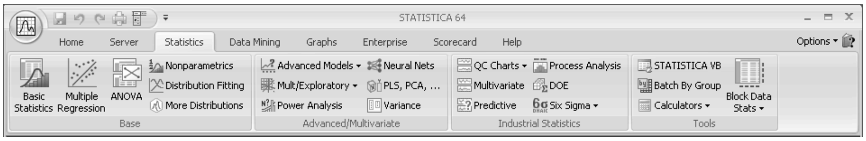
Additionally, you can send all output to an analysis report (see page 151), which produces an easily organized (via the report tree), formatted, and printed report of a specific analysis. You can also choose to send all results, regardless of what analysis it comes from, to a single report. Lastly, the output can be directed to separate windows.

To specify output options for a single analysis or session, click the Options ▾ button in the analysis or graph specification dialog and select ***Output*** to display the ***Analysis/Graph Output Manager*** dialog.

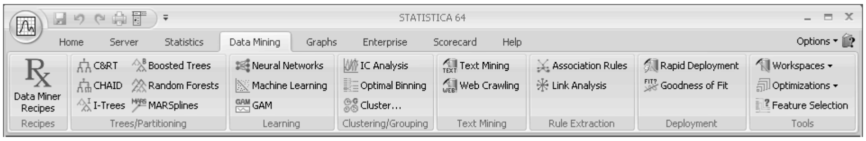
To access global output options, select the ***Tools*** tab. Click ***Options*** to display the ***Options*** dialog, and select ***Output Manager***. Or, select the ***Home*** tab and click ***Options*** in the ***Tools*** group. For more information, see the *Electronic Manual*.

Features of Analyses

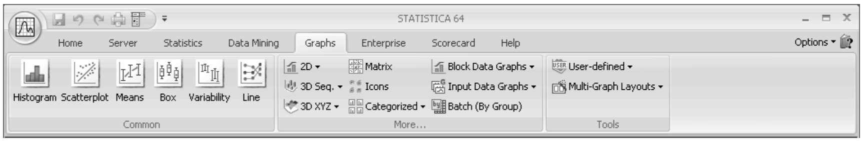
STATISTICA provides direct access to all statistical analyses via the ***Statistics*** tab:



and the ***Data Mining*** tab:



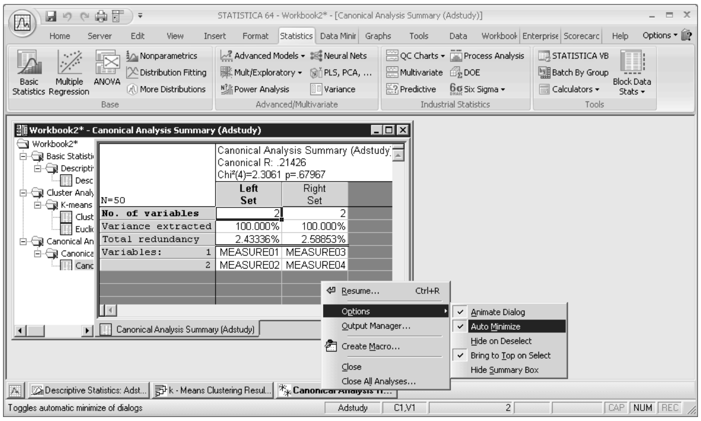
and provides direct access to all graphical analysis dialogs via the ***Graphs*** tab:



These tabs are never disabled, i.e., they are available whenever any input data document is open.

The ***Statistics*** and ***Data Mining*** tabs provide access to all available analysis types within *STATISTICA*. The ***Graphs*** tab provides direct access to a variety of commonly used graph types (e.g., scatterplots, histograms, means/error plots, etc.) as well as hierarchical access to all graph types in *STATISTICA* including ***2D Graphs***, ***3D Sequential*** and ***XYZ Graphs***, ***Categorized Graphs***, ***User-defined Graphs***, ***Block Data Graphs, Input Data Graphs***, and ***Multi-Graph Layouts***. Comprehensive discussions of all the various types of statistics and graphs offered by *STATISTICA* are available in the glossary of the *Electronic Manual*. See also, Appendix C: *STATISTICA Family of Products* (page 275) for more information on all members of the comprehensive selection of data analysis applications from the *STATISTICA* family of products.

Using the analysis bar. To take advantage of *STATISTICA*'s "multitasking" functionality (see *Multiple Analysis Support*, page 128), *STATISTICA* analyses are organized as functional units that are represented with buttons on the analysis bar at the bottom of the application window (above the status bar, see the next illustration, where *Descriptive Statistics*, *Cluster Analysis*, and *Canonical Analysis* are running simultaneously). Consecutive buttons are added as you start new analyses.

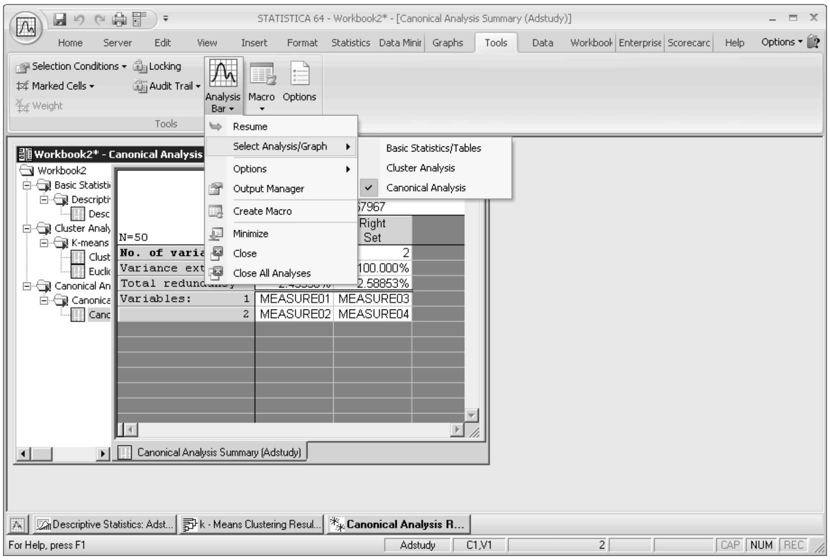


Canonical Analysis Summary (Adstudy)
Canonical R: .21426
Chi²(4)=2.3061 p=.67967

N=50	Left Set	Right Set
No. of variables	2	2
Variance extracted	100.000%	100.000%
Total redundancy	2.43336%	2.58853%
Variables: 1	MEASURE01	MEASURE03
2	MEASURE02	MEASURE04

Minimizing dialogs (and a hint for users with large screens). Depending on your preferences, you can choose to minimize all analysis dialogs when you select another window in *STATISTICA* or another application. By default the ***Auto Minimize*** command is selected; however, when your screen is large enough to accommodate several windows, it is recommended that you clear this command. This keeps the analysis dialogs on screen while the respective output created from these dialogs is produced, thus enabling you to use the dialogs as "toolbars" from which output can be selected. See page 129 for information on how to adjust this command.

Continuing analyses/graphs. It is easy to continue an analysis or graph (i.e., to change the focus to the current dialog for a particular analysis). Select the ***Tools*** tab, click ***Analysis Bar***, and select ***Resume*** from the drop-down menu; or press CTRL+R; or click the analysis/graph button on the analysis bar. When multiple analyses are running, you can also select the specific analysis from the ***Select Analysis/Graph*** submenu (as shown in the next illustration).



Hiding windows. To further facilitate the organization of windows from various analyses, you can hide all windows associated with a particular analysis when that analysis is deselected: select the ***Tools*** tab, click ***Analysis Bar***, and from the ***Options*** submenu, select ***Hide on Deselect***. By default, this command is not selected. Note that this command only applies when the results are sent to individual windows; see the discussion of the ***Output Manager*** (page 147) for more details on managing output from analyses. In addition, there is a command on the ***Home*** tab in the ***Windows*** group to close all document windows: click ***Close All*** (or press CTRL+L on your keyboard), and a command on the ***Tools*** tab to close all analyses: click ***Analysis Bar*** and select ***Close All Analyses*** from the drop-down menu.

Bringing windows to the top. On the ***Tools*** tab click ***Analysis Bar***, and from the ***Options*** submenu select ***Bring to Top on Select*** to activate (bring to the top of *STATISTICA*) all windows associated with a particular analysis when that analysis is selected, replacing whatever dialogs were on top. This command also facilitates the organization of individual windows from various analyses. By default, this command is selected. Note that this command only applies when the results are sent to individual windows; see the discussion of the ***Output Manager*** (page 147) for more details on managing output from analyses.

Hiding the summary box. By default, a summary box is located at the top of certain results dialogs (such as ***Multiple Regression Results***) and contains basic summary information about the analysis. You can hide an individual summary box by clicking the ⏫ button in the lower-right corner of the summary box. You can also suppress the display of all summary boxes globally by selecting the ***Tools*** tab, clicking ***Analysis Bar***, and selecting ***Hide Summary Box*** from the ***Options*** submenu.

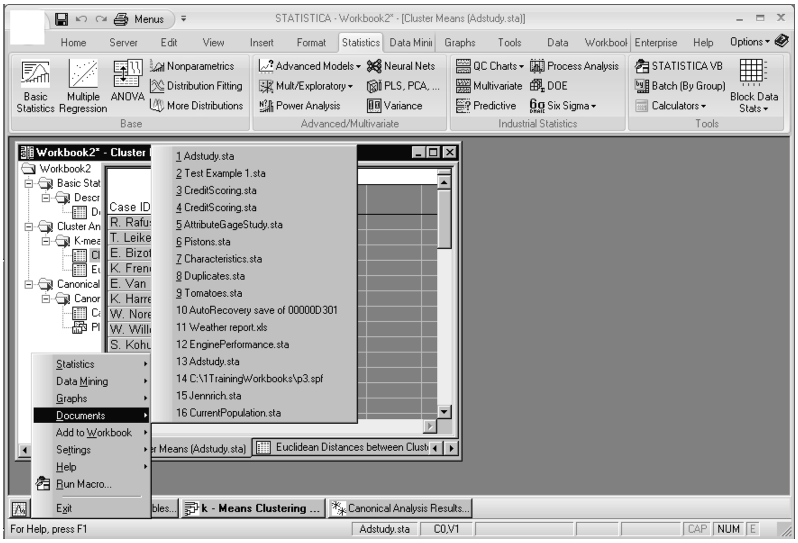
Document Types

STATISTICA uses seven principal document types:

- Workbooks (see pages 148 and 169)
- Spreadsheets (multimedia tables) (see page 173)
- Reports (see pages 151 and 180)
- Graphs (see pages 182 and 189)
- Macros (*STATISTICA* Visual Basic programs) (see pages 183 and 219)
- *STATISTICA* Project Files (see page 184)
- *Data Miner Recipes* Project Files (see page 61)

Using these seven document types, you can manage data of various types, perform data entry and analyses, generate graphs of the highest quality, develop custom applications of any degree of complexity, and create custom-formatted reports.

You can quickly access the most recently used documents. Click the *STATISTICA* Start menu (in the lower-left corner of the screen) and select ***Documents***.



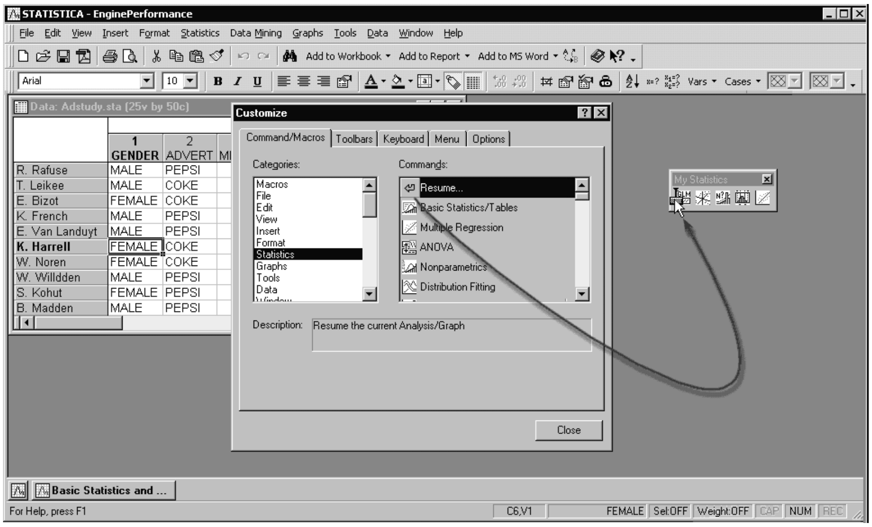
In the ***General*** options pane of the ***Options*** dialog (accessible by selecting the ***Tools*** tab and clicking ***Options***), you can specify how many recently used documents to display (the default is *16*). For more detailed information about each document type, see the overviews for workbooks, spreadsheets, reports, graphs, and macros on page 169; for further information, see the *Electronic Manual*.

Tabs related to types of active document windows. Each of the main types of *STATISTICA* document windows (see page 137) manages data in a different way and, thus, offers different customization and management options. These differences are reflected in the tabs that accompany each type of window. Menu commands and buttons for each of the main types of documents are described in detail in the *Electronic Manual*.

The tabs that are available when workbooks are open depend on the type of document that is currently selected in the workbook. Therefore, when you are editing a spreadsheet, graph, report, or macro within a workbook, the tabs relevant for that document type are available. When you select an “empty node” in the workbook tree pane, by default, the ***Workbook*** tab is displayed.

User-defined toolbars. In addition to the variety of toolbars provided on the *STATISTICA* classic menus (on the ribbon bar, click the ▤ icon in the upper-left corner to display the classic menus), you can also create user-defined toolbars. These toolbars can include any command available in *STATISTICA*, as well as special controls (i.e., font name, font size, graph styles, etc.). The toolbars can be given any name and can be designated to open depending on the active document type. Also, you can customize all toolbars (including existing toolbars) by adding commands and special controls.

To create a toolbar (or edit an existing one) use the options on the ***Toolbars*** tab of the ***Customize*** dialog, accessible by selecting ***Customize*** from the ***Tools*** menu. Customizing a toolbar is as easy as dragging commands from the dialog to the toolbar, as shown in the illustration below.



	1 GENDER	2 ADVERT
R. Rafuse	MALE	PEPSI
T. Leikee	MALE	COKE
E. Bizot	FEMALE	COKE
K. French	MALE	PEPSI
E. Van Landuyt	MALE	PEPSI
K. Harrell	FEMALE	COKE
W. Noren	FEMALE	COKE
W. Willdden	MALE	PEPSI
S. Kohut	FEMALE	PEPSI
B. Madden	MALE	PEPSI

Description: Resume the current Analysis/Graph

Shapes and locations of toolbars can be easily adjusted (e.g., all toolbars can be docked or free floating). All of these options make it possible for you to create unique toolbars that provide you with a very specialized user interface. The *Electronic Manual* includes simple-to-follow, step-by-step instructions on how to make customizations. Specifically, see *Create a New Toolbar* in the *Electronic Manual* for more details.

The Quick Access toolbar located at the top of the ribbon bar can be customized as well; see *Customize Quick Access Toolbar* in the *Electronic Manual*.

User-defined menus. Customizing the classic menus is equally easy and can be performed using the ***Menu*** tab of the ***Customize*** dialog (see the *Electronic Manual* for details).

STATISTICA VISUAL BASIC AND CONTROLLING *STATISTICA* FROM OTHER APPLICATIONS

The industry standard *STATISTICA* Visual Basic language (integrated into *STATISTICA*) provides an alternative user interface to the entire functionality of *STATISTICA*, and it offers incomparably more than just a “supplementary application programming language” that can be used to write custom extensions. *STATISTICA* Visual Basic takes full advantage of the object model architecture of *STATISTICA* and can be used to access programmatically every aspect and virtually every detail of the functionality of *STATISTICA*. Even the most complex analyses and graphs can be recorded into Visual Basic macros and later be run repeatedly or edited and used as building blocks of other applications. *STATISTICA* Visual Basic adds an arsenal of more than 14,000 new functions to the standard comprehensive syntax of Visual Basic, thus comprising one of the largest and richest development environments available. For more information on *STATISTICA* Visual Basic, see Chapter 8 (page 219).

Controlling *STATISTICA* from other applications. One of the features that makes the *STATISTICA* Visual Basic environment so powerful is the ability to integrate and manipulate various applications and their environments within a single macro. For example, you can record or write a *STATISTICA* Visual Basic program that computes predictions via the *STATISTICA Time Series* module and execute that program from within an Excel spreadsheet or a Word document. The exchange of information between different applications is accomplished by exposing those applications to the Visual Basic programs as Objects. So, for example, you can run statistical analyses in the *STATISTICA Basic Statistics* module from a Visual Basic program in Excel by declaring inside the program an object of type *Statistica.Application*.

Once an object has been created, the Visual Basic program then has access to the properties and methods contained in that object. Properties can be mostly thought of as functions, methods can be mostly thought of as subroutines that perform certain operations or computations inside the respective application object. You can call *STATISTICA* procedures directly from many other applications and programming languages (e.g., C++, Java, and others).

WEB BROWSER-BASED USER INTERFACE: *STATISTICA ENTERPRISE SERVER*

In addition to the two basic types of user interfaces described in the previous sections, the entire *STATISTICA* family of products also optionally offers a browser-based user interface, where all interactions with the application involving querying databases, data management operations, data analysis, or data mining, as well as generating reports and collaborative work, can be performed without having any *STATISTICA* application installed on the local computer, using only a browser. This alternative user interface requires that a Client-Server version of the respective *STATISTICA* application be installed.

STATISTICA Enterprise Server is a highly scalable, enterprise-level, fully Web-enabled data analysis and database gateway application system that is built on distributed processing technology and fully supports multi-tier Client-Server architecture configurations. *STATISTICA Enterprise Server* exposes the analytic, query, reporting, and graphics functionality of *STATISTICA* through easy-to-use, interactive, standard Web interfaces. Alternatively, it enables users of the desktop version (“thick client”) to offload computationally intensive analytics and database operations to the Server. It is offered as a complete, ready-to-install application with an interactive, Internet browser-based (“point-and-click”) user interface (“thin client”) that makes it possible for users to interactively create data sets, run analyses, and review output. However, *STATISTICA Enterprise Server* is built using open architecture and includes .NET-compatible development kit tools (based entirely on industry standard syntax conventions such as VB Script, C++/C#, HTML, Java, and XML) that enables IT department personnel to customize all main components of the system or expand it by building on its foundations, for example, by adding new components and/or company-specific analytic or database facilities.

As mentioned, *STATISTICA Server* is provided with an Internet browser-based user interface (in the form of simple-to-navigate and easy-to-use dialogs) enabling you to specify analyses and review results. However, tools are provided to customize these dialogs and easily set up new user interfaces or to add new functions. For example, a simple dialog with only three buttons can be created in the browser, and clicking each button will run a series of analyses and generate a detailed report. *STATISTICA Enterprise Server* applications add a new dimension and an endless array of possibilities to the entire line of *STATISTICA* Data Analysis, Data Mining, and Quality Control/Six Sigma software.

The system is compatible with all major Web server software platforms (e.g., UNIX Apache, and Microsoft IIS), works in both Microsoft .NET and Sun/Java environments, and does not require any changes to the existing firewall and Internet/Intranet security systems

For more information, please refer to *Appendix B – STATISTICA Enterprise Server*, page 263.

MICROSOFT OFFICE INTEGRATION

If Microsoft Office is installed on the same machine as *STATISTICA*, Excel spreadsheets can be opened directly within *STATISTICA* and used as a data source for analyses, and Word documents can be used as a destination for reports (see page143; see also page 154).

Excel as a data source. *STATISTICA* can open Excel documents in the *STATISTICA* workspace through the standard ***Open*** dialog. When an Excel workbook is selected, a dialog will be displayed that enables you to import the file into a standard *STATISTICA* Spreadsheet or to keep the document in Excel form, i.e., as an Excel window within *STATISTICA*.

Once the Excel document is opened, you have access to all the menus and toolbars that Excel supports. Thus, you can edit and update formulas, change the formatting, copy/paste, drag/drop – everything that you would normally do if you were within the Excel application.

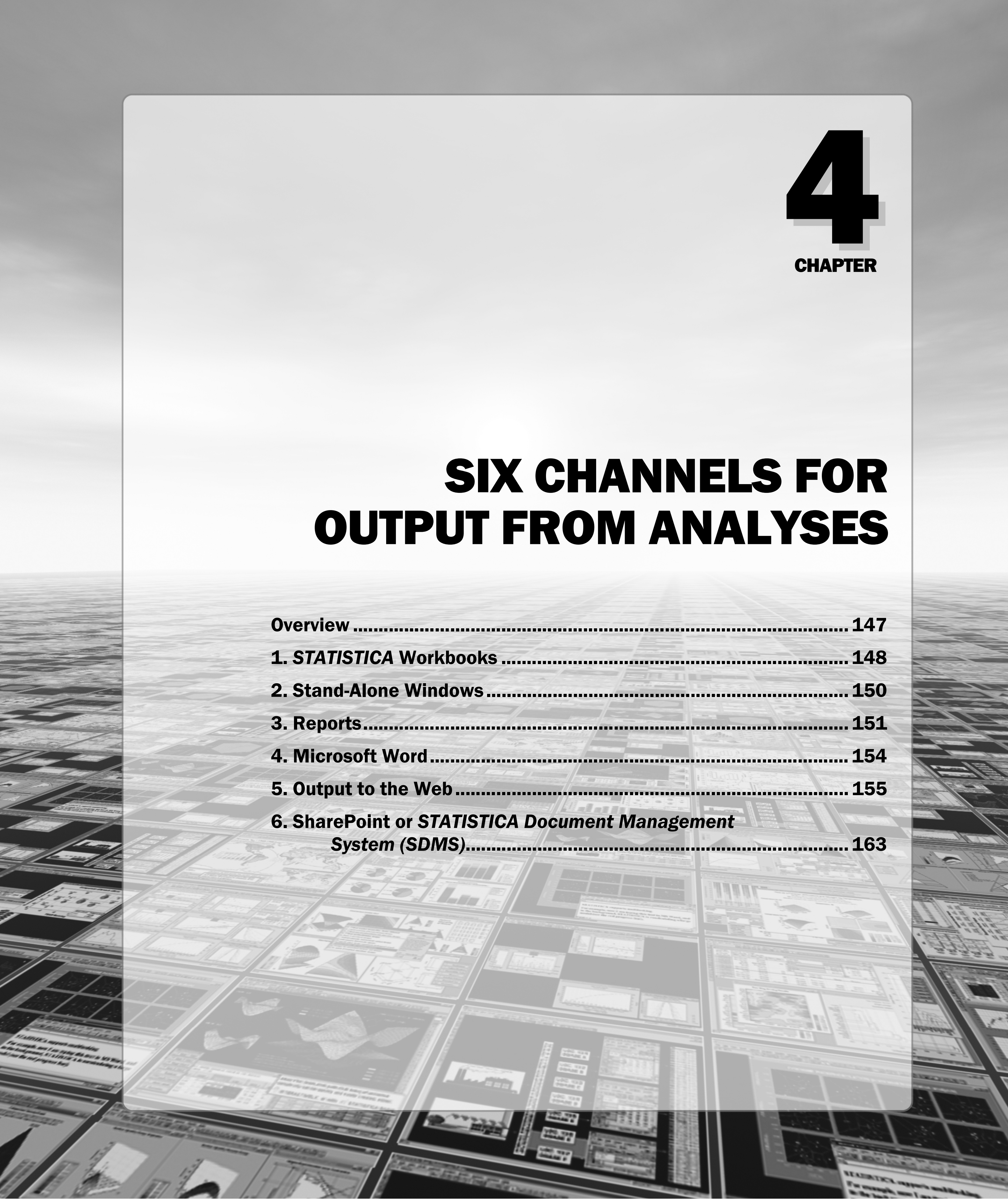
The main strength in Excel integration is that the Excel documents can be used as a data source for analyses. Simply have the Excel document window selected when starting an analysis, and the analysis will source from the Excel document. When initially running the analysis, *STATISTICA* will display a dialog in which you can specify what range of the Excel document should be used as the data source and if a particular row or column is to be used as variable names or case names. These settings are assigned to the Excel document so you will only need to specify them once.

Not only can *STATISTICA* use the Excel file as a data source, but auto updating can be specified as well. If you create an auto-updating graph and then change the Excel file by entering new data or re-evaluating formulas, the graph will also be updated.

Word as a report destination. You can also open and edit Word documents within the *STATISTICA* workspace. Word documents can be opened using the standard ***Open*** dialog, and when performing statistical analyses or creating graphs, output can being directed to a Word document. Any output that can be directed to a *STATISTICA* Report is capable of being directed to a Word document.

As with Excel, when the Word document is open, you have access to all the toolbars and menus that are supported within the Word application. You can perform any formatting and editing that Word supports within its application.

When sending spreadsheet analytical results to Word, *STATISTICA* will take advantage of Word's table editing facility and convert the spreadsheet into a table. For multi-page spreadsheets, you can control where to break the rows and columns. These spreadsheets will be broken by columns such as will be allowed without exceeding the page width. All rows for a given set of columns will be rendered before the next set of spreadsheet columns is rendered in the Word document. This solution enables the presentation of spreadsheets in Word that are natively editable in Word, display the entire contents of the spreadsheet, and print and paginate correctly.



4

CHAPTER

SIX CHANNELS FOR OUTPUT FROM ANALYSES

4

CHAPTER

SIX CHANNELS FOR OUTPUT FROM ANALYSES

OVERVIEW

When you perform an analysis, *STATISTICA* generates output in the form of multimedia tables (spreadsheets) and graphs. There are six basic channels to which you can direct all output:

1. *STATISTICA* Workbooks (page 148)
2. Stand-alone Windows (page 150)
3. Reports (page 151)
4. Microsoft Word (page 154)
5. The Web (page 155)
6. SharePoint or *STATISTICA Document Management System (SDMS)* (page 163)

The first four output channels listed above are controlled by the options in the ***Output Manager*** (accessible by selecting ***Output Manager*** from the Start button drop-down menu located in the upper-left corner of the ribbon bar, see page 23 for further details on both the global ***Output Manager*** in the ***Options*** dialog and the ***Analysis/Graph Output Manager*** dialog). There are a number of ways to output to the Web, depending on the version of *STATISTICA* you have. SharePoint is accessible from within *STATISTICA*, and *SDMS* is an additional product available from StatSoft.

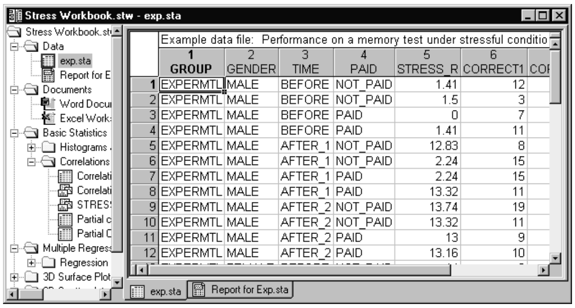
These means for output can be used in many combinations (e.g., a workbook and report simultaneously) and can be customized in a variety of ways. Also, all output objects (spreadsheets and graphs) placed in each of the output channels can contain other embedded and linked objects and documents, so *STATISTICA* output can be hierarchically organized in a variety of ways. Each of the *STATISTICA* output channels has its unique advantages, as described in the following sections. More comprehensive overviews of each of the document types associated with the respective channels of output are included in Chapter 5 - *STATISTICA Documents* (page 167).

The auto save and recovery features. All *STATISTICA* documents (i.e., input spreadsheets, workbooks, reports, and macros) that accumulate the results of your work (e.g., data entry, editing, or output collection) over an extended period of time support the ***Auto Save*** feature, which is configurable in the ***General*** options pane of the ***Options*** dialog (accessible by selecting the ***Tools*** tab and clicking ***Options***). This facility will automatically save the contents of your work periodically (e.g., every 10 minutes) and, thus, give you the option to retrieve data that otherwise could be lost in case of a power outage or a system failure.

1. *STATISTICA* WORKBOOKS

Workbooks are the default way of managing output (for more information, see page 169). Each output document (e.g., a *STATISTICA* Spreadsheet or Graph, as well as a Word or Excel document) is stored as a tab in the workbook.

Documents can be organized into hierarchies of folders or document nodes (by default, one is created for each new analysis) using a tree view, in which individual documents, folders, or entire branches of the tree can be flexibly managed.



Example data file: Performance on a memory test under stressful conditio

	1 GROUP	2 GENDER	3 TIME	4 PAID	5 STRESS_R	6 CORRECT1	COI
1	EXPERMTL	MALE	BEFORE	NOT_PAID	1.41	12	
2	EXPERMTL	MALE	BEFORE	NOT_PAID	1.5	3	
3	EXPERMTL	MALE	BEFORE	PAID	0	7	
4	EXPERMTL	MALE	BEFORE	PAID	1.41	11	
5	EXPERMTL	MALE	AFTER_1	NOT_PAID	12.83	8	
6	EXPERMTL	MALE	AFTER_1	NOT_PAID	2.24	15	
7	EXPERMTL	MALE	AFTER_1	PAID	2.24	15	
8	EXPERMTL	MALE	AFTER_1	PAID	13.32	11	
9	EXPERMTL	MALE	AFTER_2	NOT_PAID	13.74	19	
10	EXPERMTL	MALE	AFTER_2	NOT_PAID	13.32	11	
11	EXPERMTL	MALE	AFTER_2	PAID	13	9	
12	EXPERMTL	MALE	AFTER_2	PAID	13.16	10	

For example, selections of documents can be extracted (e.g., drag-copied or drag-moved) to a report window or to the application workspace (i.e., the *STATISTICA* application "background" where they will be displayed in stand-alone windows). Entire branches can be placed into other workbooks in a variety of ways in order to build specific folder organization, etc.

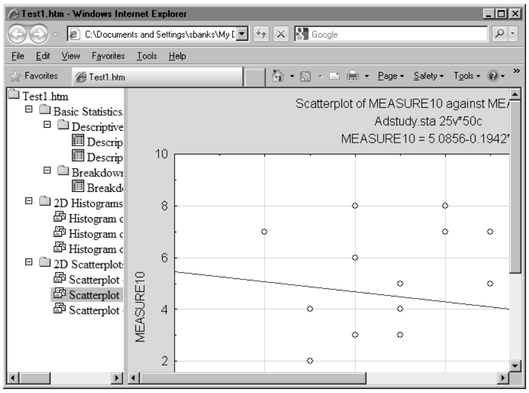
Technically speaking, workbooks are ActiveX document containers (see page 238 for information on ActiveX technology, see also the *Electronic Manual*). Workbooks are compatible with a variety of foreign file formats (e.g., Office documents) that can be easily inserted into workbooks and in-place edited.

User notes and comments in workbooks. Workbooks offer powerful options to efficiently manage even extremely large amounts of output, and they may be the best output handling solution for both novices and advanced users. It might appear that one possible drawback is that user comments (e.g., notes) and supplementary information cannot be as transparently inserted into the "stream" of the workbook output as they can in traditional, word processor style reports, such as *STATISTICA* Reports (see the next section). However, note that:

- All *STATISTICA* documents can easily be annotated, both a) directly, by typing text into graphs, tables, and reports, and b) indirectly, by entering notes into the ***Comments*** box of the ***Document Properties*** dialog (accessed by selecting ***Properties*** from the Start button drop-down menu located in the upper-left corner of the ribbon bar), and
- Formatted documents with notes and comments (in the form of text files, *STATISTICA* Report documents, WordPad or word processor documents, etc.) can easily be inserted anywhere in the hierarchical organization of output in workbooks. Moreover, such summary notes or comment documents can be made nodes for groups of subordinate objects to which the note is related to further enhance their organization.

Saving workbooks as Web pages. Workbooks can be saved as **.html* (Web) files by selecting ***Save As*** on the ***Home*** tab in the ***File*** group from the ***Save*** menu, and in the ***Save As*** dialog, choosing ***Web Page (*.htm; *.html)*** from the ***Save as type*** drop-down list. Saving as a Web page will create an **.html* file in the specified directory that can be opened with standard internet browsers such as Microsoft

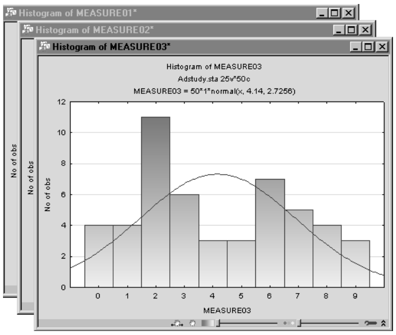
Internet Explorer. When saving the workbook as a Web page, *STATISTICA* also creates a subdirectory that contains all the images referenced by the Web page.

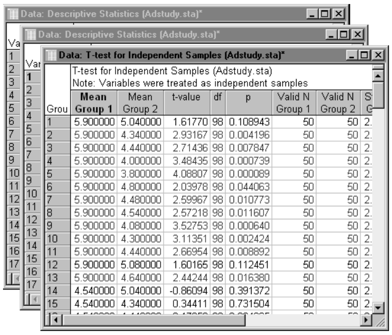


The Web page output contains an *.html*-based tree control that enables you to navigate and display the various workbook images, similar to the actual workbook.

2. STAND-ALONE WINDOWS

STATISTICA output documents can also be directed to a queue of stand-alone windows; the ***Queue Length*** can be controlled in the ***Output Manager*** options pane of the ***Options*** dialog (accessible by selecting the ***Tools*** tab and clicking ***Options***).




Data: Descriptive Statistics (Adstudy.sta)*

Data: Descriptive Statistics (Adstudy.sta)*

Data: T-test for Independent Samples (Adstudy.sta)*

T-test for Independent Samples (Adstudy.sta)
Note: Variables were treated as independent samples

Grou	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	S G
1	5.900000	5.040000	1.61770	98	0.108943	50	50	2.
2	5.900000	4.340000	2.93167	98	0.004196	50	50	2.
3	5.900000	4.440000	2.71436	98	0.007847	50	50	2.
4	5.900000	4.000000	3.48435	98	0.000739	50	50	2.
5	5.900000	3.800000	4.08807	98	0.000089	50	50	2.
6	5.900000	4.800000	2.03978	98	0.044063	50	50	2.
7	5.900000	4.480000	2.59967	98	0.010773	50	50	2.
8	5.900000	4.540000	2.57218	98	0.011607	50	50	2.
9	5.900000	4.080000	3.52753	98	0.000640	50	50	2.
10	5.900000	4.300000	3.11351	98	0.002424	50	50	2.
11	5.900000	4.440000	2.66954	98	0.008892	50	50	2.
12	5.900000	5.080000	1.60165	98	0.112451	50	50	2.
13	5.900000	4.640000	2.44244	98	0.016380	50	50	2.
14	4.540000	5.040000	-0.86094	98	0.391372	50	50	2.
15	4.540000	4.340000	0.34411	98	0.731504	50	50	2.

The clear disadvantage of this output mode is its total lack of organization and its natural tendency to clutter the application workspace (some procedures can generate hundreds of tables or graphs with a click of the button).

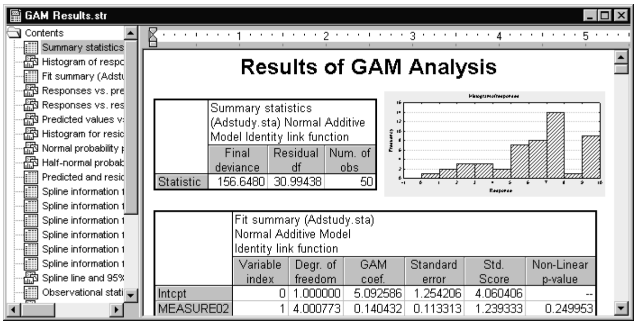
One of the advantages of this way of handling output is that you can easily custom arrange these objects within the *STATISTICA* application workspace (e.g., to create multiple, easy to identify "reference documents" to be compared to the new output). However, note that in order to achieve that effect, you do not need to configure the output ahead of time and generate a large number of (mostly unwanted) separate windows that can clutter the workspace. Instead, individual, specific output objects directed to and stored in the other two channels (workbooks and reports) can easily be dragged out from their respective tree views onto the application workspace as needed.

3. REPORTS

When performing an analysis, the ultimate goal is to create meaningful output in order to gain an understanding of the data. The manner in which the output is produced is important as well. *STATISTICA* offers a variety of methods to produce reports that accommodate the diverse needs of users.

STATISTICA Reports

STATISTICA Reports (for more information, see page 180) offer a more traditional way of handling output where each object (e.g., a *STATISTICA* Spreadsheet or Graph, or an Excel spreadsheet) is displayed sequentially in a word processor style document.



Results of GAM Analysis

	Summary statistics (Adstudy.sta) Normal Additive Model Identity link function		
	Final deviance	Residual df	Num. of obs
Statistic	156.6480	30.99438	50

	Fit summary (Adstudy.sta) Normal Additive Model Identity link function					
	Variable index	Degr. of freedom	GAM coef.	Standard error	Std. Score	Non-Linear p-value
Intcpt	0	1.000000	5.092586	1.254206	4.060406	--
MEASURE02	1	4.000773	0.140432	0.113313	1.239333	0.249953

However, the technology behind this simple editor offers you very rich functionality. For example, like the workbook (see *STATISTICA Workbooks*, page 148), the *STATISTICA* Report is also an ActiveX container (for information on

ActiveX technology, see page 238 or the *Electronic Manual*) where each of its objects (not only *STATISTICA* Spreadsheets and Graphs, but also any other ActiveX-compatible documents, e.g., Excel spreadsheets) remains active, customizable, and in-place editable.

The obvious advantages of this way of handling output (more traditional than the workbook) are the ability to insert notes and comments in between the objects as well as its support for the more traditional way of quick scrolling through and reviewing the output to which some users may be accustomed (e.g., the editor supports variable speed scrolling). Also, only the report output includes and preserves the record of the supplementary information, which contains a detailed log of the options specified for the analyses (e.g., selected variables and their labels, long names, etc., depending on the level of supplementary information specified in the ***Output Manager***, see page 25).

The obvious drawback, however, of these traditional reports is the inherent flat structure imposed by their word processor style format, although that is what some users or certain applications may favor.

Reports from Workbooks

When you have a *STATISTICA* Workbook containing analyses output, you may decide you want to transfer it to a report.

Open a *STATISTICA* Workbook and select all of the files, i.e., select the first file, press the SHIFT key on your keyboard, and select the last file. Then, click ***Add to Report*** on the ***Home*** tab in the ***Output*** group. All the files in the workbook will be duplicated in a *STATISTICA* Report.

RTF (Rich Text Format) Reports

RTF (Rich Text Format) is a Microsoft standard method of encoding formatted text and graphics for easy transfer between applications. When reports are saved in Rich Text Format (**.rtf*), all file formatting is preserved so that it can be read and interpreted by other RTF-compatible applications (e.g., Word).

The *STATISTICA* Report format (*.str*) adheres to RTF conventions; however, saving reports in the default *STATISTICA* Report format ensures that the reports will be opened in *STATISTICA*, giving you complete access to the report tree.

In order to open a *STATISTICA* report in an RTF-compatible application, open the report, select the ***Home*** tab, click the ***Save*** arrow, and select ***Save As*** from the drop-down menu to display the ***Save As*** dialog. From ***the Save as type*** drop-down list, select ***Rich Text Files (*.rtf)***, enter a name in the ***File name*** field, and click the ***Save*** button. You can then open the file in any RTF-compatible application.

Acrobat (PDF) Reports

PDF is the acronym for Portable Document Format; it is the industry-standard format for storing textual and graphical data. PDF offers a graphically rich appearance and structure that makes it ideal for presentation purposes. Additionally, PDF documents can be viewed in both image and textual mode, i.e., you can either select data as a formatted image or as regular text.

PDF is platform independent, and most operating systems offer free PDF viewing applications (e.g., Adobe Acrobat on Windows and Ghostscript on Linux).

PDF has been approved as an acceptable document storage format for regulated environments according to the FDA’s 21 CFR Part 11.

To save a *STATISTICA* Report as a PDF file, open the report, select the ***Home*** tab, and then select ***Save As PDF*** from the ***Save*** menu. The ***Output Options*** dialog will be displayed, where you can choose whether to output spreadsheets as ***Objects (as they are sized in the report window)*** or ***Full-sized Spreadsheets (on separate pages)***. If you always want to output spreadsheets in the same manner, select the ***Use the current setting and do not display this dialog again*** check box. Click the ***OK*** button to close the ***Output Options*** dialog and display the ***Save report as PDF*** dialog. Use the ***Save in*** field to select the appropriate location in which to save the document, enter a name in the ***File name*** field, and click the ***Save*** button. *STATISTICA* Reports, Spreadsheets, and Graphs can all be saved in PDF format.

Note that these are not simplified PDF files (representing compressed bitmaps of the respective document page images) but full-featured PDF files that support such operations as selective copying of text information.

HTML Reports

You may want to post a *STATISTICA* Report or Workbook on the Internet for others to review. With *STATISTICA*, you can save reports and workbooks in HTML (an acronym for HyperText Markup Language) format. HTML uses tags to identify elements of the document, such as text or graphics.

Open a *STATISTICA* Report or Workbook, and select ***Save As*** from the ***Save*** menu (located on the ***Home*** tab in the ***File*** group) to display the ***Save As*** dialog. From the ***Save as type*** drop-down list, select ***Web Page (*.html; *.htm)*** to save the file with an **.htm* extension.

Note that graphs in the report or workbook are saved as **.png* files in the same folder as the HTM file. You can save graphs as JPG files, instead. To do this, click ***Options*** (on the ***Home*** tab in the ***Tools*** group) to display the ***Options*** dialog. Select either ***Reports*** or ***Workbooks*** in the tree view, according to which document you intend to send to an *.htm* document, select the ***JPEG format*** option button in the ***Export HTML images as*** group box, and click ***OK***.

4. MICROSOFT WORD

With *STATISTICA*, you can also route output directly to Word via the Office Integration features. When Word is open within *STATISTICA,* Word toolbars and menus are also available through standard Active X Document interfaces technology. In *STATISTICA*, you can perform any formatting and editing that Word supports in its application.

When sending spreadsheet analytical results to Word, *STATISTICA* will take advantage of Word’s table editing facility, and convert the spreadsheet to a table. For multi-page spreadsheets, you can control where to break the rows and columns. These spreadsheets will be broken by columns such as will be allowed without exceeding the page width. All rows for a given set of columns will be rendered before the next set of spreadsheet columns is rendered in the Word document. This solution enables the presentation of spreadsheets in Word that are natively editable in Word, displays the entire contents of the spreadsheet, and prints and paginates correctly.

As with standard *STATISTICA* Reports (see page 151), Word documents can store and preserve the record of supplementary information (e.g., selected variables, long names, etc.).

To send output to a Word document, use the options in the ***Output Manager*** (accessible by selecting ***Output Manager*** from the Start button drop-down menu located in the upper-left corner of the ribbon bar; or by selecting the ***Home*** tab, clicking ***Options*** in the ***Tools*** group, and selecting ***Output Manager*** in the ***Options*** dialog tree view). In the ***Microsoft Word Output*** drop-down list, select either ***Multiple Word documents (one for each analysis/graph)***, ***Common Word document (one shared for all analyses/graphs)***, or ***[Select File]*** to browse to a preexisting Word document.

Although Word documents do not provide the navigational tree of a *STATISTICA* Workbook or Report, the advantages in sending output to Word documents are many. By sending results to a Word document, you have all the word processing features of Word available. For example, you can attach templates to create customized documents, add tables of content and indices, track changes, etc.

When inserting a large spreadsheet into a Word document, *STATISTICA* automatically detects how many variables can fit on each page and partitions the spreadsheet into several Word tables. If the spreadsheet uses case names, those names will be the first column in each table.

Additional benefits of sending results to a Word document include increased printing functionality (e.g., printing to files, manual duplex) and the ability to save results as Web pages.

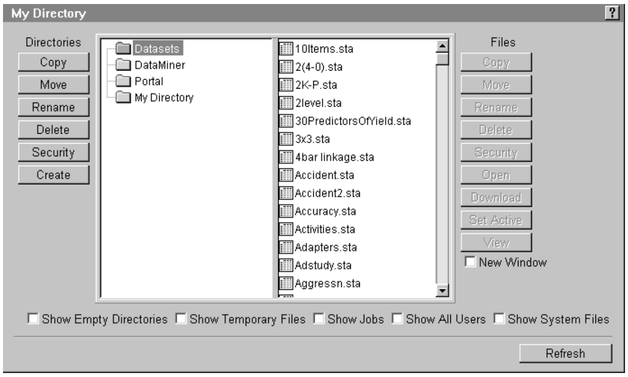
5. OUTPUT TO THE WEB

Knowledge Portal

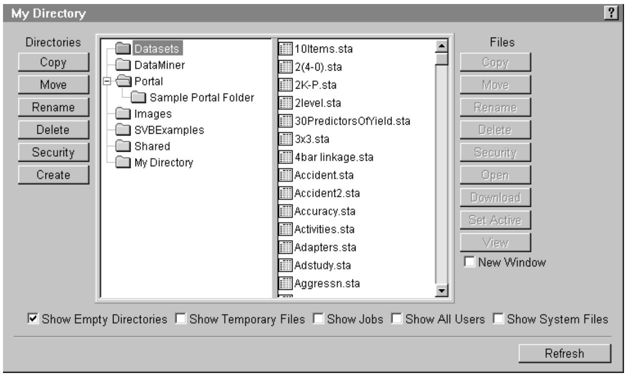
STATISTICA Enterprise Server Reports, or any *STATISTICA* Reports (see *HTML Reports*, page 154), can be distributed through the Knowledge Portal. The

Knowledge Portal enables you to publish *STATISTICA* documents (spreadsheets, graphs, reports, or workbooks) to the Internet. Users with limited Knowledge Portal permissions can then view those documents. You can control who can access these documents by setting permissions on the documents and directories using standard *STATISTICA Enterprise Server* repository tools.

To publish content in the Knowledge Portal, first create a directory in the *STATISTICA Enterprise Server* repository in the *Portal* folder: log on to the *STATISTICA Enterprise Server* as a user with Administrator rights, and from the ***File*** menu, select ***My Directory Operations*** to display the ***My Directory*** dialog; the content will look similar to the following illustration.



To create a folder in the *Portal* directory to contain your reports, select the *Portal* folder, and then click the ***Create*** button to display the ***Explorer User Prompt*** dialog. In the edit field, enter the new directory name of *Sample Portal Folder*, and click ***OK***. A dialog will be displayed confirming that the directory */Portal/Sample Portal Folder* was created. Click the ***Show My Directory*** button, and you will be returned to the ***My Directory*** dialog. Select the ***Show Empty Directories*** check box, and then click the ***Refresh*** button. Expand the *Portal* directory by clicking the + next to that folder, and the new *Sample Portal Folder* will be displayed.



☐ New Window

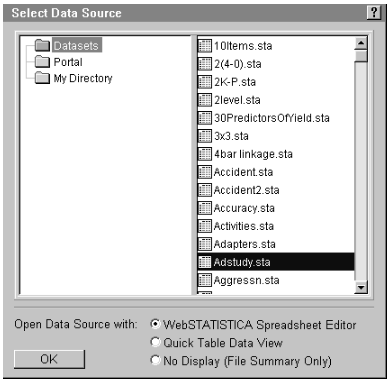
☑ Show Empty Directories ☐ Show Temporary Files ☐ Show Jobs ☐ Show All Users ☐ Show System Files

Note that you can control who can read and write to this folder by selecting the *Sample Portal Folder*, clicking the ***Security*** button, and using the options to set the user and group permissions for the folder appropriately.

Publishing Content from *STATISTICA Enterprise Server*

Now that the folder has been created, you can add analysis results to it for *Portal* users to view using either *STATISTICA Enterprise Server* or *STATISTICA*.

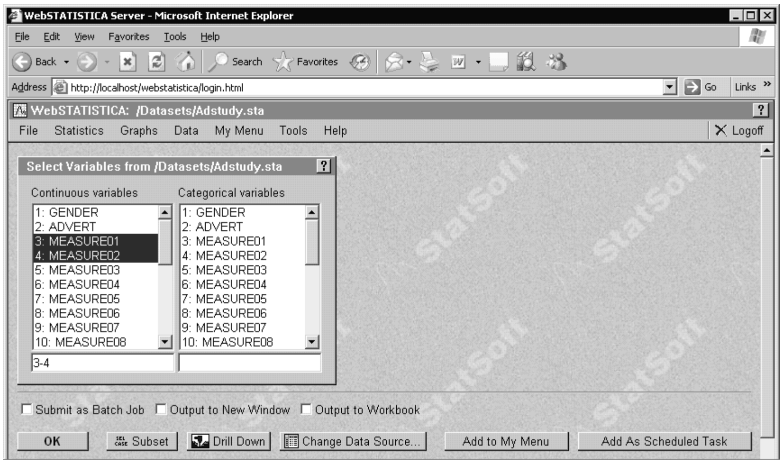
In *STATISTICA Enterprise Server*, start with a typical analysis. From the ***File*** menu, select ***Open Data Spreadsheet***. In the ***Select Data Source*** dialog, select the *Datasets* folder in the left pane, select the data file *Adstudy.sta* in the right pane, and click ***OK***.



Open Data Source with: ● WebSTATISTICA Spreadsheet Editor
○ Quick Table Data View
○ No Display (File Summary Only)

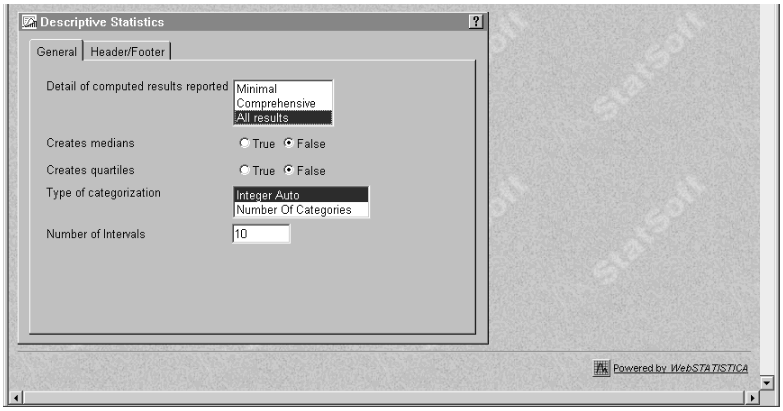
Close the resulting Spreadsheet Editor window (we won't need it in this example), leaving just the browser window displaying the active data source summary information for *Adstudy.sta.*

From the ***Statistics - Basic Statistics and Tables*** submenu, select ***Descriptive Statistics*** to display the variable selection dialog and the ***Descriptive Statistics*** specifications dialog. In the variable selection dialog, select *MEASURE01* and *MEASURE02* in the ***Continuous variables*** column.


WebSTATISTICA Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost/webstatistica/login.html
Go Links
WebSTATISTICA: /Datasets/Adstudy.sta
File Statistics Graphs Data My Menu Tools Help
Logoff
Select Variables from /Datasets/Adstudy.sta
Continuous variables
Categorical variables
1: GENDER
2: ADVERT
3: MEASURE01
4: MEASURE02
5: MEASURE03
6: MEASURE04
7: MEASURE05
8: MEASURE06
9: MEASURE07
10: MEASURE08
□ Submit as Batch Job □ Output to New Window □ Output to Workbook
OK
Subset
Drill Down
Change Data Source...
Add to My Menu
Add As Scheduled Task

Continuous variables: 3-4

In the ***Descriptive Statistics*** specifications dialog, select ***All results*** in the ***Detail of computed results reported*** field.



Detail of computed results reported: All results

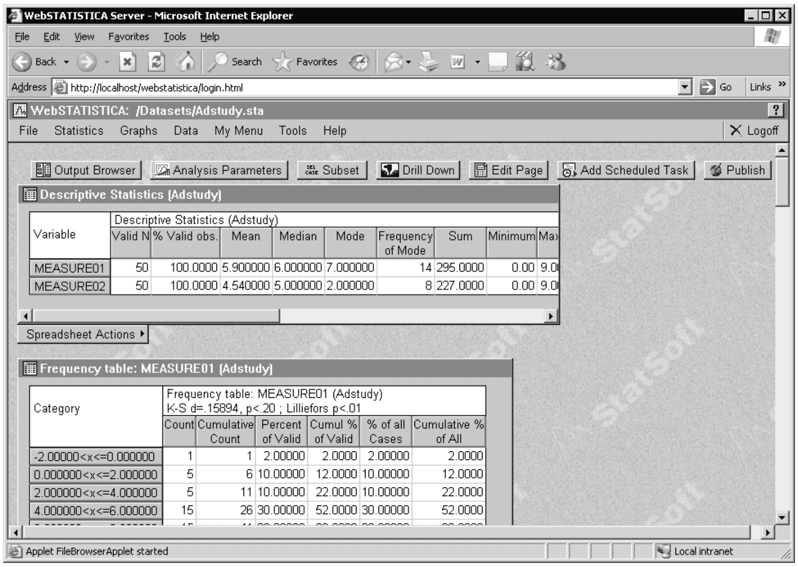
Creates medians ○ True ● False

Creates quartiles ○ True ● False

Type of categorization: Integer Auto

Number of Intervals: 10

Click ***OK*** to display the results for this analysis, consisting of several spreadsheets and graphs.


WebSTATISTICA Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost/webstatistica/login.html
Go Links
WebSTATISTICA: /Datasets/Adstudy.sta
File Statistics Graphs Data My Menu Tools Help
Logoff
Output Browser
Analysis Parameters
Subset
Drill Down
Edit Page
Add Scheduled Task
Publish
Spreadsheet Actions
Applet FileBrowserApplet started
Local intranet

Descriptive Statistics (Adstudy)

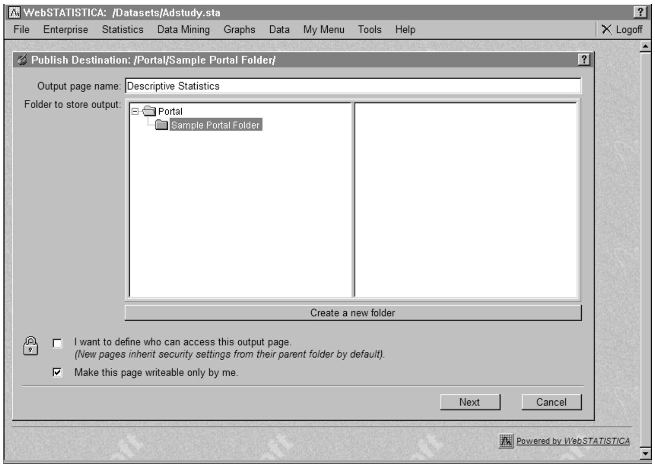
Variable	Valid N	% Valid obs.	Mean	Median	Mode	Frequency of Mode	Sum	Minimum	Max
MEASURE01	50	100.0000	5.900000	6.000000	7.000000	14	295.0000	0.00	9.0
MEASURE02	50	100.0000	4.540000	5.000000	2.000000	8	227.0000	0.00	9.0

Frequency table: MEASURE01 (Adstudy)

Frequency table: MEASURE01 (Adstudy)
K-S d=.15894, p<.20 ; Lilliefors p<.01

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
-2.00000<x<=0.000000	1	1	2.00000	2.0000	2.00000	2.0000
0.000000<x<=2.000000	5	6	10.00000	12.0000	10.00000	12.0000
2.000000<x<=4.000000	5	11	10.00000	22.0000	10.00000	22.0000
4.000000<x<=6.000000	15	26	30.00000	52.0000	30.00000	52.0000

Now, to publish this page so that other users can see it from the *Knowledge Portal*, click the ***Publish*** button in the upper-right portion of the window. The ***Publish Destination*** dialog will be displayed. Here you can select the *Sample Portal Folder* that you created. You also can control who can have access to this particular page by selecting the ***I want to define who can access this output page*** check box.


WebSTATISTICA: /Datasets/Adstudy.sta
File Enterprise Statistics Data Mining Graphs Data My Menu Tools Help
Logoff
Publish Destination: /Portal/Sample Portal Folder/
Portal
Sample Portal Folder
Create a new folder
Next
Cancel
Powered by WebSTATISTICA

Output page name: Descriptive Statistics

Folder to store output: /Portal/Sample Portal Folder

☐ I want to define who can access this output page.
(New pages inherit security settings from their parent folder by default).

☑ Make this page writeable only by me.

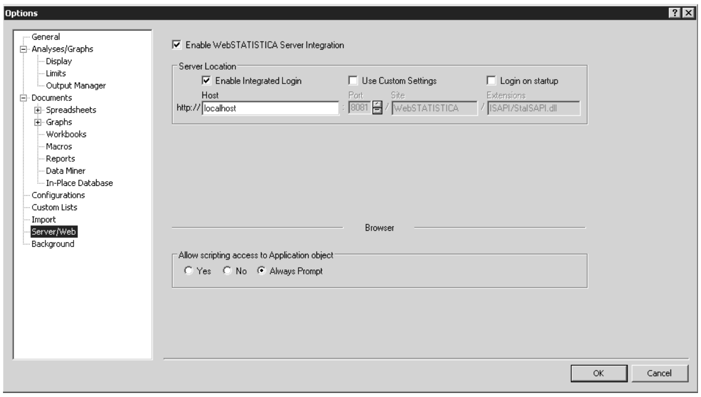
Click the ***Next*** button, and the page will be saved to the selected destination.

Now, when a *Knowledge Portal* user logs on, they will see the new *Sample Portal Folder* in their output browser, from which they can select the newly added *Descriptive Statistics* page.

Publishing Content from *STATISTICA* Desktop Applications

With the *STATISTICA Enterprise Server* integration feature of desktop *STATISTICA*, you can also publish *STATISTICA* documents (spreadsheets, graphs, reports, and workbooks) to the *Knowledge Portal* directly from within the *STATISTICA* application.

The first step is to enable *STATISTICA Enterprise Server* integration. Select the ***Home*** tab, and in the ***Tools*** group click ***Options*** to display the ***Options*** dialog. Select ***Server/Web*** in the tree view, and in the options pane, select the ***Enable STATISTICA Enterprise Server Integration*** check box. Then, specify the URL of the *STATISTICA Enterprise Server* and any optional custom configuration settings that may have been defined by your system administrator when installing *STATISTICA Enterprise Server*. In the following illustration, *STATISTICA Enterprise Server* has been installed on *serverx23*; the information in your dialog will be different depending on where *STATISTICA Enterprise Server* is installed on your network.



☑ Enable WebSTATISTICA Server Integration

Server Location

☑ Enable Integrated Login ☐ Use Custom Settings ☐ Login on startup

Host: http:// localhost : Port 8081 / Site WebSTATISTICA / Extensions ISAPI/StatSAPI.dll

Browser

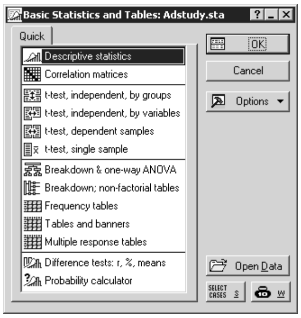
Allow scripting access to Application object

○ Yes ○ No ● Always Prompt

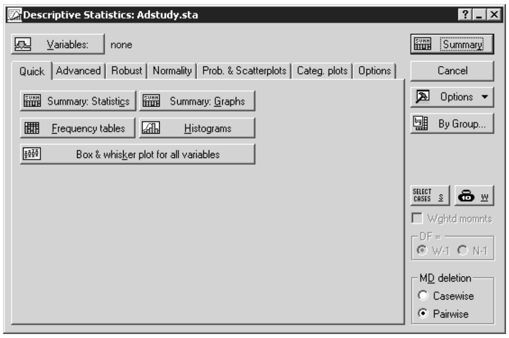
After you click the ***OK*** button in the ***Options*** dialog, note that there is a now a ***Server*** tab displayed in *STATISTICA* next to the ***Home*** tab. The only command on the ***Server*** tab that is available initially is ***Log In***; select that command. If you have enabled integrated log in (and your Windows account is enabled on the *STATISTICA Enterprise Server*), you will be logged in automatically. Otherwise, you will be prompted for a *STATISTICA Enterprise Server* user name and password. Once you have logged in, the other commands are available on the ***Server*** tab.

Now, we will create an analysis and upload the results to the Knowledge Portal. Open the *Adstudy.sta* data file: select the ***Home*** tab, click the ***Open*** arrow, and select ***Open Examples*** from the drop-down menu; in the ***Open a STATISTICA Data File*** dialog, double-click on the *Datasets* folder, and then double-click on the *Adstudy.sta* file to open that spreadsheet for use in *STATISTICA*.

Next, select the ***Statistics*** tab, and in the ***Base*** group, click ***Basic Statistics*** to display the ***Basic Statistics and Tables*** Startup Panel. Select ***Descriptive statistics***.

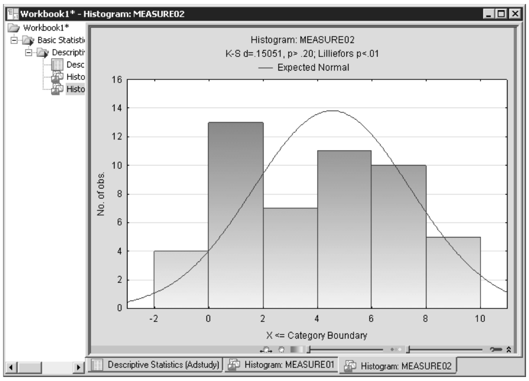

Basic Statistics and Tables: Adstudy.sta
Quick
Descriptive statistics
Correlation matrices
t-test, independent, by groups
t-test, independent, by variables
t-test, dependent samples
t-test, single sample
Breakdown & one-way ANOVA
Breakdown; non-factorial tables
Frequency tables
Tables and banners
Multiple response tables
Difference tests: r, %, means
Probability calculator
OK
Cancel
Options
Open Data
SELECT CASES
W

Click ***OK*** to display the ***Descriptive Statistics*** dialog.

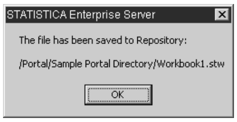


To ensure that all the output from this analysis will be sent to a workbook, click the ***Options*** button on the right side of the dialog, and from the drop-down list, select ***Output***. In the ***Analysis/Graph Output Manager***, verify that the ***Workbook*** option button is selected in the ***Place all results (Spreadsheets, Graphs) in*** group box. Then click ***OK*** to return to the ***Descriptive Statistics*** dialog.

Click the ***Variables*** button to display the variable selection dialog, select *MEASURE01* and *MEASURE02*, and click ***OK*** to return to the ***Descriptive Statistics*** dialog. On the ***Quick*** tab, click the ***Summary: Statistics*** button to send those results to the workbook. The ***Descriptive Statistics*** dialog will be minimized so you can see the results; restore it by clicking the ***Descriptive Statistics*** button on the Analysis Bar in the lower-left of the screen. Now click the ***Histograms*** button to generate histograms for each selected variable. The analysis dialog is minimized again, and the workbook should look as follows.



This is the document we want to publish to the Knowledge Portal. On the ***Server*** tab in the ***File*** group, click ***Save As***. The ***STATISTICA Enterprise Repository*** dialog will be displayed, containing a list of folders you can reference in the *STATISTICA Enterprise Server*. Open the *Portal* folder, select *Sample Portal Folder*, and click the ***OK*** button. This will upload the workbook to that Knowledge Portal directory.

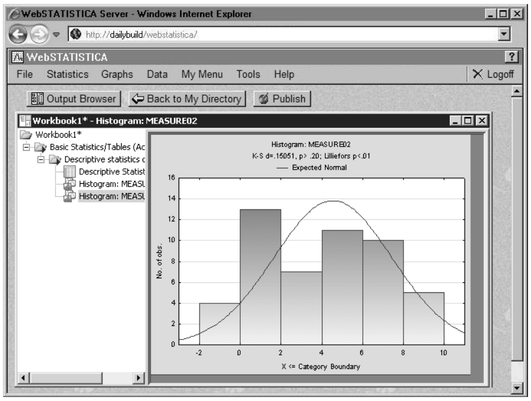

STATISTICA Enterprise Server

The file has been saved to Repository:

/Portal/Sample Portal Directory/Workbook1.stw

You can review the document from within *STATISTICA* by opening a browser window inside of the *STATISTICA* workspace. On the ***Server*** tab in the ***Tools*** group, select ***Open in Browser***, and a new browser window will be opened, allowing you to log on to the *STATISTICA Enterprise Server*.

From the *STATISTICA Enterprise Server* ***File*** menu, choose ***My Directory Operations***; in *My Directory*, you can navigate to the *Sample Portal Directory*, and see the *Workbook1.stw* file that was uploaded. Select this file and click the ***View*** button, and the workbook will be opened within the browser.



6. SHAREPOINT OR *STATISTICA DOCUMENT MANAGEMENT SYSTEM (SDMS)*

With *STATISTICA*, you can also route output to either Microsoft SharePoint or to the *STATISTICA Document Management System (SDMS)*.

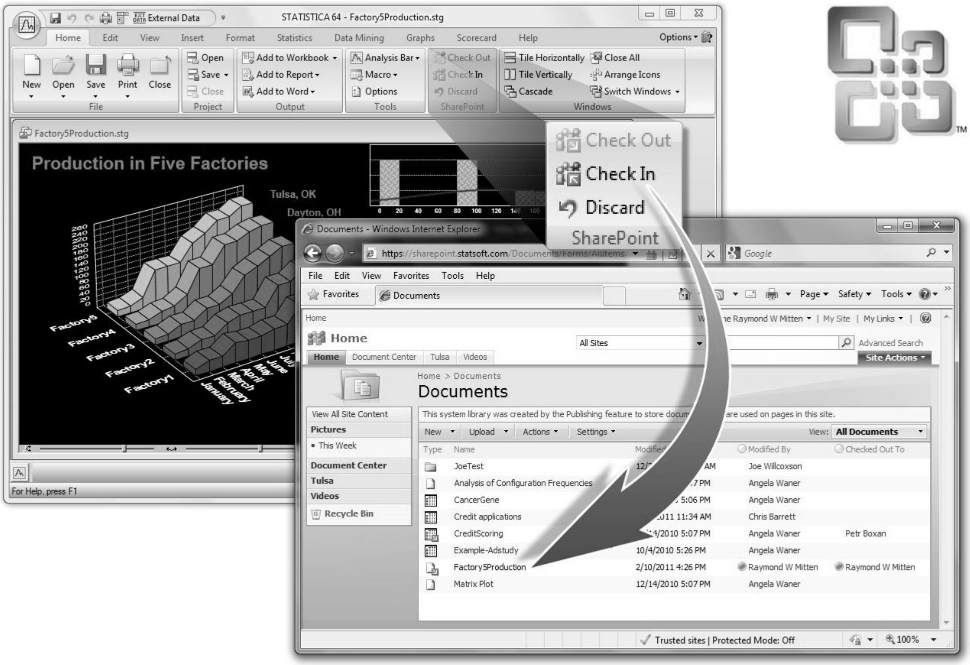
SharePoint

With *STATISTICA* SharePoint integration, you can open, check out, check in, and upload new *STATISTICA* files to SharePoint.

To open a document in *STATISTICA* that is located in SharePoint, select the ***Home*** tab. Click the ***Open*** arrow, and select ***Open Document***. In the ***Open*** dialog, in the ***Look in*** drop-down list, select the Web Folder to the SharePoint server location (see page 165), and then navigate to the document you want to open. You will need to log on to SharePoint.

To save a *STATISTICA* document (spreadsheet, workbook, macro, etc.) to SharePoint, select the ***Home*** tab. Click the ***Save*** arrow, and select ***Save As***. In the ***Save As*** dialog, in the ***Save in*** drop-down list, select the Web Folder to the SharePoint server location, and then navigate to the location in which you want to save the document. You will need to log on to SharePoint.

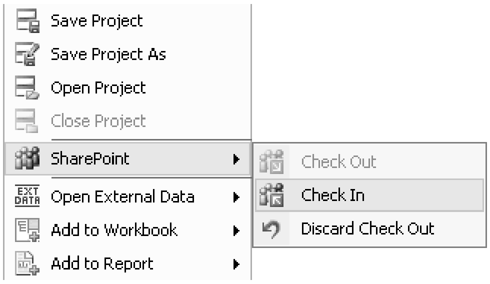
The SharePoint options ***Check Out***, ***Check In***, and ***Discard*** are located on the ***Home*** tab in the ***SharePoint*** group.



This system library was created by the Publishing feature to store docum are used on pages in this site.

Type	Name	Modified	Modified By	Checked Out To
	JoeTest	12/ AM	Joe Willcoxson	
	Analysis of Configuration Frequencies	7 PM	Angela Waner	
	CancerGene	5:06 PM	Angela Waner	
	Credit applications	011 11:34 AM	Chris Barrett	
	CreditScoring	4/2010 5:07 PM	Angela Waner	Petr Boxan
	Example-Adstudy	10/4/2010 5:26 PM	Angela Waner	
	Factory5Production	2/10/2011 4:26 PM	Raymond W Mitten	Raymond W Mitten
	Matrix Plot	12/14/2010 5:07 PM	Angela Waner	

These options can also be accessed by clicking the Start button located in the upper-left corner of the ribbon bar. These options become available after you have opened a document from SharePoint.



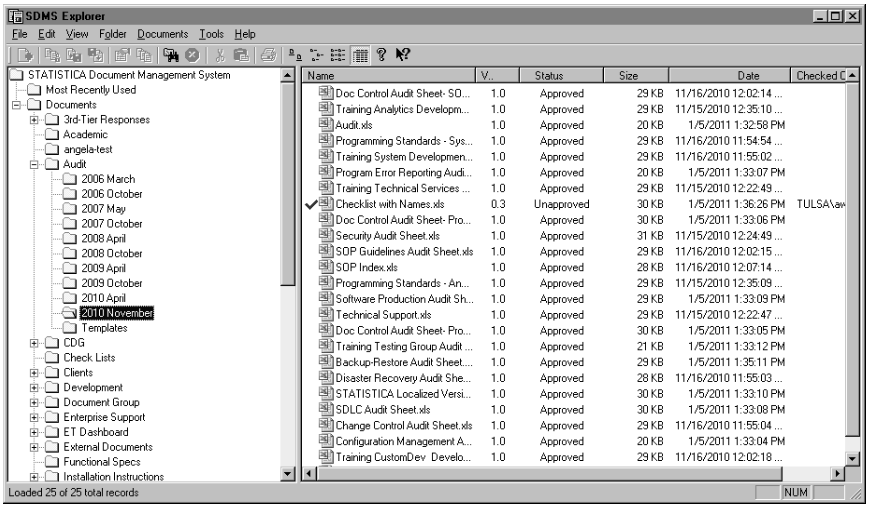
Before using these options, you must first create a Web Folder to the SharePoint server location. To do this, click the Start button in the lower-left corner of the Windows taskbar, and click *Computer*. Right-click in any open area in the right pane of the *Computer* dialog, and from the shortcut menu, select *Add a network location* to display the *Add Network Location* dialog. Click the *Next* button.

Double-click *Choose a custom network location*. In the *Internet or network address* field, enter the Web address of your SharePoint location: https://sharepoint..., or click the *Browse* button to browse to and select the location. Click *Next*.

Log on to SharePoint, and click *OK*. Enter a name for the Web Folder in the *Type a name for this network location* field, and click *Next*. You will see *Completing the Add Network Location Wizard*; select the *Open this network location when I click Finish* check box, and then click *Finish*. A Network Location Web Folder has been created in the *Network Location* section of *Computer* with the label you chose.

STATISTICA Document Management System (SDMS)

STATISTICA Document Management System (SDMS) is a complete database solution package for managing documents. *SDMS* enables you to quickly, efficiently, and securely save documents of any type to a secure repository database, and then manage them [e.g., find them, access them, search for content, review, organize, edit (with trail logging and versioning), approve, etc.].



	Name	V..	Status	Size	Date	Checked C
	Doc Control Audit Sheet- SO...	1.0	Approved	29 KB	11/16/2010 12:02:14 ...	
	Training Analytics Developm...	1.0	Approved	29 KB	11/15/2010 12:35:10 ...	
	Audit.xls	1.0	Approved	20 KB	1/5/2011 1:32:58 PM	
	Programming Standards - Sys...	1.0	Approved	29 KB	11/16/2010 11:54:54 ...	
	Training System Developmen...	1.0	Approved	29 KB	11/16/2010 11:55:02 ...	
	Program Error Reporting Audi...	1.0	Approved	20 KB	1/5/2011 1:33:07 PM	
	Training Technical Services ...	1.0	Approved	29 KB	11/15/2010 12:22:49 ...	
✓	Checklist with Names.xls	0.3	Unapproved	30 KB	1/5/2011 1:36:26 PM	TULSA\aw
	Doc Control Audit Sheet- Pro...	1.0	Approved	30 KB	1/5/2011 1:33:06 PM	
	Security Audit Sheet.xls	1.0	Approved	31 KB	11/15/2010 12:24:49 ...	
	SOP Guidelines Audit Sheet.xls	1.0	Approved	29 KB	11/16/2010 12:02:15 ...	
	SOP Index.xls	1.0	Approved	28 KB	11/16/2010 12:07:14 ...	
	Programming Standards - An...	1.0	Approved	29 KB	11/15/2010 12:35:09 ...	
	Software Production Audit Sh...	1.0	Approved	29 KB	1/5/2011 1:33:09 PM	
	Technical Support.xls	1.0	Approved	29 KB	11/15/2010 12:22:47 ...	
	Doc Control Audit Sheet- Pro...	1.0	Approved	30 KB	1/5/2011 1:33:05 PM	
	Training Testing Group Audit ...	1.0	Approved	21 KB	1/5/2011 1:33:12 PM	
	Backup-Restore Audit Sheet....	1.0	Approved	29 KB	1/5/2011 1:35:11 PM	
	Disaster Recovery Audit She...	1.0	Approved	28 KB	11/16/2010 11:55:03 ...	
	STATISTICA Localized Versi...	1.0	Approved	30 KB	1/5/2011 1:33:10 PM	
	SDLC Audit Sheet.xls	1.0	Approved	30 KB	1/5/2011 1:33:08 PM	
	Change Control Audit Sheet.xls	1.0	Approved	29 KB	11/16/2010 11:55:04 ...	
	Configuration Management A...	1.0	Approved	20 KB	1/5/2011 1:33:04 PM	
	Training CustomDev Develo...	1.0	Approved	29 KB	11/16/2010 12:02:18 ...	

Loaded 25 of 25 total records

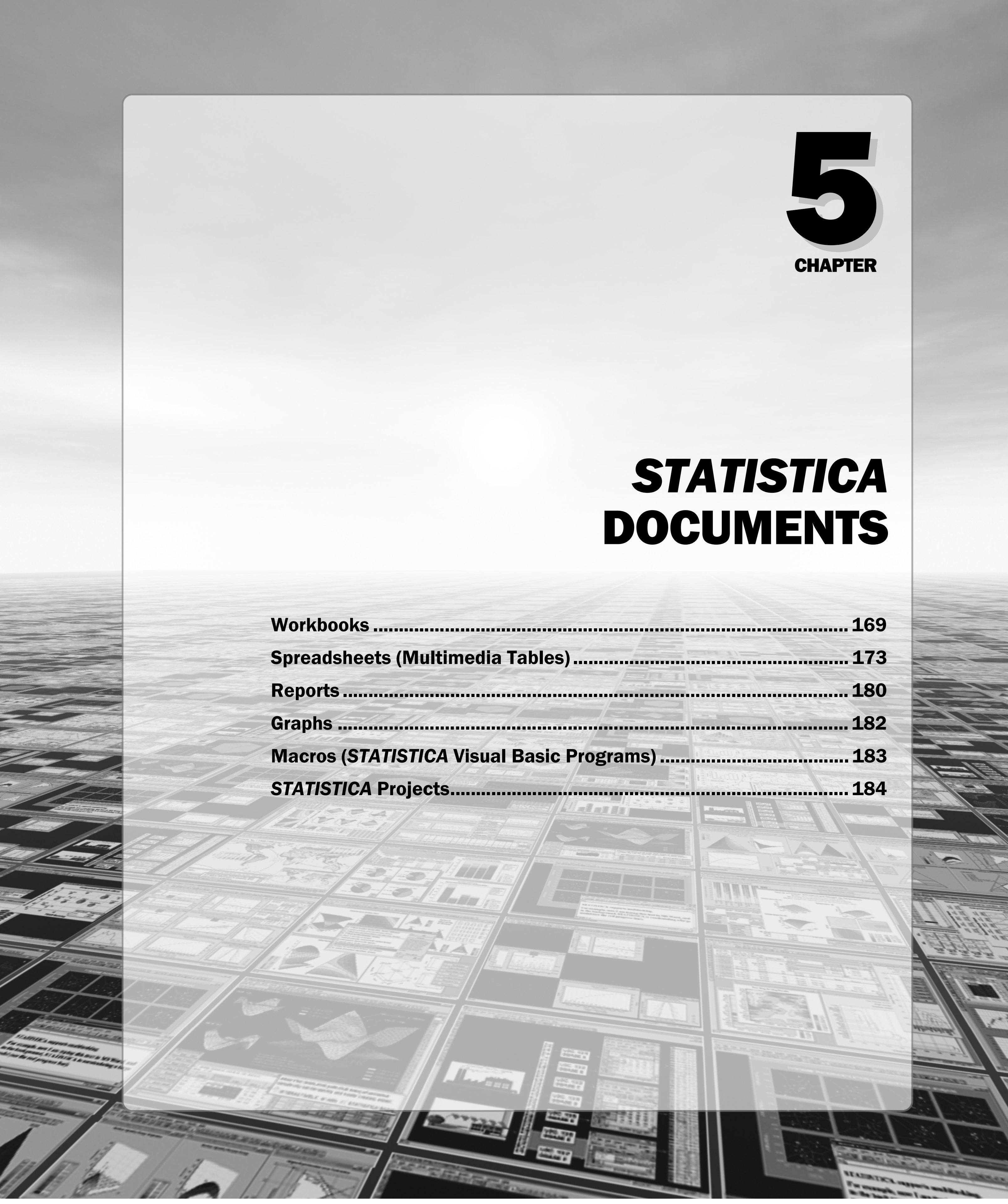
The intuitive user interface of *SDMS* makes it easy to perform all document management operations from any computer on your network or even via the Internet.

In the *STATISTICA Document Management System*, everything is documented and traceable. For example, documents are never deleted. When a document is edited, a new version of that document is created, properly authenticated, and annotated with electronic signatures. Authorized users can be required to explicitly check out the documents from the repository and check the new versions into the repository with notes and documentation regarding the nature and purpose of the edits.

SDMS is specifically designed to ensure compliance with FDA 21 CFR Part 11 regulations and Sarbanes-Oxley legislation, as well as ISO 9000, 9001, 14001 documentation requirements.

STATISTICA Document Management System seamlessly integrates with all *STATISTICA* products, from desktop and network versions, to enterprise-wide installations such as *STATISTICA Enterprise Server*-based worldwide installations or *STATISTICA Enterprise*/QC (for process analysis and quality control/improvement). *SDMS* can also be used as a stand-alone system.

SDMS is highly configurable, and its functionality is compatible with other applications, so the system can be customized to accommodate your specific tasks and can be integrated seamlessly into existing systems for data and document management.



5

CHAPTER

STATISTICA DOCUMENTS

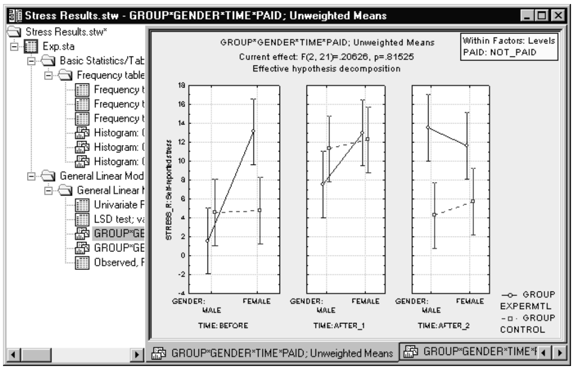
5

CHAPTER

STATISTICA DOCUMENTS

WORKBOOKS

Workbooks (introduced briefly on page 148) are the default way of managing output. They store each output document (e.g., a *STATISTICA* Spreadsheet or Graph, as well as a Word or Excel document) as a tab.



Technically speaking, *STATISTICA* Workbooks are optimized ActiveX (see page 238) containers that can efficiently handle large numbers of documents. The documents can be organized into hierarchies of folders or document nodes (by default, one is created for each new analysis) using a tree view, in which individual documents, folders, or entire branches of the tree can be flexibly managed.

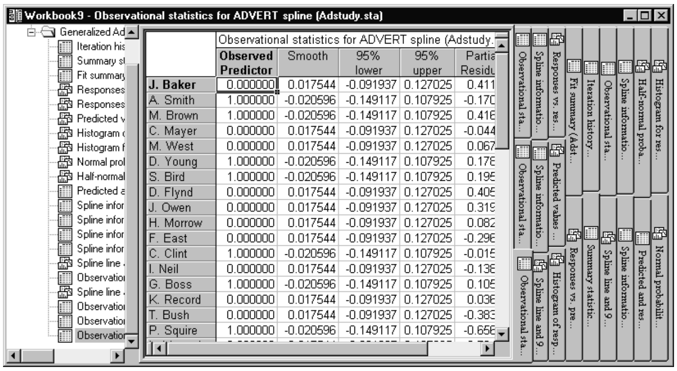
For example, selections of documents can be extracted (e.g., drag-copied or drag-moved) to the report window or to the application workspace (i.e., the *STATISTICA* application "background" where they are displayed in stand-alone windows).

Entire branches can be placed into other workbooks in a variety of ways in order to build a specific folder organization, etc.

Each workbook contains two panels: an Explorer-style navigation tree on the left and a document viewer on the right.

The navigation tree (workbook tree) can be split into various nodes that are used to organize files in logical groupings (e.g., all analysis outputs or all macros created for a project).

Tabs at the bottom of the document viewer (workbook viewer) are used to easily navigate the children of the currently selected node. You can move the tabs to the top, right, or left of the workbook viewer by right-clicking on one of the tabs and selecting a different location from the shortcut menu. One advantage of the side placement of tabs is that multiple rows (rather than one long row) are provided (as shown below). This makes it easy to select the desired tab.



Observational statistics for ADVERT spline (Adstudy.

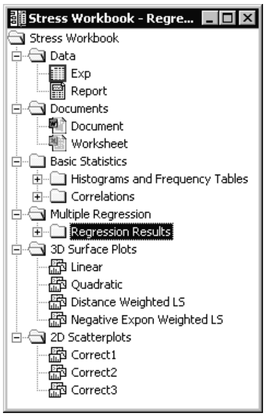
	Observed Predictor	Smooth	95% lower	95% upper	Partia Residu
J. Baker	0.000000	0.017544	-0.091937	0.127025	0.411
A. Smith	1.000000	-0.020596	-0.149117	0.107925	-0.170
M. Brown	1.000000	-0.020596	-0.149117	0.107925	0.416
C. Mayer	0.000000	0.017544	-0.091937	0.127025	-0.044
M. West	0.000000	0.017544	-0.091937	0.127025	0.067
D. Young	1.000000	-0.020596	-0.149117	0.107925	0.178
S. Bird	1.000000	-0.020596	-0.149117	0.107925	0.195
D. Flynd	0.000000	0.017544	-0.091937	0.127025	0.405
J. Owen	0.000000	0.017544	-0.091937	0.127025	0.319
H. Morrow	0.000000	0.017544	-0.091937	0.127025	0.082
F. East	0.000000	0.017544	-0.091937	0.127025	-0.296
C. Clint	1.000000	-0.020596	-0.149117	0.107925	-0.015
I. Neil	0.000000	0.017544	-0.091937	0.127025	-0.138
G. Boss	1.000000	-0.020596	-0.149117	0.107925	0.105
K. Record	0.000000	0.017544	-0.091937	0.127025	0.036
T. Bush	0.000000	0.017544	-0.091937	0.127025	-0.383
P. Squire	1.000000	-0.020596	-0.149117	0.107925	-0.656

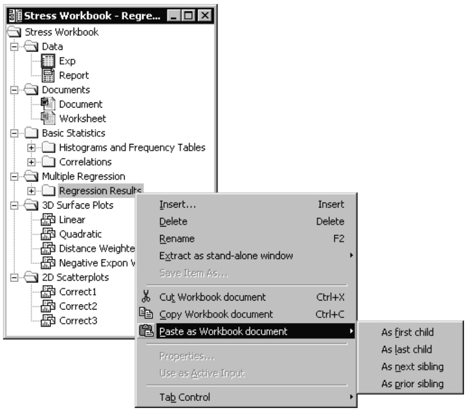
Displaying tabs can also be suppressed to save space. Unlike many Explorer-style navigation and organization applications that only allow folders to have children, the *STATISTICA* Workbook allows any item in the tree to have children. For example, you can add a spreadsheet to your workbook, and then add all the graphs produced using the data in the spreadsheet as children to the spreadsheet. A variety of drag-and-drop features and Clipboard procedures are available to aid you in organizing the workbook tree.

The workbook can hold all native *STATISTICA* documents including spreadsheets, graphs, reports, and macros. It can contain other types of ActiveX documents as well, including Excel spreadsheets, Word documents, and others. If you want to edit these documents, you can do so using the workbook viewer pane. To edit a Word document, double-click on the object in the workbook tree. The Word document opens in the viewer, and the workbook menu bar merges with the Word menu bar giving you access to all of the editing features you need. Workbooks can also be used to store all output from a particular analysis.

Navigating the Workbook Tree

The workbook tree displays the organization of files and folders in the workbook, displayed in an Explorer-style format. Items with plus signs next to them indicate folders or files that have children associated with them. To expand the tree for a particular folder or file, click the plus sign next to it. The workbook can support an unlimited number of levels, and individual items from the tree view or entire branches can be flexibly (interactively) managed (e.g., dragging to copy or move between workbooks or reports, etc., or via the shortcut menu, as shown below in the second image).





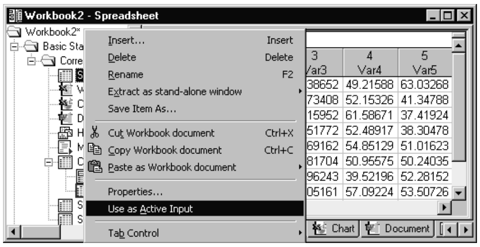
To select a workbook item for review or editing, simply locate the file in the workbook tree and click on its associated icon. The document will be displayed in the workbook viewer pane. Note that you can also navigate through the children of the currently selected node using the navigation tabs available (by default) at the bottom of the workbook viewer. You can easily move these navigation tabs to the top, right, or left of the workbook viewer by right-clicking on one of the tabs and selecting a different location from the shortcut menu or selecting the appropriate command from the ***Workbook*** tab, ***Tools*** group, ***Tab Control*** menu.

Note that tabs at the top and bottom of the viewer scroll sideways, while multiple rows of tabs are used when tabs are placed to the left or right of the viewer.

Items in the tree are identified by the icon next to them. The folder icon represents a folder that can contain a variety of documents and subfolders. The folder icon with a red arrow on it indicates that the script that produced the results in that folder has been attached to the folder. This enables *STATISTICA* to rerun or resume the analysis (for more details, see Chapter 8 – *STATISTICA Visual Basic*). The spreadsheet, report, macro, and graph icons represent *STATISTICA* Spreadsheet, Report, Macro, and Graph documents, respectively. The *Data Miner* icon represents a *Data Miner* workspace.

All non-*STATISTICA* documents are represented by their respective document icons. For example, Word documents are represented by the Word icon, and Excel spreadsheet files are represented by the Excel spreadsheet icon.

Commands for inserting, extracting, renaming, and removing items from the workbook tree are available from the workbook tree shortcut menu (accessed by right-clicking anywhere in the tree).



3 Var3	4 Var4	5 Var5
38652	49.21588	63.03268
73408	52.15326	41.34788
15952	61.58671	37.41924
51772	52.48917	38.30478
69162	54.85129	51.01623
81704	50.95575	50.24035
96243	39.52196	52.28152
05161	57.09224	53.50726

These commands are also accessible on the ***Workbook*** tab.

The workbook tree can be organized and modified using drag-and-drop features (as well as Clipboard procedures). Use keys on your keyboard to specify whether an item is to be moved or copied, and whether an item is to be inserted as a child (i.e., one level below) or as a sibling (i.e., on the same level).

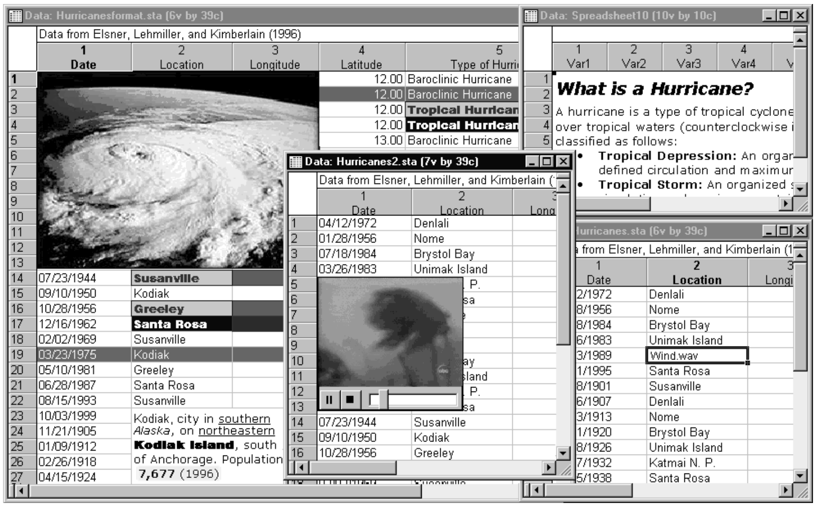
The following table illustrates four drag-and-drop options:

Action	Key Press	Cursor	Effect
Move Child	(none)		Move the first selected item **one level below** the second selected item.
Move Sibling	**SHIFT**		Move the first selected item **directly below and on the same level as** the second selected item.
Copy Child	**CTRL**		Copy the first selected item **one level below** the second selected item.
Copy Sibling	**SHIFT + CTRL**		Copy the first selected item **directly below and on the same level as** the second selected item.

First, select the item(s) that you want to move or copy. Drag the selection to its new location and drop it. To select a single item, click on the item (e.g., spreadsheet, graph, or report). To select a parent node and all of its children, click on the folder. Note that horizontal and/or vertical scrolling within the workbook tree can be utilized during a drag-and-drop operation.

SPREADSHEETS (MULTIMEDIA TABLES)

STATISTICA Spreadsheets are based on StatSoft's proprietary multimedia table technology and are used to manage both input data and the numeric or text (and, optionally, any other type of) output. The basic form of the spreadsheet is a simple two-dimensional table that can handle a practically unlimited number of cases (rows) and variables (columns), and each cell can contain a virtually unlimited number of characters. Sound, video, graphs, animations, reports with embedded objects, or any ActiveX compatible documents can also be attached.



Data: Hurricanesformat.sta (6v by 39c)

Data from Elsner, Lehmiller, and Kimberlain (1996)

	1 Date	2 Location	3 Longitude	4 Latitude	5 Type of Hurri
1				12.00	Baroclinic Hurricane
2				12.00	Baroclinic Hurricane
3				12.00	**Tropical Hurrican**
4				12.00	**Tropical Hurrican**
5				13.00	Baroclinic Hurricane
6					
7					
8					
9					
10					
11					
12					
13					
14	07/23/1944	**Susanville**			
15	09/10/1950	Kodiak			
16	10/28/1956	**Greeley**			
17	12/16/1962	**Santa Rosa**			
18	02/02/1969	Susanville			
19	03/23/1975	Kodiak			
20	05/10/1981	Greeley			
21	06/28/1987	Santa Rosa			
22	08/15/1993	Susanville			
23	10/03/1999	Kodiak, city in southern			
24	11/21/1905	*Alaska*, on northeastern			
25	01/09/1912	**Kodiak Island**, south			
26	02/26/1918	of Anchorage. Population			
27	04/15/1924	**7,677** (1996)			

Data: Hurricanes2.sta (7v by 39c)

Data from Elsner, Lehmiller, and Kimberlain (1

	1 Date	2 Location	3 Long
1	04/12/1972	Denlali	
2	01/28/1956	Nome	
3	07/18/1984	Brystol Bay	
4	03/26/1983	Unimak Island	
5		P.	
6		sa	
7		e	
8			
9			
10		ay	
11		sland	
12		P.	
13		sa	
14	07/23/1944	Susanville	
15	09/10/1950	Kodiak	
16	10/28/1956	Greeley	

Data: Spreadsheet10 (10v by 10c)

1 Var1	2 Var2	3 Var3	4 Var4	V

What is a Hurricane?

A hurricane is a type of tropical cyclone over tropical waters (counterclockwise i classified as follows:

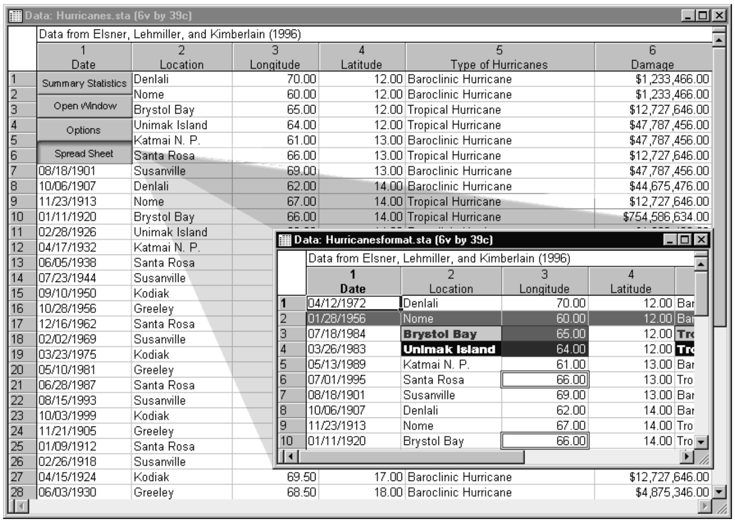
- **Tropical Depression:** An orgar defined circulation and maximur
- **Tropical Storm:** An organized s

urricanes.sta (6v by 39c)

a from Elsner, Lehmiller, and Kimberlain (1

1 Date	2 Location	3 Longi
2/1972	Denlali	
8/1956	Nome	
8/1984	Brystol Bay	
6/1983	Unimak Island	
3/1989	Wind.wav	
1/1995	Santa Rosa	
8/1901	Susanville	
6/1907	Denlali	
3/1913	Nome	
1/1920	Brystol Bay	
8/1926	Unimak Island	
7/1932	Katmai N. P.	
5/1938	Santa Rosa	

Because *STATISTICA* Spreadsheets can also contain macros and any user-defined user interface, these multimedia tables can be used as a framework for custom applications (e.g., with a list box of options or a series of buttons placed in the upper-left corner), self-running presentations, animations, simulations, etc.



Data: Hurricanes.sta (6v by 39c)

Data from Elsner, Lehmiller, and Kimberlain (1996)

	1 Date	2 Location	3 Longitude	4 Latitude	5 Type of Hurricanes	6 Damage
1	Summary Statistics	Denlali	70.00	12.00	Baroclinic Hurricane	$1,233,466.00
2	Open Window	Nome	60.00	12.00	Baroclinic Hurricane	$1,233,466.00
3	Options	Brystol Bay	65.00	12.00	Tropical Hurricane	$12,727,646.00
4		Unimak Island	64.00	12.00	Tropical Hurricane	$47,787,456.00
5	Spread Sheet	Katmai N. P.	61.00	13.00	Baroclinic Hurricane	$47,787,456.00
6		Santa Rosa	66.00	13.00	Tropical Hurricane	$12,727,646.00
7	08/18/1901	Susanville	69.00	13.00	Baroclinic Hurricane	$47,787,456.00
8	10/06/1907	Denlali	62.00	14.00	Baroclinic Hurricane	$44,675,476.00
9	11/23/1913	Nome	67.00	14.00	Tropical Hurricane	$12,727,646.00
10	01/11/1920	Brystol Bay	66.00	14.00	Tropical Hurricane	$754,586,634.00
11	02/28/1926	Unimak Island				
12	04/17/1932	Katmai N. P.				
13	06/05/1938	Santa Rosa				
14	07/23/1944	Susanville				
15	09/10/1950	Kodiak				
16	10/28/1956	Greeley				
17	12/16/1962	Santa Rosa				
18	02/02/1969	Susanville				
19	03/23/1975	Kodiak				
20	05/10/1981	Greeley				
21	06/28/1987	Santa Rosa				
22	08/15/1993	Susanville				
23	10/03/1999	Kodiak				
24	11/21/1905	Greeley				
25	01/09/1912	Santa Rosa				
26	02/26/1918	Susanville				
27	04/15/1924	Kodiak	69.50	17.00	Baroclinic Hurricane	$12,727,646.00
28	06/03/1930	Greeley	68.50	18.00	Baroclinic Hurricane	$4,875,346.00

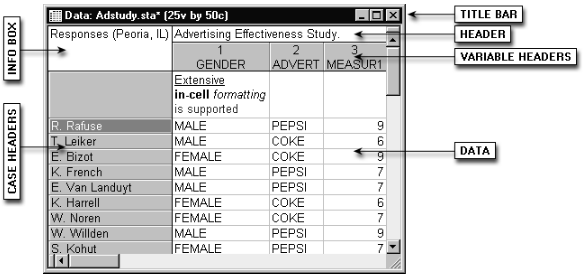
Data: Hurricanesformat.sta (6v by 39c)

Data from Elsner, Lehmiller, and Kimberlain (1996)

	1 Date	2 Location	3 Longitude	4 Latitude	
1	04/12/1972	Denlali	70.00	12.00	Bar
2	01/28/1956	Nome	60.00	12.00	Bar
3	07/18/1984	**Brystol Bay**	65.00	12.00	**Tro**
4	03/26/1983	**Unimak Island**	64.00	12.00	**Tro**
5	05/13/1989	Katmai N. P.	61.00	13.00	Bar
6	07/01/1995	Santa Rosa	66.00	13.00	Tro
7	08/18/1901	Susanville	69.00	13.00	Bar
8	10/06/1907	Denlali	62.00	14.00	Bar
9	11/23/1913	Nome	67.00	14.00	Tro
10	01/11/1920	Brystol Bay	66.00	14.00	Tro

Data file layout in spreadsheets. *STATISTICA* data are organized into cases and variables. If you are unfamiliar with this notation, you can think of cases as the equivalent of records in a database management program (or rows of a spreadsheet), and variables as the equivalent of fields (or columns of a spreadsheet). Each case consists of a set of values of variables, and the first column in the file can (optionally) contain names of cases.

The spreadsheet window comprises several basic components.



Data: Adstudy.sta* (25v by 50c)

Responses (Peoria, IL)	Advertising Effectiveness Study.		
	1 GENDER	2 ADVERT	3 MEASUR1
	<u>Extensive</u> **in-cell** *formatting* is supported		
R. Rafuse	MALE	PEPSI	9
T. Leiker	MALE	COKE	6
E. Bizot	FEMALE	COKE	9
K. French	MALE	PEPSI	7
E. Van Landuyt	MALE	PEPSI	7
K. Harrell	FEMALE	COKE	6
W. Noren	FEMALE	COKE	7
W. Willden	MALE	PEPSI	9
S. Kohut	FEMALE	PEPSI	7

Title bar. The title bar displays the name of the spreadsheet followed by the spreadsheet extension (*.sta*). If the spreadsheet is an input spreadsheet, the title bar also displays the number of variables by number of cases (e.g., *25v by 50c*). In the image shown above, the title bar contains the text *Data: Adstudy.sta (25v by 50c)*.

Info box. You can select the entire spreadsheet by clicking once in the lower-right corner (the mouse pointer will be the default arrow) of the info box, which is located in the upper-left corner of the spreadsheet window. To select the info box only (for formatting), click once in the upper-left corner of the info box (the mouse pointer will be an outlined plus sign ✜). Double-click in the info box to enter or edit the text in the info box (e.g., additional details about the spreadsheet). In the image shown above, the info box contains the text *Responses (Peoria, IL)*.

Header. The header is located immediately above the variable headers at the top of the window. Double-click the header to enter or edit text information. To select the header only (for formatting), click once in the upper-left corner (the mouse pointer will be an outlined plus sign ✜). Press CTRL+ENTER or ALT+ENTER to enter a new line (note that you need to extend the height of the field to see new lines that you are adding). In the image shown above, the header contains the text *Advertising Effectiveness Study*.

Case headers. These cells, located at the far left of the window, contain header information for each case. Double-click on any case header cell to enter or edit text information. To select the case header only (for formatting), click once on the left side of the case header (the mouse pointer will be an outlined plus sign ✜). To select the case row (for editing), click once on the middle or right side of the case header (the mouse pointer will be an outlined plus sign with an arrow ✙⇨). To select a block of case headers, (without selecting their respective rows), click on the left side of a case header and drag the mouse pointer to include all desired case headers. To autofit the case headers, double-click on the far-right side of any case header (the mouse pointer will be a cross with a double-headed arrow ↔). In the previous image, the case header cells contain the first initials and last names of the respondents in the study. Note that case headers are optional and you can choose not to display them (select the ***View*** tab, in the ***Display*** group click ***Display Options***, and toggle off the ***Case Names*** command); if they are not displayed, the case numbers are shown.

Variable headers. These cells, located at the top of each column, contain header information for each variable. To display details about an individual variable, double-click on the variable header cell. To select the variable header only (for formatting) click once in the upper portion of the variable header (the mouse pointer will be an outlined plus sign ✙). To select the variable column (for editing) click once in the lower portion of the variable header (the mouse pointer will be an outlined plus sign with an arrow ✙⇩). To autofit the variable column, double-click on the right side of the variable header (the mouse pointer will be a cross with a double-headed arrow ↔). In the previous image, the first two variable header cells contain the text *GENDER* and *ADVERT*. You have the option to change how the variable header cells display information so that they show the column number associated with the variable, the variable long name, and/or an abbreviation of the display types for the variables in the spreadsheet. Each of these options is available on the ***View*** tab in the ***Display*** group; click ***Variable Headers***.

Data (and in-cell formatting options). The remainder of the spreadsheet contains data that pertain to the cases and variables and any optional attached or linked objects (multimedia objects, macros, custom user interface). Text in cells can be of practically unlimited length (in most *STATISTICA* configurations it is limited to 1,000 characters to protect against inadvertent pasting of unwanted large amounts of data into one cell). Text in cells can be extensively formatted including wrapping the text, different fonts, and font attributes.

Input vs. Output Spreadsheets

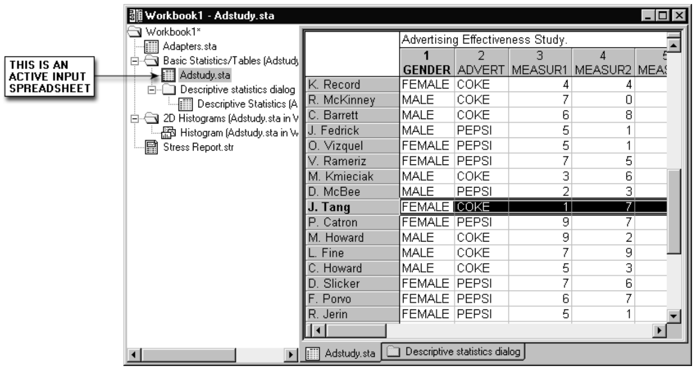
STATISTICA offers the ability to open and use many spreadsheets at the same time, allowing you to work with several different input data files simultaneously. In addition to storing data, *STATISTICA* uses spreadsheets to display the numeric output from its analyses. Because *STATISTICA* makes no distinction in the features supported for an input spreadsheet (from which *STATISTICA* retrieves its data) and an output spreadsheet (where the results of an analysis are displayed), it is easy to use the results of one analysis as input data for further analyses.

Any spreadsheet opened from a disk file is automatically treated as an input spreadsheet, and any number of input spreadsheets can be open at a time. To avoid confusion, however, an output spreadsheet (containing the results of an analysis) is not automatically available as input data for analysis. It must first be designated as an input spreadsheet before being used for further analyses. Additionally, input spreadsheets report the number of variables and cases for that spreadsheet in the title bar. For example, if *Exp.sta (88v by 48c)* is in the title bar, it is an input spreadsheet; if *Exp.sta* is in the title bar, it is not an input spreadsheet.

To designate an output spreadsheet as an input spreadsheet, select the spreadsheet (i.e., ensure the spreadsheet has the focus). Then, on the ***Data*** tab in the ***Mode*** group, select the ***Input*** check box. Now you can begin an analysis, and *STATISTICA* will use the data from the specified input spreadsheet for the analysis. Note that if you switch back to another spreadsheet that has previously been designated as an input spreadsheet, it can still be used for analyses as well.

In a workbook, only one spreadsheet can be selected for analyses at a time, even if the workbook contains several input spreadsheets. This spreadsheet is called the Active Input spreadsheet, and its icon (in the workbook tree) is framed in red.

By default, when an output spreadsheet is designated as an input spreadsheet, *STATISTICA* automatically selects it as the Active Input spreadsheet. To select another input spreadsheet for active input, select the ***Active Input*** check box on the ***Workbook*** tab in the ***Items*** group, or select ***Use as Active Input*** from the workbook tree shortcut menu.



THIS IS AN ACTIVE INPUT SPREADSHEET

Workbook1 - Adstudy.sta
Workbook1*
Adapters.sta
Basic Statistics/Tables (Adstudy
Adstudy.sta
Descriptive statistics dialog
Descriptive Statistics (A
2D Histograms (Adstudy.sta in V
Histogram (Adstudy.sta in W
Stress Report.str
Adstudy.sta
Descriptive statistics dialog

Advertising Effectiveness Study.

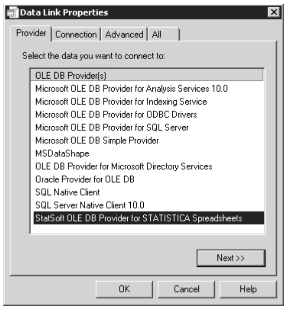
	1 GENDER	2 ADVERT	3 MEASUR1	4 MEASUR2	5 MEAS
K. Record	FEMALE	COKE	4	4	
R. McKinney	MALE	COKE	7	0	
C. Barrett	MALE	COKE	6	8	
J. Fedrick	MALE	PEPSI	5	1	
O. Vizquel	FEMALE	PEPSI	5	1	
V. Rameriz	FEMALE	PEPSI	7	5	
M. Kmieciak	MALE	COKE	3	6	
D. McBee	MALE	PEPSI	2	3	
J. Tang	FEMALE	COKE	1	7	
P. Catron	FEMALE	PEPSI	9	7	
M. Howard	MALE	COKE	9	2	
L. Fine	MALE	COKE	7	9	
C. Howard	MALE	COKE	5	3	
D. Slicker	FEMALE	PEPSI	7	6	
F. Porvo	FEMALE	PEPSI	6	7	
R. Jerin	FEMALE	PEPSI	5	1	

It is also possible to leave a stand-alone spreadsheet open but designate it as unavailable for analysis. To do this, select the spreadsheet, and clear the ***Input*** check box on the ***Data*** tab in the ***Mode*** group. Now *STATISTICA* automatically defaults to the most recently selected input spreadsheet for analysis, ignoring all spreadsheets that are not designated as input spreadsheets.

STATISTICA Spreadsheet OLE DB Provider

In addition to using spreadsheets as data sources for analyses in *STATISTICA*, spreadsheets can also supply data to other database-aware applications by using the ***StatSoft OLE DB Provider for STATISTICA Spreadsheets***. This OLE DB driver is installed with *STATISTICA*, and allows read-only access to data in *STATISTICA* Spreadsheets using the industry-standard Structured Query Language (SQL). You can access the OLE DB Provider at any point the system allows you to choose a database connection, using the standard Microsoft Data Link Properties.

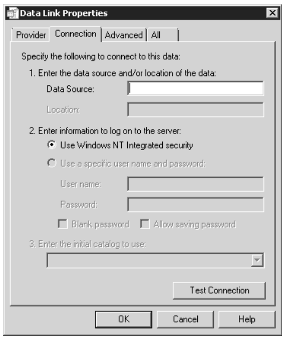
To access this functionality, select the ***Data*** tab. In the ***Manage*** group, click ***External Data*** and from the drop-down list, select ***Create Query***. In the ***Database Connection*** dialog, click the ***New*** button to display the ***Data Link Properties*** dialog, where you select ***StatSoft OLE DB Provider for STATISTICA Spreadsheets***.



Select the data you want to connect to:

OLE DB Provider(s)
Microsoft OLE DB Provider for Analysis Services 10.0
Microsoft OLE DB Provider for Indexing Service
Microsoft OLE DB Provider for ODBC Drivers
Microsoft OLE DB Provider for SQL Server
Microsoft OLE DB Simple Provider
MSDataShape
OLE DB Provider for Microsoft Directory Services
Oracle Provider for OLE DB
SQL Native Client
SQL Server Native Client 10.0
StatSoft OLE DB Provider for STATISTICA Spreadsheets

Click the ***Next*** button to display the ***Connection*** tab.



Specify the following to connect to this data:

1. Enter the data source and/or location of the data:

Data Source:

Location:

2. Enter information to log on to the server:

● Use Windows NT Integrated security

○ Use a specific user name and password:

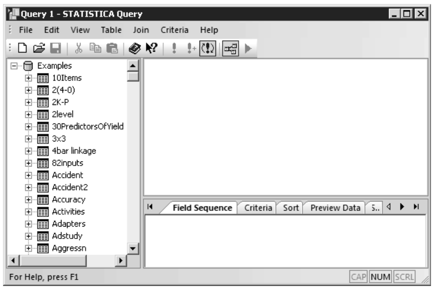
User name:

Password:

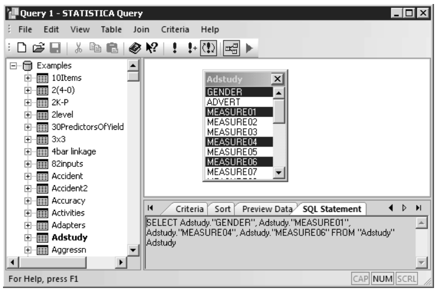
□ Blank password □ Allow saving password

3. Enter the initial catalog to use:

The ***Data Source*** field specifies the directory path where the spreadsheet is located. When creating the query, you can choose individual spreadsheet files within that directory. The following example uses *STATISTICA Query*, and has defined a connection to the Spreadsheet OLE DB, specifying the path of the *STATISTICA Examples* folder. Each spreadsheet within the folder shows up as a potential table.



These spreadsheets can be referenced in *FROM* clauses, specific variable names selected as fields in *SELECT* clauses, and cases defined with *WHERE* clauses. Joins between multiple spreadsheets are supported as well, using standard *JOIN* clauses.

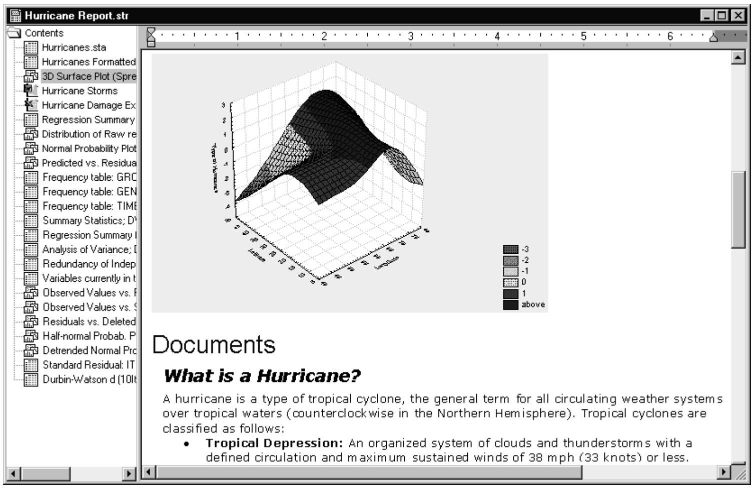


```
SELECT Adstudy."GENDER", Adstudy."MEASURE01",
Adstudy."MEASURE04", Adstudy."MEASURE06" FROM "Adstudy"
Adstudy
```

Using the ***StatSoft OLE DB Provider for STATISTICA Spreadsheets*** enables you to provide *STATISTICA* Spreadsheet data to any application (including *STATISTICA* itself) that can use the industry-standard OLE DB interface for querying data.

REPORTS

Reports (briefly introduced on page 150) in *STATISTICA* offer a more traditional way of handling output (compared to workbooks) as each object (e.g., a *STATISTICA* Spreadsheet or Graph, or an Excel spreadsheet) is displayed sequentially in a word processor style document.



Documents

What is a Hurricane?

A hurricane is a type of tropical cyclone, the general term for all circulating weather systems over tropical waters (counterclockwise in the Northern Hemisphere). Tropical cyclones are classified as follows:

- **Tropical Depression:** An organized system of clouds and thunderstorms with a defined circulation and maximum sustained winds of 38 mph (33 knots) or less.

However, the technology behind this simple report offers you rich functionality. For example, like the workbook, each *STATISTICA* Report is also an ActiveX (see page 238) container where each of its objects (not only *STATISTICA* Spreadsheets and Graphs, but also any other ActiveX-compatible documents, e.g., Word documents) is active, customizable, and in-place editable. Reports are stored in the STR file format, which is a StatSoft extension of the Microsoft RTF (Rich Text Format, **.rtf*) format. STR files share the RTF formatting information and additionally they include the tree view information (which cannot be stored in the standard RTF files). Hence, report files are by default saved with the file name extension **.str*, but they can also be saved as standard RTF files (in which case the tree information will not be preserved).

The obvious advantages of this way of handling output (more traditional than the workbook) are the ability to insert notes and comments "in between" the objects as well as its support for the more traditional way of quickly scrolling through and reviewing the output to which some users may be accustomed. Also, only the report output includes and preserves a record of the supplementary information, which contains a detailed log of the options specified for the analyses (e.g., selected variables and their labels, long names, etc., depending on the level of supplementary information specified in the ***Output Manager***, see page 25).

The obvious drawback, however, of these traditional reports is the inherent flat structure imposed by their word processor style format, though that is what some users of certain applications may favor.

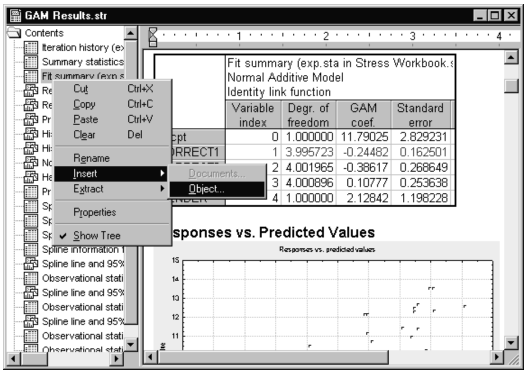
Navigating the Report Tree

The report tree displays the organization of files in the report. The files are displayed in an Explorer-style format; however, unlike workbooks that can support any number of levels, the report supports only one level of files.

You can embed any type of *STATISTICA* document in a report, including spreadsheets, graphs, and analyses. In addition to *STATISTICA* document types, you can embed other types of ActiveX/OLE objects in a report, including Excel spreadsheets, Word documents, bitmap images, and others. To edit one of these types of embedded documents, double-click on the document. The file opens in the viewer, and the report toolbar merges with the toolbar from the embedded file's native application, giving you access to all of the editing features you need.

Items in the tree are identified by the icon next to them. The spreadsheet, macro, and graph icons represent *STATISTICA* Spreadsheet, Macro, and Graph documents, respectively. The *Data Miner* icon represents a *Data Miner* workspace. All non-*STATISTICA* documents are represented by their document icons. For example, Word documents are represented by the Word icon, and Excel spreadsheet files are represented by the Excel spreadsheet icon.

The report tree can be organized and modified using drag-and-drop features as well as Clipboard procedures.



Fit summary (exp.sta in Stress Workbook.s
Normal Additive Model
Identity link function

	Variable index	Degr. of freedom	GAM coef.	Standard error
cpt	0	1.000000	11.79025	2.829231
RRECT1	1	3.995723	-0.24482	0.162501
	2	4.001965	-0.38617	0.268649
	3	4.000896	0.10777	0.253638
NDER	4	1.000000	2.12842	1.198228

sponses vs. Predicted Values

Commands for inserting, extracting, renaming, and removing items from the report tree are available from the report tree shortcut menu (accessed by right-clicking anywhere in the tree, as shown above).

GRAPHS

Graphs represent another distinctive type of *STATISTICA* documents, and they offer rich functionality both in terms of the variety of ways in which graphs can be created in *STATISTICA* and in the selection of graph customization tools.

Similar to the other *STATISTICA* documents, graphs are ActiveX containers (see page 238), which means that they can contain a variety of compatible documents (e.g., Visio drawings, Adobe illustrations, Excel spreadsheets, etc.). *STATISTICA* Graphs are also ActiveX objects and, therefore, can be linked to or embedded into other compatible documents (e.g., Word documents) where they can be in-place edited by simply double-clicking on them.

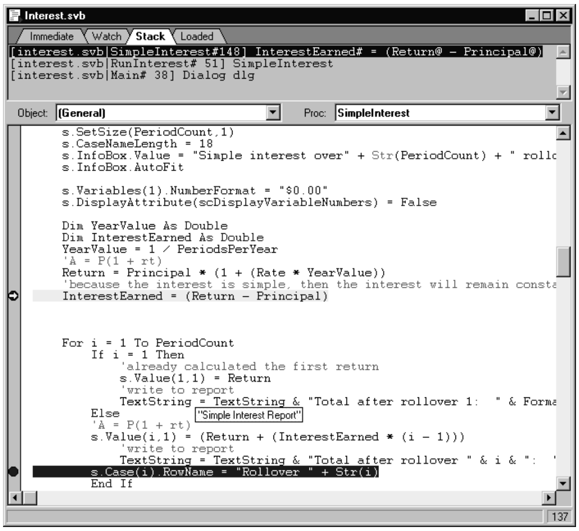
Graphs are discussed in more detail in Chapter 6 – *Graphs*.

MACROS *(STATISTICA* VISUAL BASIC PROGRAMS)

The industry standard *STATISTICA* Visual Basic (SVB) language (integrated into *STATISTICA*) offers another (alternative) user interface to the functionality of *STATISTICA*, and it offers incomparably more than just a "supplementary application programming language" that can be used to write custom extensions.

Note that *STATISTICA* Visual Basic is not Microsoft Visual Basic 6.0. StatSoft owns and maintains the code for *STATISTICA* Visual Basic. SVB is compatible with Microsoft's VB.NET, Microsoft's Visual Basic for Applications (VBA), and also with Microsoft's Visual Basic 6.0 (VB6). SVB scripting language is unique in terms of its flexibility and compatibility, and it is also very powerful. It provides access to Visual Basic for Applications (used for scripting Microsoft Office products) and access to the .NET Framework within the same file (see Chapter 10 – Programming *STATISITCA* from .NET, page 247). Other APIs can also be accessed and leverage the flexibility of SVB such as, for example, Yahoo's Stock Quote API or Google Analytics API. SVB offers a powerful 64-bit solution for system integration, expansion, and custom development.

STATISTICA Visual Basic takes full advantage of the object model architecture of *STATISTICA* and is used to access programmatically every aspect and virtually every detail of the functionality of *STATISTICA*. Even the most complex analyses and graphs can be recorded into Visual Basic macros and later be run repeatedly or edited and used as building blocks of other applications. *STATISTICA* Visual Basic adds an arsenal of more than 14,000 new functions to the standard comprehensive syntax of Visual Basic, thus comprising one of the largest and richest development environments available.



```
[interest.svb|SimpleInterest#148] InterestEarned# = (Return@ - Principal@)
[interest.svb|RunInterest# 51] SimpleInterest
[interest.svb|Main# 38] Dialog dlg
```

```
    s.SetSize(PeriodCount,1)
    s.CaseNameLength = 18
    s.InfoBox.Value = "Simple interest over" + Str(PeriodCount) + " rollc
    s.InfoBox.AutoFit

    s.Variables(1).NumberFormat = "$0.00"
    s.DisplayAttribute(scDisplayVariableNumbers) = False

    Dim YearValue As Double
    Dim InterestEarned As Double
    YearValue = 1 / PeriodsPerYear
    'A = P(1 + rt)
    Return = Principal * (1 + (Rate * YearValue))
    'because the interest is simple, then the interest will remain consta
    InterestEarned = (Return - Principal)



    For i = 1 To PeriodCount
        If i = 1 Then
            'already calculated the first return
            s.Value(1,1) = Return
            'write to report
            TextString = TextString & "Total after rollover 1:  " & Forma
        Else
        'A = P(1 + rt)
        s.Value(i,1) = (Return + (InterestEarned * (i - 1)))
            'write to report
            TextString = TextString & "Total after rollover " & i & ":  '
        s.Case(i).RowName = "Rollover " + Str(i)
        End If
```

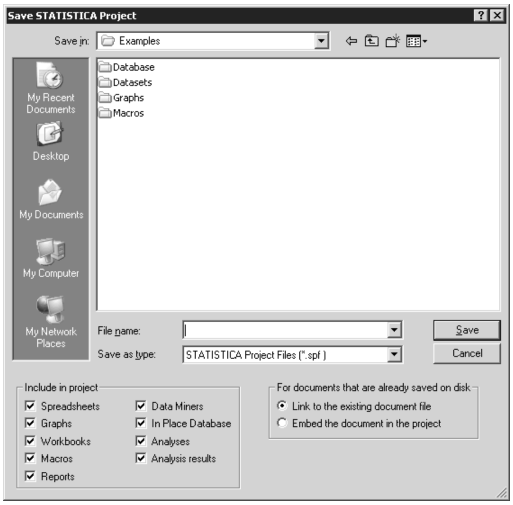
STATISTICA Macros can be saved in several formats, depending on how you intend to use them (see the *STATISTICA Visual Basic Primer* and the *Electronic Manual* for more information). You can also copy them to the Clipboard and paste them into other programs or documents.

STATISTICA Visual Basic is discussed in more detail in Chapter 8 (page 219).

STATISTICA PROJECTS

When performing statistical analyses and working with *STATISTICA* documents, you will often have many different windows open, and even different analyses in different stages of progress. *STATISTICA* provides a means for saving your workspace, including any analyses in progress. You can close *STATISTICA* at any point during an analysis, and when you later re-open the project, the previously opened files and in-process analyses will be restored.

To save a *STATISTICA* Project, select the ***Home*** tab, click the ***Save*** arrow in the ***Project*** group, and select ***Save Project As*** to display the ***Save STATISTICA Project*** dialog.



Save as type: STATISTICA Project Files (*.spf)

Include in project
☑ Spreadsheets
☑ Graphs
☑ Workbooks
☑ Macros
☑ Reports
☑ Data Miners
☑ In Place Database
☑ Analyses
☑ Analysis results

For documents that are already saved on disk
● Link to the existing document file
○ Embed the document in the project

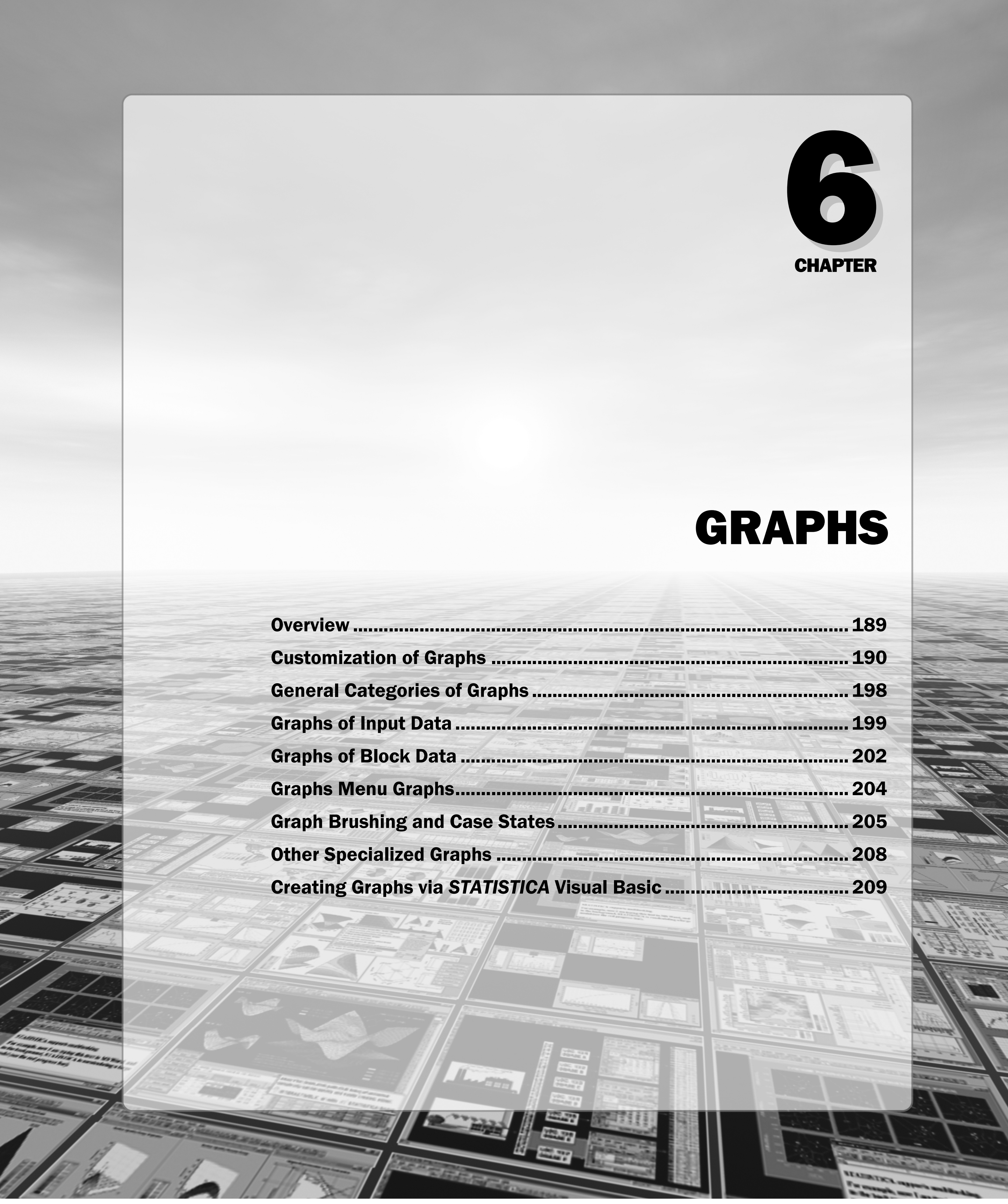
In this dialog, specify the path and file name of the *STATISTICA* Project file (a project's extension is *.spf*). You can also specify what items to include in the project. All *STATISTICA* document types can be selected (***Spreadsheets***, ***Graphs***, ***Workbooks***, ***Macros***, ***Reports***, ***Data Miner*** projects, ***In-Place Database*** projects, ***Analyses***, and ***Analysis results***). For those *STATISTICA* documents that are already stored on disk, you have the option to either ***Link to the existing document file***, or to store a copy of the document within the *STATISTICA* Project file (***Embed the document in the project***).

In addition to *STATISTICA* documents, project files will also save all in-progress analyses. The project file will store the recorded scripts that are automatically created when every analysis is run. When the project is re-opened, the scripts for the analyses are re-run against the original data and the analyses dialogs are made visible again in exactly the state they were when the project file was saved.

Project files are a convenient way to send in-progress analysis steps and results back and forth between users if you elect to embed the saved documents in the project file. One user can run analyses to a certain point, and then save the project file and pass it to another user, who can open the project file and continue exactly where the first user stopped the analyses.

Unless you configure it otherwise, *STATISTICA* will automatically display a prompt asking if you want to save a project file when quitting the program, and will automatically re-open the last-saved project file when starting. Thus, *STATISTICA* makes it easy to quit for the day and start the next session right where you left off.

Note that a project is a state of an instance of *STATISTICA*. Thus, projects are not like other documents in that you cannot open more than one project in a single instance of *STATISTICA*. A different (second) project can be opened in a second instance of *STATISTICA*.



6

CHAPTER

GRAPHS

6

CHAPTER

GRAPHS

OVERVIEW

The most common application of graphs is to efficiently present and communicate information (typically, numerical data). However, graphical techniques also provide powerful analytical tools for the exploration of data and verification of hypotheses.

A broad selection of graphics options. *STATISTICA* includes a comprehensive selection of graphical methods for both data analysis and the presentation of results. All graphs in *STATISTICA* include a broad selection of built-in, interactive analytic techniques and extensive customization tools that enable you to interactively control virtually all aspects of the display. Also, flexible graphics management facilities are available that are used to integrate various graphical displays and to build dynamic links between applications (e.g., using OLE-Object Linking and Embedding).

Comprehensive support for Visual Basic and other languages. *STATISTICA* graphical options can also be accessed programmatically (using built-in *STATISTICA* Visual Basic or other compatible languages), which creates practically unlimited possibilities for producing highly customized graphical displays. These custom graphs can later be permanently added to *STATISTICA*'s user interface (e.g., assigned to buttons on toolbars or added to the menus).

General categories of graphs. The *STATISTICA* system offers a variety of methods in which graphs can be requested or defined. These methods (constituting broad categories of graphs, such as input data, block data, and specialized) are reviewed in *General Categories of Graphs* on page 198; they complement each other, providing a high level of integration between numbers (such as raw data, intermediate results, or final results) and graphical displays. For example, specialized graphs can be requested as part of the automatic output from statistical procedures, but they can also be requested via integrated tools to visualize virtually any combination of numbers (and/or labels) that are displayed or generated by *STATISTICA*.

CUSTOMIZATION OF GRAPHS

Interactive graph customization. The customization options in *STATISTICA* graphics include hundreds of features and tools that can be used to adjust every detail of the display and associated data processing. However, these options are arranged in a hierarchical manner, so those used most often are accessible directly via shortcuts by double-clicking or right-clicking on the respective element of the graph.

Permanent settings and automation options. The initial (default) settings of all of these features can be easily adjusted so that even the default appearance and behavior of *STATISTICA* graphs will match your specific needs and/or will require very little intervention on your part. Following are some of the ways to make these adjustments:

1. **Options dialog.** Perhaps the most straightforward way to adjust the default appearance of graphs is by modifying the graph options in the ***Options*** dialog (select the ***Tools*** tab and click ***Options***). Most commonly used settings can be easily adjusted there (select ***Display*** or ***Settings***, located under ***Graphs***), and the results will be reflected in the default styles (see number 2 below) that will be used by the system and as such, they will be automatically saved in the *STATISTICA* configuration file (e.g., different settings can be used for different projects). For further details, see the documentation for the ***Configurations*** options pane of the ***Options*** dialog in the *Electronic Manual*.
2. **Graph style system.** All of the numerous features that affect the appearance of the graph (from as elementary as the color of the font in the footnote to as general as the global features of the graph document) can be

saved as individual "styles." These styles can be given custom names and later be reapplied using simple shortcuts (such as pressing a specific key combination or clicking a button on a custom toolbar). An intelligent system internally manages these thousands of styles and their combinations in *STATISTICA* and helps you achieve your customization objectives with a minimum amount of effort. All user-defined or modified styles will be saved automatically in the *STATISTICA* configuration file (e.g., different sets or systems of styles can be used for different projects). For further details, see the documentation for the ***Configurations*** options pane of the ***Options*** dialog in the *Electronic Manual*.

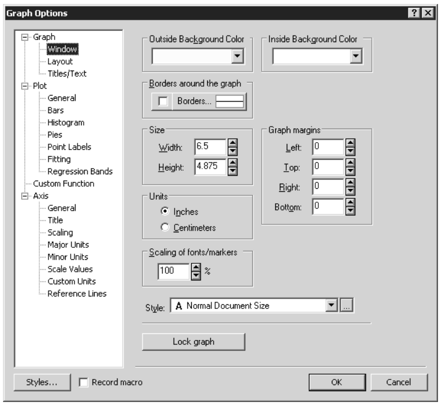
3. **User-defined graphs.** New types of graphs can be defined in a variety of ways and can be added to the menus, dialogs, or toolbars. If a custom graph that you intend to use repeatedly is not built "from scratch" but is based on one of the ***Graphs*** menu graphs and is produced by some combination of the existing graph customization options, then adding it to the ***Graphs*** menu as a new type of graph is as simple as clicking the ***Add As User-defined Graph to Menu*** button on the ***Options 2*** tab of the graph specification dialog. All user-defined graph specifications will be saved automatically in the *STATISTICA* configuration file (e.g., different sets of custom graphs can be used for different projects). For further details, see the documentation for the ***Configurations*** options pane of the ***Options*** dialog in the *Electronic Manual*.

4. ***STATISTICA* Visual Basic.** Finally, note that there are no limits to how "deeply customized" your *STATISTICA* custom graphs can be, because *STATISTICA* Visual Basic (with all its powerful custom drawing tools as well as the *STATISTICA*-based library of graphics procedures) can be used to produce virtually any graphics or multimedia output supported by the contemporary computer hardware. Those custom developed displays or multimedia output can be assigned to *STATISTICA* toolbars, menus, or dialogs and become a permanent part of "your" *STATISTICA* application.

See the *Electronic Manual* (*STATISTICA* Help) for further details on these graph customization methods. The *Electronic Manual* also contains topics devoted to specific categories of graphs, includes conceptual overviews and examples of typical applications, and discusses distinctive functional properties of the respective types of graphs.

The default settings of most graphs offered in *STATISTICA* follow the established conventions that are either explicitly described in the literature on statistical and technical graphing, or they represent standards that are commonly accepted by major scientific journals (e.g., *SCIENCE*). However, practically all default settings of *STATISTICA* can be customized to meet specific requirements of unusual applications (see page 190). *STATISTICA*'s graphics facilities were designed to play the role of flexible tools, capable of producing effects that go far beyond established patterns and templates.

In addition to a comprehensive selection of standard statistical and technical graphs, *STATISTICA* includes numerous unique types of graphs and graph customization facilities. The ***Graph Options*** dialog, accessible by double-clicking in the background of a graph, or selecting the ***Tools*** tab and clicking ***Graph*** in the ***Options*** group, contains options that address all of the relevant customizable features for a particular graph. The options are grouped in clusters containing logically related items, and are an all-inclusive "superset" of graph shortcut options accessed by double-clicking specific graph features.



Size
Width: 6.5
Height: 4.875

Graph margins
Left: 0
Top: 0
Right: 0
Bottom: 0

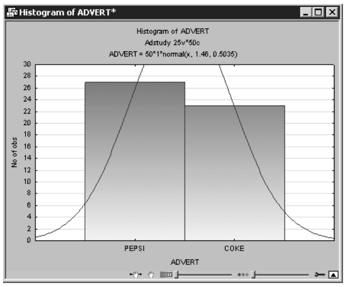
Units
● Inches
○ Centimeters

Scaling of fonts/markers
100 %

Style: A Normal Document Size

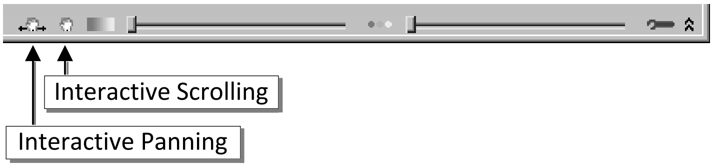
□ Record macro

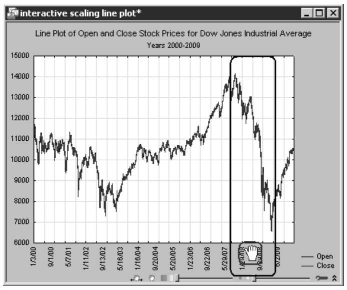
Located at the bottom of graphs, you'll find the interactive graphics controls (see the next illustrations), which enable you to adjust the transparency of the plot areas and markers, and to scroll and pan in order to interactively scale the graph. More controls are located in 3D graphs to enable interactive rotation. Click the wrench icon adjacent to the sliders to display the ***Graph Options*** dialog.



Left: 2D Graph

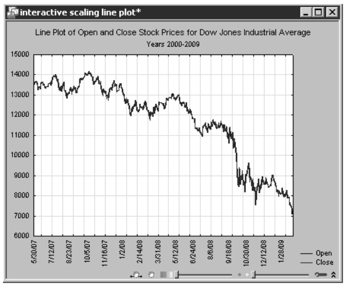
Below: Enlarged image of Panning (scaling), Scrolling, and Transparency Controls

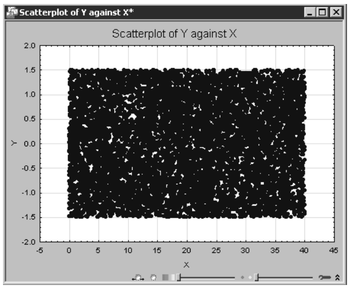



interactive scaling line plot*
Line Plot of Open and Close Stock Prices for Dow Jones Industrial Average
Years 2000-2009
Open
Close

Left: Section to be scaled is circled

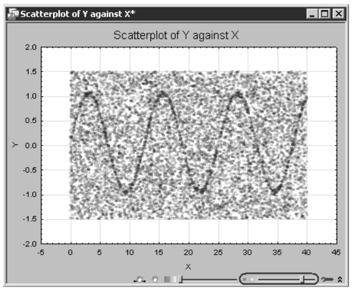
Right: Scaled view of left graph's circled area

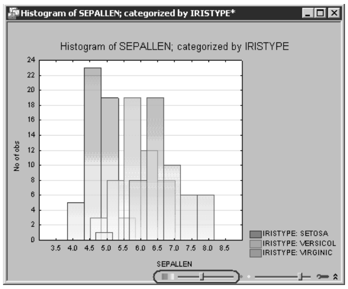

interactive scaling line plot*
Line Plot of Open and Close Stock Prices for Dow Jones Industrial Average
Years 2000-2009
Open
Close



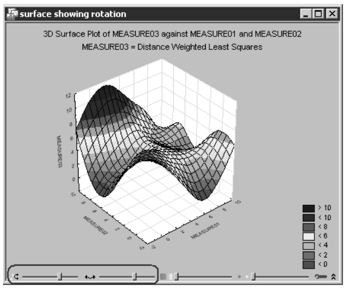
Left: Scatterplot with dense concentration of data points

Right: Transparency Control reveals hidden trends



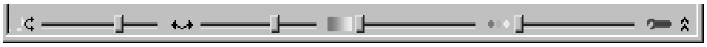

Histogram of SEPALLEN; categorized by IRISTYPE*
Histogram of SEPALLEN; categorized by IRISTYPE
No of obs
SEPALLEN
IRISTYPE: SETOSA
IRISTYPE: VERSICOL
IRISTYPE: VIRGINIC

Left: Plot Area Transparency Control circled; making plot areas transparent allows portions of the plot to overlap while still being visible


surface showing rotation
3D Surface Plot of MEASURE03 against MEASURE01 and MEASURE02
MEASURE03 = Distance Weighted Least Squares
MEASURE03
MEASURE02
MEASURE01
> 10
< 10
< 8
< 6
< 4
< 2
< 0

Left: 3D Graph; Rotation Controls circled

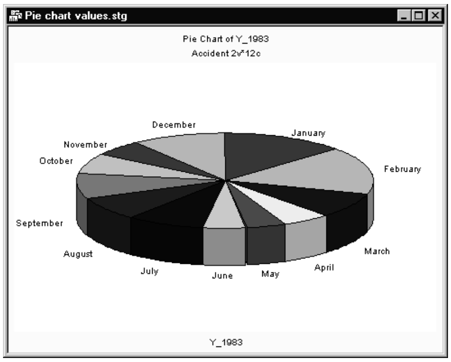
Below: Enlarged image of Rotation and Transparency Controls



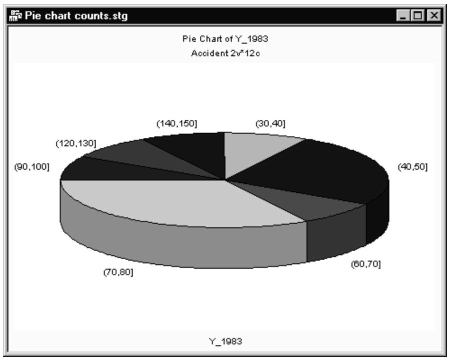
While StatSoft statisticians designed most of the graph customization options, it is important to say that *STATISTICA* users have played a significant role in their creation. In fact, the selection of graphics options included in *STATISTICA* is the result of input from thousands of users who provided their comments in response to StatSoft's inquiries. Many unique facilities of *STATISTICA* Graphs were introduced in response to users' ideas and requests. We at StatSoft are very grateful for the input from our users.

As mentioned previously (and discussed in detail on page 198), there are various methods to specify *STATISTICA* Graphs. You could say that these methods represent different types of "interfaces" between numbers and graphs.

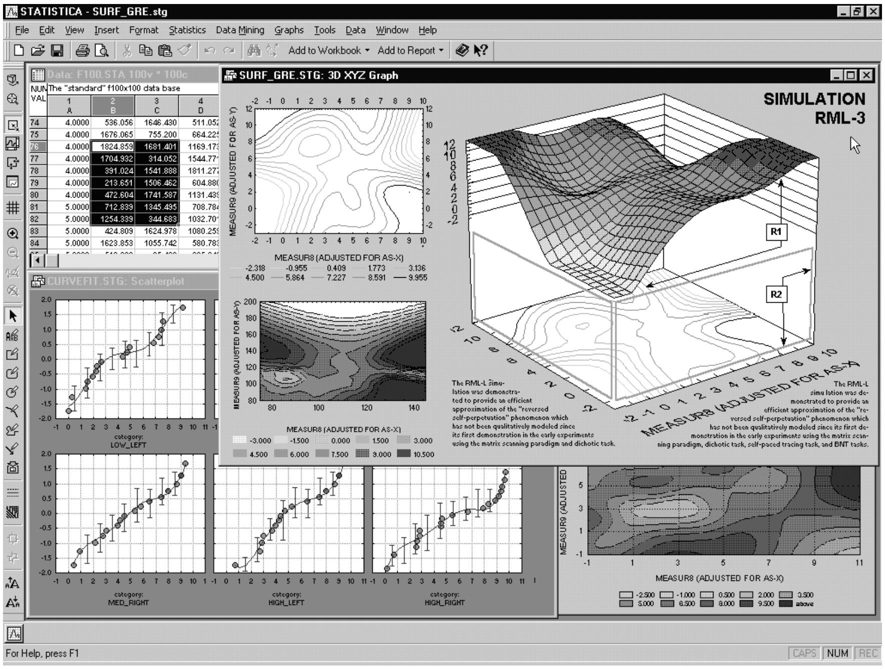
For example, the numbers represented in a pie chart can simply depict values of a spreadsheet column (e.g., variable *Sales*) in the consecutive cases of the spreadsheet (e.g., cases labeled: *Year 2008, Year 2009, Year 2010,* ..., etc.).



The numbers in a similar pie chart, however, can represent results of calculations. For example, the slices of the pie can represent relative frequencies of observations that belong to certain categories calculated by one of the histogram or frequency categorization procedures (e.g., numbers of years when the *Sales* were below $10 million, between $10 and $20 million, and above $20 million).



Regardless of the method that was used to create a graph (i.e., regardless of where the numbers represented in the graph were obtained or how they were calculated), all *STATISTICA* Graph customization and multigraphics management facilities can be used to change the appearance of the graph or integrate it with other graphs or documents.



The "standard" f100x100 data base

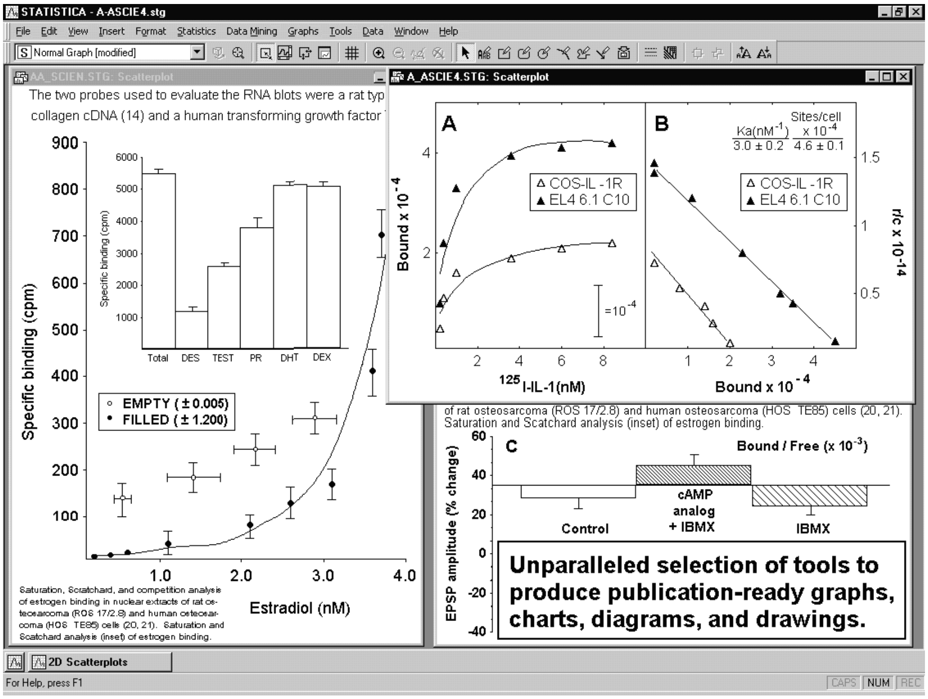
NUM VAL	1 A	2 B	3 C	4 D
74	4.0000	536.056	1646.430	511.052
75	4.0000	1676.065	755.200	664.225
76	4.0000	1824.859	1681.401	1169.173
77	4.0000	1704.932	314.052	1544.771
78	4.0000	391.024	1541.888	1811.277
79	4.0000	213.651	1506.462	604.880
80	4.0000	472.604	1741.587	1131.439
81	5.0000	712.839	1345.495	708.784
82	5.0000	1254.339	344.683	1032.701
83	5.0000	424.809	1624.978	1080.259
84	5.0000	1623.853	1055.742	580.783

The RML-L Simulation was demonstrated to provide an efficient approximation of the "reversed self-perpetuation" phenomenon which has not been qualitatively modeled since its first demonstration in the early experiments using the matrix scanning paradigm and dichotic task.

The RML-L simu lation was demonstrated to provide an efficient approximation of the "reversed self-perpetuation" phenomenon which has not been qualitatively modeled since its first demonstration in the early experiments using the matrix scanning paradigm, dichotic task, self-paced tracing task, and BNT tasks.

Also, all integrated analytic facilities that are accessible from within graphs in *STATISTICA* (such as function fitting, smoothing, rotation, brushing, analytical zooming, etc.) are available and can be applied to the graph regardless of the source of the numbers in the graph or the method that was used to create it.

The graph editing facilities offered in *STATISTICA* enable you to create not only highly customized scientific and technical publication-ready displays:



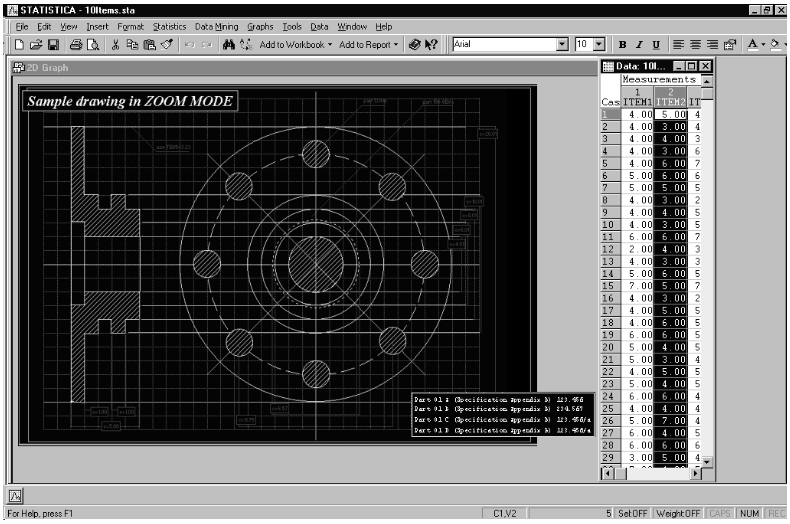
The two probes used to evaluate the RNA blots were a rat typ
collagen cDNA (14) and a human transforming growth factor

Saturation, Scratchard, and competition analysis of estrogen binding in nuclear extracts of rat osteosarcoma (ROS 17/2.8) and human osteosarcoma (HOS TE85) cells (20, 21). Saturation and Scatchard analysis (inset) of estrogen binding.

of rat osteosarcoma (ROS 17/2.8) and human osteosarcoma (HOS TE85) cells (20, 21). Saturation and Scatchard analysis (inset) of estrogen binding.

Unparalleled selection of tools to produce publication-ready graphs, charts, diagrams, and drawings.

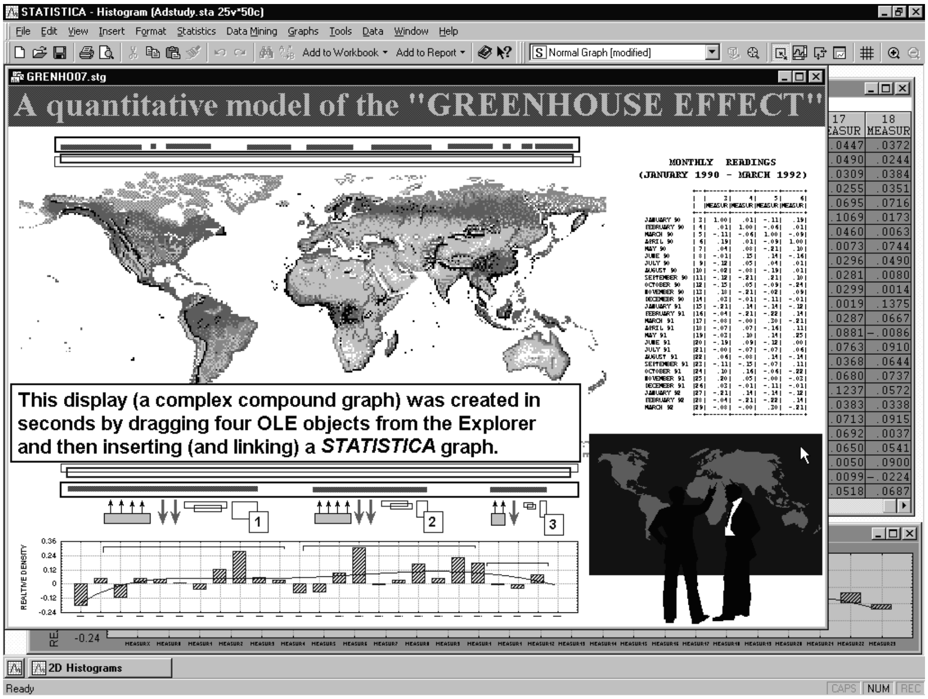
and precise drawings:


STATISTICA - 10Items.sta
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
Add to Workbook Add to Report
Arial
10
B I U
2D Graph
Sample drawing in ZOOM MODE
Data: 10I...
For Help, press F1
C1,V2
5
Sel:OFF
Weight:OFF
CAPS
NUM
REC

Part 01 A (Specification Appendix B) 123.456
Part 01 B (Specification Appendix B) 234.567
Part 01 C (Specification Appendix B) 123.456/a
Part 01 D (Specification Appendix B) 123.456/a

	Measurements		
Cas	1 ITEM1	2 ITEM2	IT
1	4.00	5.00	4
2	4.00	3.00	4
3	4.00	4.00	3
4	4.00	3.00	6
5	4.00	6.00	7
6	5.00	6.00	6
7	5.00	5.00	5
8	4.00	3.00	2
9	4.00	4.00	5
10	4.00	3.00	5
11	6.00	6.00	7
12	2.00	4.00	3
13	4.00	3.00	3
14	5.00	6.00	5
15	7.00	5.00	7
16	4.00	3.00	2
17	4.00	5.00	5
18	4.00	6.00	5
19	6.00	6.00	5
20	5.00	4.00	5
21	5.00	3.00	4
22	4.00	5.00	5
23	5.00	4.00	5
24	6.00	6.00	4
25	4.00	4.00	4
26	5.00	7.00	4
27	6.00	4.00	5
28	6.00	6.00	6
29	3.00	5.00	4

but also presentation-quality diagrams, posters, business charts, and other displays:


STATISTICA - Histogram (Adstudy.sta 25v*50c)
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
Add to Workbook Add to Report
Normal Graph [modified]
GRENH007.stg
A quantitative model of the "GREENHOUSE EFFECT"
1
2
3
0.36
0.24
0.12
0
-0.12
-0.24
REALTIVE DENSITY
-0.24
2D Histograms
Ready
CAPS
NUM
REC

This display (a complex compound graph) was created in seconds by dragging four OLE objects from the Explorer and then inserting (and linking) a *STATISTICA* graph.

```
                 MONTHLY  READINGS
          (JANUARY 1990 - MARCH 1992)

                +--+------+------+------+------+
                |  |     3|     4|     5|     6|
                |  |MEASUR|MEASUR|MEASUR|MEASUR|
                +--+------+------+------+------+
JANUARY 90      | 3|  1.00|   .01|  -.11|   .19|
FEBRUARY 90     | 4|   .01|  1.00|  -.06|   .01|
MARCH 90        | 5|  -.11|  -.06|  1.00|  -.09|
APRIL 90        | 6|   .19|   .01|  -.09|  1.00|
MAY 90          | 7|   .04|   .08|  -.21|   .10|
JUNE 90         | 8|  -.01|   .15|   .14|  -.16|
JULY 90         | 9|  -.12|   .05|   .04|   .01|
AUGUST 90       |10|  -.02|  -.08|  -.19|   .01|
SEPTEMBER 90    |11|  -.12|  -.21|   .21|   .10|
OCTOBER 90      |12|  -.15|   .05|  -.09|  -.24|
NOVEMBER 90     |13|   .18|  -.21|  -.02|   .09|
DECEMEBR 90     |14|   .03|  -.01|  -.11|  -.01|
JANUARY 91      |15|  -.21|   .14|  -.14|  -.12|
FEBRUARY 91     |16|  -.04|  -.21|  -.22|   .14|
MARCH 91        |17|  -.08|  -.00|   .30|  -.21|
APRIL 91        |18|  -.07|   .07|  -.16|   .11|
MAY 91          |19|  -.03|   .10|   .14|   .25|
JUNE 91         |20|  -.19|   .09|  -.12|   .00|
JULY 91         |21|  -.00|  -.07|  -.07|   .06|
AUGUST 91       |22|   .06|  -.08|   .14|  -.14|
SEPTEMBER 91    |23|  -.11|  -.15|  -.07|   .11|
OCTOBER 91      |24|   .10|   .16|  -.06|  -.22|
NOVEMBER 91     |25|   .20|   .05|  -.00|  -.03|
DECEMEBR 91     |26|   .03|  -.01|  -.11|  -.01|
JANUARY 92      |27|  -.21|   .14|  -.14|  -.12|
FEBRUARY 92     |28|  -.04|  -.21|  -.22|   .14|
MARCH 92        |29|  -.08|  -.00|   .30|  -.21|
                +--+------+------+------+------+
```

17 EASUR	18 MEASUR
.0447	.0372
.0490	.0244
.0309	.0384
.0255	.0351
.0695	.0716
.1069	.0173
.0460	.0063
.0073	.0744
.0296	.0490
.0281	.0080
.0299	.0014
.0019	.1375
.0287	.0667
.0881	-.0086
.0763	.0910
.0368	.0644
.0680	.0737
.1237	.0572
.0383	.0338
.0713	.0915
.0692	.0037
.0650	.0541
.0050	.0900
.0099	-.0224
.0518	.0687

that are designed to communicate information in an effective and attractive manner.

Graphs that are saved into files or that in any other way have been temporarily detached from the *STATISTICA* application (e.g., copied to the Clipboard or linked to a document in another application) are complete “objects” (technically speaking, ActiveX objects, see page 238) that contain not only all customization features and other embedded objects, but also all data that are necessary to continue editing all aspects of the display or the analysis of its contents (fitting, smoothing, etc.).

Because *STATISTICA* Graphs are ActiveX objects, they can easily be linked to or embedded in other compatible documents (e.g., Excel or Word documents), where they can be in-place edited by double-clicking on them. *STATISTICA* Graphs are also ActiveX containers and, therefore, can contain a wide variety of embedded or linked documents such as Visio drawings, Adobe illustrations, Excel spreadsheets, or Word documents. Moreover, *STATISTICA* supports hierarchies of embedded objects up to four levels, which means that it can manage “documents containing documents, containing documents, which contain documents.”

GENERAL CATEGORIES OF GRAPHS

In addition to the specialized statistical graphs that are available from the output dialogs in all statistical procedures (see page 208), there are two general categories or classes of graphs both accessible from the ***Graphs*** tab, shortcut menus, and the *STATISTICA* Start button menu:

- Input data graphs (***Graphs of Input Data***, see page 199) and ***Graphs*** menu graphs, (see page 204) and
- ***Graphs of Block Data*** (see page 202).

The most important difference between these two general categories lies in the data that the graph types utilize for generating plots.

Input data graphs. ***Graphs of Input Data*** and their expanded version on the ***Graphs*** tab produce statistical summaries or other representations of the raw data in the current input data spreadsheet (typically for all the variables, or for subsets if case selection conditions are used). Note that if graphs of this general category are produced using a shortcut menu from within a spreadsheet of results that does not contain the actual data (e.g., a correlation matrix), *STATISTICA* will still reach to the respective input (raw) data to produce the graph (e.g., a scatterplot of the variables identified by the selected cell in the correlation matrix from which the shortcut menu was opened).

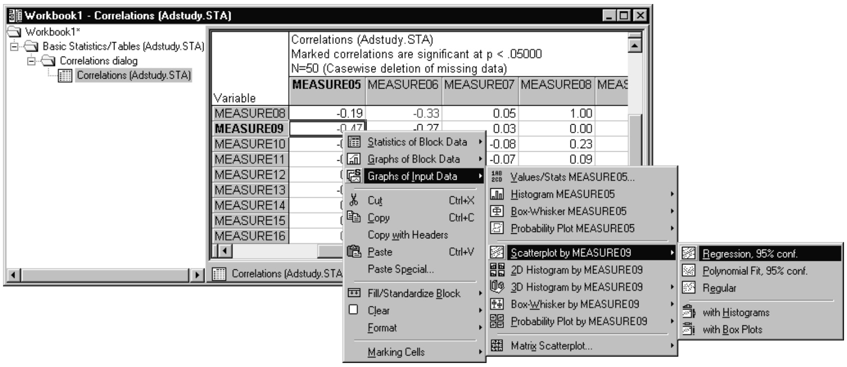
Graphs of Block Data. ***Graphs of Block Data***, however, are entirely independent of the concept of "input data" or "data file." They provide a general tool to visualize numeric values in the currently selected block of any spreadsheet (which can contain values from custom defined subsets of numerical output or arbitrarily selected subsets of raw data).

Common features of the two categories of graphs. These two general categories of graphs offer the same customization options and the same selection of types of graphs. For example, you can create the same highly specialized categorized ternary graph from the input (raw) data set and from a custom defined block of values representing results of a particular test.

These two general categories of graphs will be briefly discussed in the next two sections, followed by a section on the ***Graphs*** tab, which contains an exhaustive selection of all graphs from the first category (input data graphs, often referred to as ***Graphs*** menu graphs), as well as access to ***Graphs of Block Data*** and other options.

GRAPHS OF INPUT DATA

The ***Graphs of Input Data*** command is available from the shortcut menu of all spreadsheets, and it offers quick and simplified access to the most commonly used types of graphs based on the current input data set.



Workbook1 - Correlations (Adstudy.STA)

Correlations (Adstudy.STA)
Marked correlations are significant at p < .05000
N=50 (Casewise deletion of missing data)

Variable	MEASURE05	MEASURE06	MEASURE07	MEASURE08	MEAS
MEASURE08	-0.19	-0.33	0.05	1.00	
MEASURE09	-0.47	-0.27	0.03	0.00	
MEASURE10	-		-0.08	0.23	
MEASURE11	-		-0.07	0.09	
MEASURE12					
MEASURE13	-				
MEASURE14					
MEASURE15					
MEASURE16					

Note that all these graphs are also available on the ***Graphs*** tab, from the *STATISTICA* Start menu on the status bar, or by clicking the ***Graphs Gallery*** button on any graph specification dialog. ***Graphs of Input Data*** do not offer as many options as the corresponding ***Graphs*** menu graphs; however, they are quicker to select because unlike ***Graphs*** menu graphs:

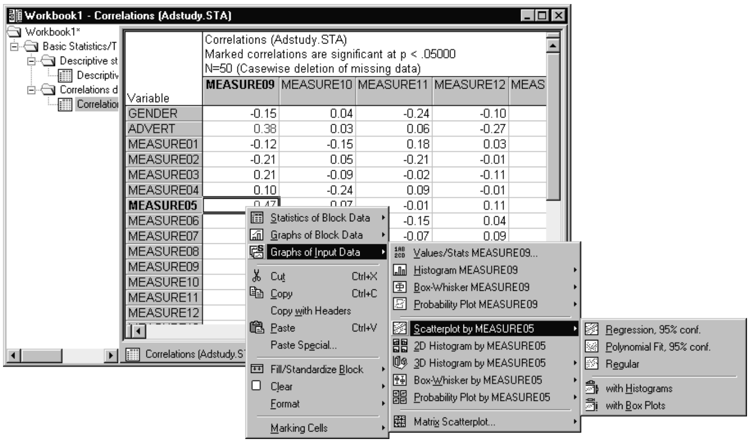
- ***Graphs of Input Data*** can be called directly from the spreadsheet shortcut menus,
- ***Graphs of Input Data*** do not require you to select variables (the variable selection is determined by the current cursor position within a spreadsheet), and
- ***Graphs of Input Data*** do not require you to select options from any intermediate dialogs (default formats of the respective graphs are produced).

Graphs of Input Data process data directly from the current input data file, and they take their cues as to which variables to use from the current cursor position (in any type of spreadsheet).

For example, if you right-click a single correlation in a results spreadsheet and create a ***Scatterplot by…*** graph, *STATISTICA* generates a 2D scatterplot using the original raw values of the two variables represented by that correlation (see the *Introductory Example* on page 11 for a more detailed example).

Although the most convenient way to select ***Graphs of Input Data*** is via the spreadsheet shortcut menu, you can also select them from the ***Graphs*** tab or the *STATISTICA* Start menu . Either method will display a submenu from which you can choose one of the statistical graphs applicable to the current variable (i.e., to the variable indicated by the current cursor position in the spreadsheet).

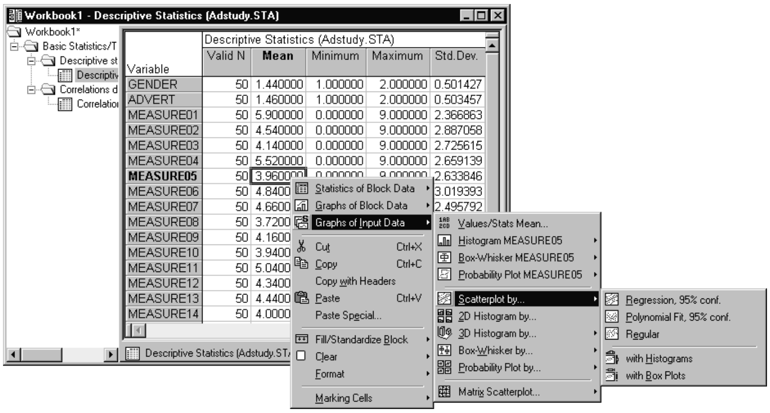
If the spreadsheet has a matrix format or a format where a cursor position indicates not one but two variables (as in the illustration showing a correlation matrix, below), then predefined bivariate graphs for the specified pair of variables will be directly available from the ***Graphs of Input Data*** submenus.



Correlations (Adstudy.STA)
Marked correlations are significant at p < .05000
N=50 (Casewise deletion of missing data)

Variable	MEASURE09	MEASURE10	MEASURE11	MEASURE12	MEAS
GENDER	-0.15	0.04	-0.24	-0.10	
ADVERT	0.38	0.03	0.06	-0.27	
MEASURE01	-0.12	-0.15	0.18	0.03	
MEASURE02	-0.21	0.05	-0.21	-0.01	
MEASURE03	0.21	-0.09	-0.02	-0.11	
MEASURE04	0.10	-0.24	0.09	-0.01	
MEASURE05	0.47	0.07	-0.01	0.11	
MEASURE06			-0.15	0.04	
MEASURE07			-0.07	0.09	
MEASURE08					
MEASURE09					
MEASURE10					
MEASURE11					
MEASURE12					

Otherwise, i.e., when the current cursor position indicates only one variable as in a table of descriptive statistics (as shown in the next illustration), and if you select any of the bivariate graphs in the menu, *STATISTICA* will prompt you to select the second variable. For example, if you select ***Scatterplot by***, the ***Select second variable*** dialog will be displayed, where you specify by which variable *Measure05* is going to be plotted.



Descriptive Statistics (Adstudy.STA)

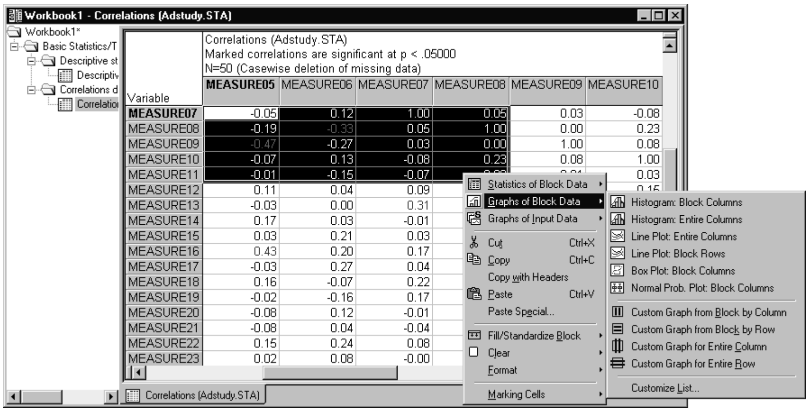
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
GENDER	50	1.440000	1.000000	2.000000	0.501427
ADVERT	50	1.460000	1.000000	2.000000	0.503457
MEASURE01	50	5.900000	0.000000	9.000000	2.366863
MEASURE02	50	4.540000	0.000000	9.000000	2.887058
MEASURE03	50	4.140000	0.000000	9.000000	2.725615
MEASURE04	50	5.520000	0.000000	9.000000	2.659139
MEASURE05	50	3.960000	0.000000	9.000000	2.633846
MEASURE06	50	4.8400			3.019393
MEASURE07	50	4.6600			2.495792
MEASURE08	50	3.7200			
MEASURE09	50	4.1600			
MEASURE10	50	3.9400			
MEASURE11	50	5.0400			
MEASURE12	50	4.3400			
MEASURE13	50	4.4400			
MEASURE14	50	4.0000			

If more than one variable is indicated by a highlighted section (i.e., when a block is selected), then the ***Graphs of Input Data*** menu will apply to the first selected variable.

When generating ***Graphs of Input Data***, *STATISTICA* takes into account the current case selection and weighting conditions for the variables that are being plotted. Note, however, that the case selection or weighting conditions need to be specified for the current spreadsheet (i.e., via the ***Tools*** tab ***Selection Conditions - Edit*** options and the ***Tools*** tab ***Weight*** options) and not just “locally” for an analysis (i.e., selected from the respective analysis/graph specification dialogs using the SELECT CASES and W buttons). The latter conditions will be ignored by the ***Graphs of Input Data***. For more information on specific types of ***Graphs of Input Data***, see the *Electronic Manual*.

GRAPHS OF BLOCK DATA

Unlike ***Graphs of Input Data***, ***Graphs of Block Data*** use the currently selected (continuous) block of data in the active spreadsheet to specify input data for the graph.



Correlations (Adstudy.STA)
Marked correlations are significant at p < .05000
N=50 (Casewise deletion of missing data)

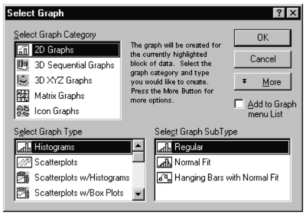
Variable	MEASURE05	MEASURE06	MEASURE07	MEASURE08	MEASURE09	MEASURE10
MEASURE07	-0.05	0.12	1.00	0.05	0.03	-0.08
MEASURE08	-0.19	-0.33	0.05	1.00	0.00	0.23
MEASURE09	-0.47	-0.27	0.03	0.00	1.00	0.08
MEASURE10	-0.07	0.13	-0.08	0.23	0.08	1.00
MEASURE11	-0.01	-0.15	-0.07			0.03
MEASURE12	0.11	0.04	0.09			0.15
MEASURE13	-0.03	0.00	0.31			
MEASURE14	0.17	0.03	-0.01			
MEASURE15	0.03	0.21	0.03			
MEASURE16	0.43	0.20	0.17			
MEASURE17	-0.03	0.27	0.04			
MEASURE18	0.16	-0.07	0.22			
MEASURE19	-0.02	-0.16	0.17			
MEASURE20	-0.08	0.12	-0.01			
MEASURE21	-0.08	0.04	-0.04			
MEASURE22	0.15	0.24	0.08			
MEASURE23	0.02	0.08	-0.00			

Note that these graphs are entirely independent from the concept of “input data.” They process values (numbers) from whatever is currently selected in the block and ignore the “meaning” of those numbers (e.g., the numbers can be raw data or values of correlation coefficients). These graphs offer an effective means of visualizing, exploring, and efficiently summarizing numeric output from analyses displayed in results spreadsheets (e.g., histograms of Monte Carlo output scores in the *SEPATH* module, or a box plot of aggregated means from a multivariate multiple classification table in the *ANOVA* module).

Although the most convenient way to select ***Graphs of Block Data*** is via the shortcut menu associated with the block selected in a spreadsheet, ***Graphs of Block Data*** are also available from the ***Graphs*** tab or the *STATISTICA* Start menu . When creating ***Graphs of Block Data***, you can select from default graphs (e.g., ***Histogram: Block Columns*** or ***Line Plot: Block Rows***), or you can create your own custom graphs for either the selected cells in the rows or columns, or of all cells in the selected rows or columns (i.e., going beyond the values that are selected in the block).

Default graphs. Using the default graphs (the first six commands on the ***Graphs of Block Data*** submenu, shown in the illustration above), you can create specified graphs with a single click. For specific information on each default graph, refer to the *Electronic Manual*.

Custom graphs. Select any of the four ***Custom Graph*** commands to display the ***Select Graph*** dialog, which provides a variety of options for creating customized graph.



The graph will be created for the currently highlighted block of data. Select the graph category and type you would like to create. Press the More Button for more options.

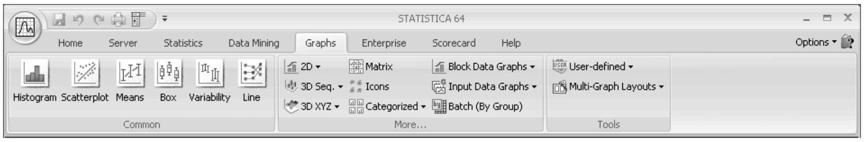
For specific information on custom graphs, refer to the *Electronic Manual*.

Customizing graphs. As with most features of *STATISTICA*, ***Graphs of Block Data*** are fully customizable. Select ***Customize List*** from the ***Block Data Graphs*** menu to display the ***Customize Graph Menu*** dialog, which provides options to remove, rename, or edit the currently listed graphs as well as to add new (user-defined) graphs to the ***Graphs of Block Data*** menu.

For example, if you want to include a normal fit on the histograms created using ***Histogram: Block Columns***, select ***Histogram: Block Columns*** in the ***Customize Graph Menu*** dialog, click the ***Edit*** button, and switch the ***Graph SubType*** to ***Normal Fit***. All subsequently created ***Histogram: Block Columns*** plots will include a normal fit to the data.

GRAPHS MENU GRAPHS

The ***Graphs*** tab provides a complete selection of all statistical graphs available in *STATISTICA*. These options are available from not only the ***Graphs*** tab, but also the *STATISTICA* Start menu , and offer hundreds of types of graphical representations and analytic summaries of data.



Note that, unlike ***Graphs of Block Data*** (which are also included on this tab in order to offer a full complement of all graphical options accessible from a single control), all other graph types from the ***Graphs*** tab are not limited to the values in the current output spreadsheet. Instead, they process data directly from the current input spreadsheet, in the same way the (previously discussed) ***Graphs of Input Data*** do. They represent either standard methods to graphically summarize raw data (e.g., various scatterplots, histograms, or plots of central tendencies such as medians) or standard graphical analytic techniques (e.g., categorized normal probability plots, detrended probability plots, or plots of confidence intervals of regression lines). When generating these graphs, *STATISTICA* takes into account the current case selection and weighting conditions for the variables selected to be plotted.

Graphs menu graphs include ***2D Graphs***, ***3D Sequential Graphs***, ***3D XYZ Graphs***, ***Matrix Plots***, ***Icon Plots***, ***Categorized Graphs***, and ***User-Defined Graphs***. Note that the ***Common*** group on the ***Graphs*** tab includes the most commonly used types of graphs (***Histograms***, ***Scatterplots***, ***Mean/Error Plots***, etc.), and the ***More*** group contains a comprehensive list of all graph types. See also, *Types of Graphs Menu Graphs* in the *Electronic Manual*.

GRAPH BRUSHING AND CASE STATES

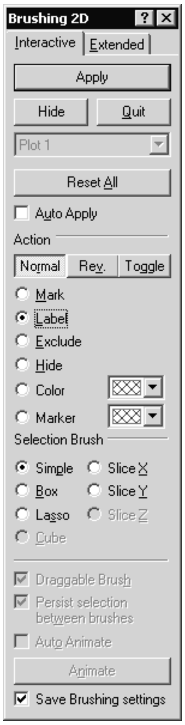
Graphs that are created from the ***Graphs*** tab are highly interactive with the spreadsheet from which they were created. You can identify and select points in the graph and specify that they are to be highlighted in the source spreadsheet, and vice versa.

In addition to selecting points in graphs and spreadsheets, you can identify properties of a case in a spreadsheet that will be used when the graph is created from that data. These properties include the point marker style and color, and whether the point is to be excluded from the graph and/or fit calculations.

To start brushing within a graph, click the brushing button on the ***Edit*** tab in the ***Customize Graph*** group, or right-click in the background of a graph and select ***Show Brushing*** from the shortcut menu to display the ***Brushing*** dialog, which is shown in the illustration to the right.

With the default ***Selection Brush***, which is ***Simple***, you can draw a rectangle on the graph to select the points contained in the rectangle. The following illustration demonstrates this for the example data set *Adstudy.sta*, with a 2D scatterplot of *MEASURE01* by *MEASURE02*.

Note that the upper-left three points have been selected by the brushing tool, which highlights the points in the graph as well as the corresponding cases in the spreadsheet from which the graph was created.



Brushing 2D

Interactive | Extended

Apply

Hide | Quit

Plot 1

Reset All

□ Auto Apply

Action

Normal | Rev. | Toggle

○ Mark
● Label
○ Exclude
○ Hide
○ Color
○ Marker

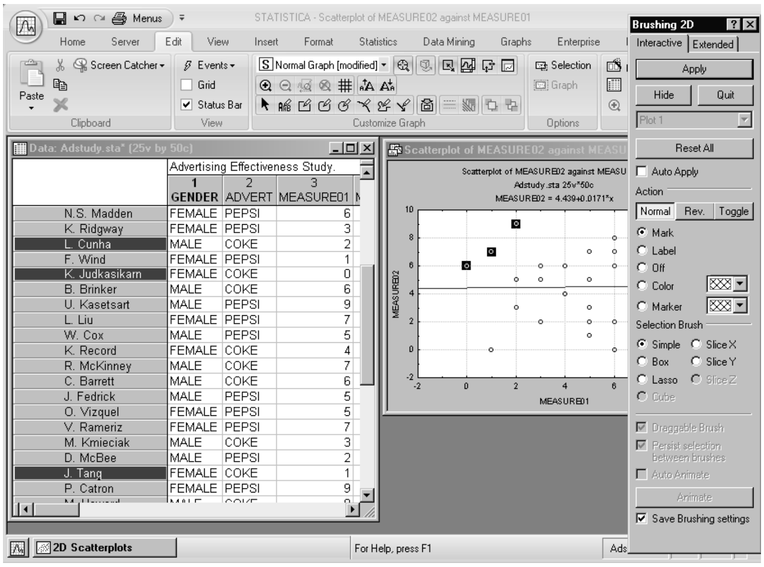
Selection Brush

● Simple ○ Slice X
○ Box ○ Slice Y
○ Lasso ○ Slice Z
○ Cube

☑ Draggable Brush
☑ Persist selection between brushes
□ Auto Animate

Animate

☑ Save Brushing settings



Data: Adstudy.sta* (25v by 50c)

Advertising Effectiveness Study.

	1 GENDER	2 ADVERT	3 MEASURE01
N.S. Madden	FEMALE	PEPSI	6
K. Ridgway	FEMALE	PEPSI	3
L. Cunha	MALE	COKE	2
F. Wind	FEMALE	PEPSI	1
K. Judkasikarn	FEMALE	COKE	0
B. Brinker	MALE	COKE	6
U. Kasetsart	MALE	PEPSI	9
L. Liu	FEMALE	PEPSI	7
W. Cox	MALE	PEPSI	5
K. Record	FEMALE	COKE	4
R. McKinney	MALE	COKE	7
C. Barrett	MALE	COKE	6
J. Fedrick	MALE	PEPSI	5
O. Vizquel	FEMALE	PEPSI	5
V. Rameriz	FEMALE	PEPSI	7
M. Kmieciak	MALE	COKE	3
D. McBee	MALE	PEPSI	2
J. Tang	FEMALE	COKE	1
P. Catron	FEMALE	PEPSI	9

Brushing 2D

Interactive | Extended

Apply

Hide | Quit

Plot 1

Reset All

☐ Auto Apply

Action

Normal | Rev. | Toggle

● Mark
○ Label
○ Off
○ Color ▨
○ Marker ▨

Selection Brush

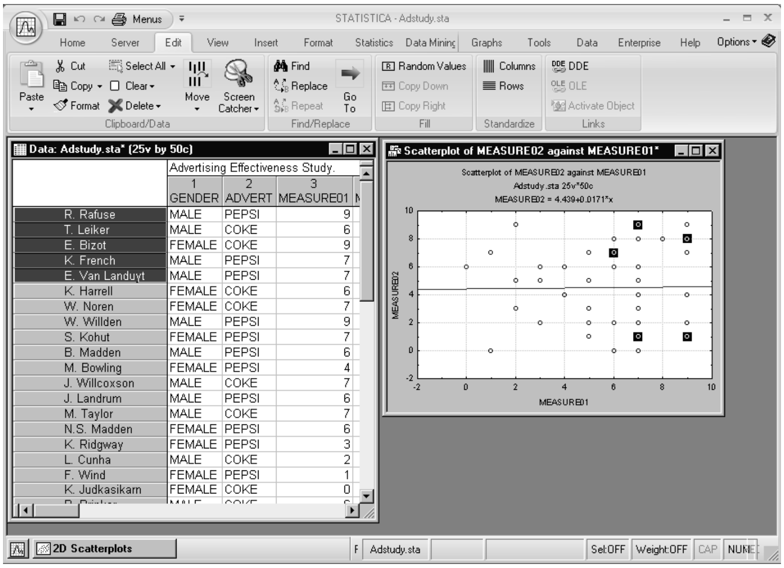
● Simple ○ Slice X
○ Box ○ Slice Y
○ Lasso ○ Slice Z
○ Cube

☑ Draggable Brush
☑ Persist selection between brushes
☐ Auto Animate

Animate

☑ Save Brushing settings

Alternatively, instead of using the ***Brushing*** facilities, you can select cases in the spreadsheet (click on the far-left side of the case name) and the corresponding points will be marked in the graph, as shown in the following illustration, where the first five cases in the *Adstudy.sta* spreadsheet have been selected.

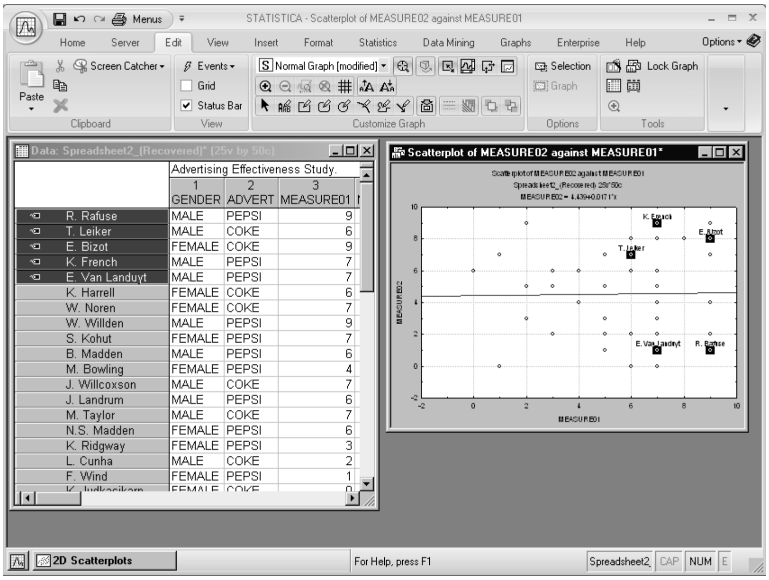


Data: Adstudy.sta* (25v by 50c)

Advertising Effectiveness Study.

	1 GENDER	2 ADVERT	3 MEASURE01
R. Rafuse	MALE	PEPSI	9
T. Leiker	MALE	COKE	6
E. Bizot	FEMALE	COKE	9
K. French	MALE	PEPSI	7
E. Van Landuyt	MALE	PEPSI	7
K. Harrell	FEMALE	COKE	6
W. Noren	FEMALE	COKE	7
W. Willden	MALE	PEPSI	9
S. Kohut	FEMALE	PEPSI	7
B. Madden	MALE	PEPSI	6
M. Bowling	FEMALE	PEPSI	4
J. Willcoxson	MALE	COKE	7
J. Landrum	MALE	PEPSI	6
M. Taylor	MALE	COKE	7
N.S. Madden	FEMALE	PEPSI	6
K. Ridgway	FEMALE	PEPSI	3
L. Cunha	MALE	COKE	2
F. Wind	FEMALE	PEPSI	1
K. Judkasikarn	FEMALE	COKE	0

You can specify spreadsheet case states from either a spreadsheet or a graph. In a *STATISTICA* Spreadsheet, right-click on a case name to display the shortcut menu, which contains commands including ***Off***, ***Label***, ***Marked Points***, and ***Case State***s. Similar commands are available from the shortcut menu displayed when you right-click on the points in a graph. The graph will use these options when displaying the points represented by this case. For example, if you select ***Label***, the corresponding points will be labeled, as shown in the next illustration. Note that the spreadsheet cases are marked with a case state icon to indicate that the case points are labeled.



Advertising Effectiveness Study.

	1 GENDER	2 ADVERT	3 MEASURE01
R. Rafuse	MALE	PEPSI	9
T. Leiker	MALE	COKE	6
E. Bizot	FEMALE	COKE	9
K. French	MALE	PEPSI	7
E. Van Landuyt	MALE	PEPSI	7
K. Harrell	FEMALE	COKE	6
W. Noren	FEMALE	COKE	7
W. Willden	MALE	PEPSI	9
S. Kohut	FEMALE	PEPSI	7
B. Madden	MALE	PEPSI	6
M. Bowling	FEMALE	PEPSI	4
J. Willcoxson	MALE	COKE	7
J. Landrum	MALE	PEPSI	6
M. Taylor	MALE	COKE	7
N.S. Madden	FEMALE	PEPSI	6
K. Ridgway	FEMALE	PEPSI	3
L. Cunha	MALE	COKE	2
F. Wind	FEMALE	PEPSI	1
K. Judkasikarn	FEMALE	COKE	0

Right-click on a case name, and from the shortcut menu select ***Case States - Edit Case States*** to change the case marker and/or color.

Note that the selection of points is available for graph types other than Scatterplots. For histograms, brushing/selecting a histogram bar will select the corresponding points to that bar in the spreadsheet. The same is true of the boxes in box plots.

Using case states and brushing and selecting points is particularly useful with the ***Hidden*** and ***Excluded*** case states options. First, to make these options available, display the ***Options*** dialog (select the ***Tools*** tab and click ***Options***), and in the tree view select ***Navigation / Defaults*** (located under ***Spreadsheets***). Clear the

Combine Excluded and Hidden Case States into Off state check box, and click the ***OK*** button.

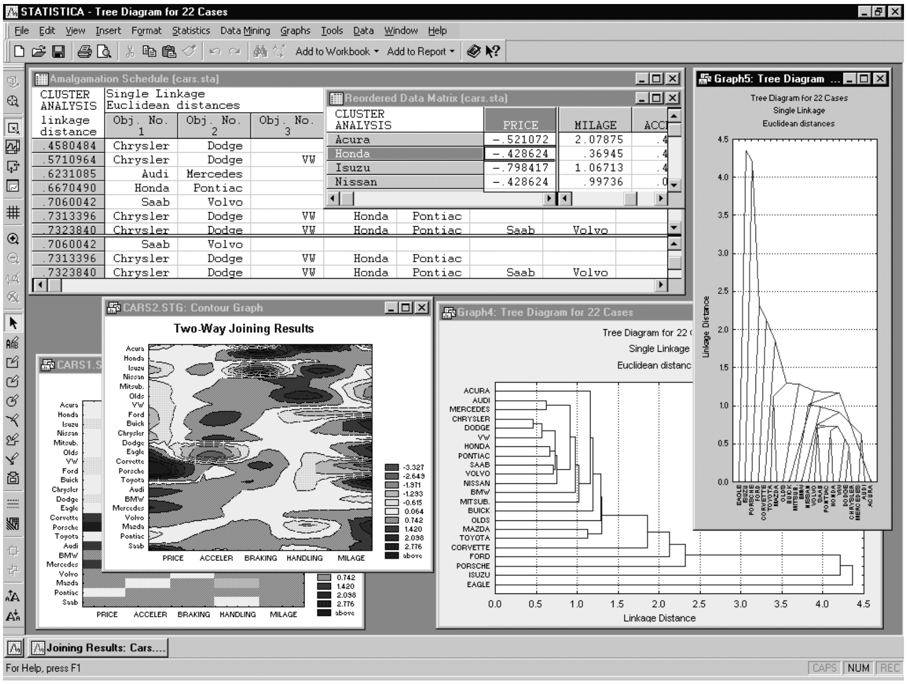
Then, select the ***Data*** tab, and in the ***Cases*** group click ***Cases***. From the ***Case States*** submenu, select ***Hidden*** to mark a case as hidden, i.e., the case will not be visible in graphs, but will be used in analyses. You can also right-click on a case name, and from the shortcut menu select ***Case States - Edit Case States*** to display the ***Case State*** dialog, where you can select the ***Hidden*** check box.

Select ***Excluded*** to mark a case as excluded, i.e., the case will not be used in the computations; however, the case will be displayed in most graph types. The case point marker is displayed, but the case is removed from computations. The ***Excluded*** case state also works in conjunction with spreadsheet selection conditions; any case that has the ***Excluded*** case state set will be treated as if the case were excluded by selection conditions. Therefore, using graph brushing and case states is a convenient tool to interactively remove outliers and then rerun analyses with the points removed.

When the ***Combine Excluded and Hidden Case States into Off state*** check box is selected in the ***Options*** dialog - ***Navigation / Defaults*** options pane, the ***Hidden*** and ***Excluded*** options are replaced with the ***Off*** option. Select ***Off*** to mark a case as ***Hidden*** and ***Excluded***; the point will be excluded from computations and from graphs.

OTHER SPECIALIZED GRAPHS

In addition to the standard selection of ***Graphs of Input Data***, ***Graphs of Block Data***, and ***Graphs*** menu graphs, other specialized statistical graphs that are related to a type of analysis (e.g., cluster analysis results) are accessible directly from results dialogs (i.e., the dialogs that contain output options from the current analysis).


STATISTICA - Tree Diagram for 22 Cases
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
Add to Workbook Add to Report
Amalgamation Schedule (cars.sta)
Reordered Data Matrix (cars.sta)
Graph5: Tree Diagram
CARS2.STG: Contour Graph
Two-Way Joining Results
PRICE ACCELER BRAKING HANDLING MILAGE
CARS1.S
Graph4: Tree Diagram for 22 Cases
Tree Diagram for 22 Cases
Single Linkage
Euclidean distances
Linkage Distance
Joining Results: Cars....
For Help, press F1
CAPS NUM REC

CLUSTER ANALYSIS linkage distance	Single Linkage Euclidean distances Obj. No. 1	Obj. No. 2	Obj. No. 3					
.4580484	Chrysler	Dodge						
.5710964	Chrysler	Dodge	VW					
.6231085	Audi	Mercedes						
.6670490	Honda	Pontiac						
.7060042	Saab	Volvo						
.7313396	Chrysler	Dodge	VW	Honda	Pontiac			
.7323840	Chrysler	Dodge	VW	Honda	Pontiac	Saab	Volvo	
.7060042	Saab	Volvo						
.7313396	Chrysler	Dodge	VW	Honda	Pontiac			
.7323840	Chrysler	Dodge	VW	Honda	Pontiac	Saab	Volvo	

CLUSTER ANALYSIS	PRICE	MILAGE	ACC
Acura	-.521072	2.07875	.4
Honda	-.428624	.36945	.4
Isuzu	-.798417	1.06713	.4
Nissan	-.428624	.99736	.0

The specialized graphs are described in the documentation for the analyses from which they can be produced; for information, refer to the *Electronic Manual*.

CREATING GRAPHS VIA *STATISTICA* VISUAL BASIC

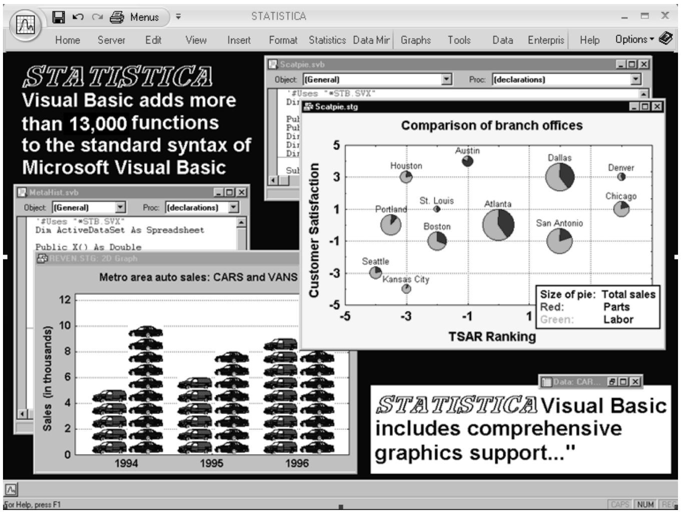
STATISTICA graphical options can also be accessed programmatically using the built-in *STATISTICA* Visual Basic (SVB) or other compatible languages. Therefore, there are no limits to how "deeply customized" your *STATISTICA* graphs can be, because SVB (with all its powerful custom drawing tools as well as the *STATISTICA*-based library of graphics procedures) can be used to produce virtually any graphics or multimedia output supported by the contemporary computer hardware.

An application written in *STATISTICA* Visual Basic can operate on graphs in three ways:

- Create a new graph and then modify, print, or save it;
- Access an existing graph and then modify it;

- Open an existing graph file and then modify, print, or save it.

Every graph available in *STATISTICA* can be produced by *STATISTICA* Visual Basic and then customized using *STATISTICA* procedures or general options offered in this comprehensive language.



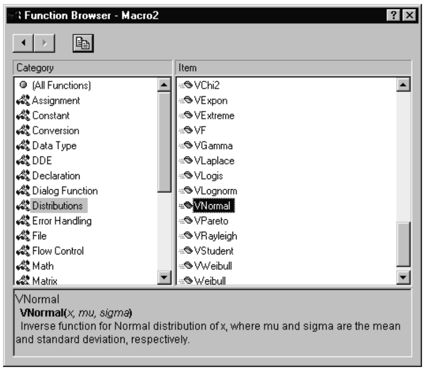
STATISTICA Visual Basic adds more than 13,000 functions to the standard syntax of Microsoft Visual Basic

```
'#Uses "*STB.SVX"
Dim ActiveDataSet As Spreadsheet

Public X() As Double
```

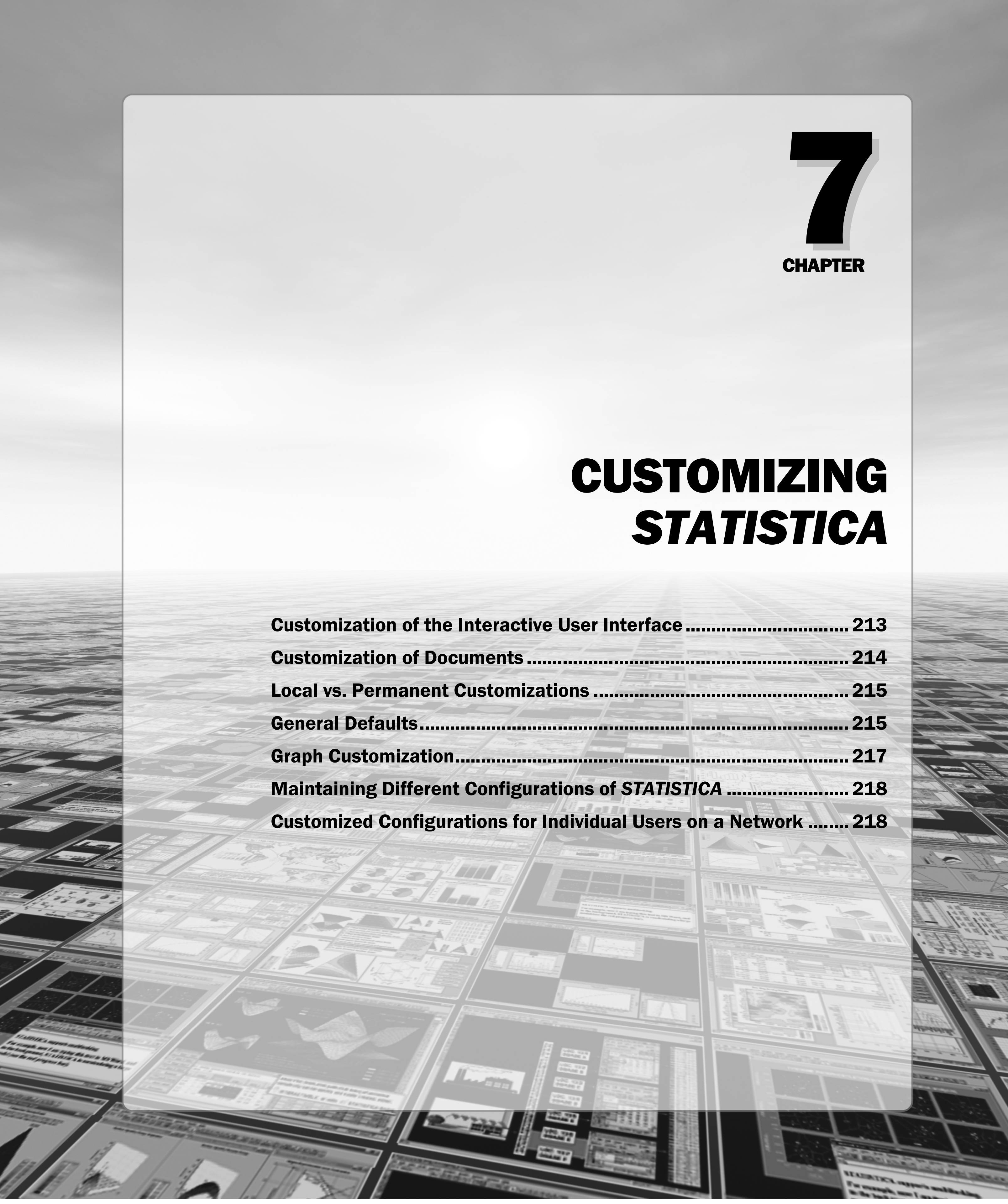
STATISTICA Visual Basic includes comprehensive graphics support..."

As with all other functions in *STATISTICA* Visual Basic, functions to access the graphics library of *STATISTICA* can be easily incorporated into *STATISTICA* Visual Basic programs via a hierarchically organized ***Function Browser***. It contains short descriptions of all functions and options that can be inserted directly into the source code of your program (i.e., into the *STATISTICA* Visual Basic Editor, see page 225).



VNormal
VNormal(*x, mu, sigma***)**
Inverse function for Normal distribution of x, where mu and sigma are the mean and standard deviation, respectively.

For more information on accessing the graphics libraries of *STATISTICA* via the *STATISTICA* Visual Basic programming language, refer to the *Electronic Manual*.



7

CHAPTER

CUSTOMIZING *STATISTICA*

7

CHAPTER

CUSTOMIZING *STATISTICA*

STATISTICA offers the flexibility of fully customizable user interfaces and supports the necessary adjustment of the standard user interface to better suit your specific needs. In fact, *STATISTICA* “anticipates” your needs in that it remembers various choices as you make them. For example, if you launch an analysis from the ***Advanced*** tab on an analysis specification dialog, the ***Advanced*** tab will be selected (instead of the ***Quick*** tab) the next time you display that dialog.

Practically all aspects of the user interface can be customized starting with such elementary controls as the classic menus, Quick Access toolbar, and the keyboard. The process for customizing these screen components is quick and straightforward (for example, see the illustration of customizing the toolbar on page 139). You can set both global and local customizations for graphs, spreadsheets, workbooks, reports, etc., and maintain different configurations of *STATISTICA* (for a single user as well as for network users). You can also define entirely new user interfaces (see pages 139 and 140).

CUSTOMIZATION OF THE INTERACTIVE USER INTERFACE

As mentioned before, *STATISTICA* contains facilities to define entirely new user interfaces (see page 139), including the Internet browser-based user interfaces (see page 141). However, practically all aspects of the default, interactive user interface can also be adjusted easily in a variety of ways. For example, you can add to the default options, simplify them, or keep changing them as your needs change. Depending on the requirements of the tasks to be performed, as well as your personal preferences for particular “modes” of work (and aesthetic choices), you can suppress all icons, toolbars, status bars, long menus, workbook facilities, drag-and-drop facilities, dynamic (automatic) links between graphs and data, 3D effects in tables, and 3D effects in dialog boxes; request “bare-bones” sequential output with simple, paper-white spreadsheets and monochrome graphs; and set the system to automatically maintain no more than one simple report at a time.

Or alternatively, you can define elaborate local and global toolbars; take full advantage of all special tools and controls, icons, toolbars, macros (e.g., assign particular tasks to specific new classic menu commands, the Quick Access toolbar, or keys), elaborate multimedia tables, workbook facilities, and drag-and-drop facilities; establish multiple dynamic (automatic) links between graphs and data and internal links between graphical objects; customize the output windows with colors, special fonts, and highlights; adjust the default graph styles and their display modes; and send the results to separate hierarchically organized workbooks to create an elaborate, “multi-layered” data analysis environment that facilitates the exploration of complex data files and allows you to compare different aspects of the output.

CUSTOMIZATION OF DOCUMENTS

There is a variety of comprehensive, specialized tools to customize the layout and operation of *STATISTICA* documents (see Chapter 5 – *STATISTICA Documents*, page 167). For example, *STATISTICA* has a comprehensive system of managing defaults of every aspect of graphs and combining customizations into hierarchically organized “styles.” Similarly, you can create custom layouts and formats for spreadsheets (multimedia tables) and even customize events (e.g., what happens when you double-click on a table). See the *Electronic Manual* for further details.

LOCAL VS. PERMANENT CUSTOMIZATIONS

Many aspects of the appearance of *STATISTICA* can be adjusted from both the ***View*** and ***Tools*** tabs. Each of these two methods, however, has a different function.

View tab. The changes specified on the ***View*** tab affect the current appearance of *STATISTICA* (e.g., hides the Status Bar) or the current document window (e.g., spreadsheet grid lines).

Options dialog. The options available in the ***Options*** dialog (select the ***Tools*** tab and click ***Options***) are used to adjust the permanent program defaults (discussed in more detail in the next section). Note, however, that the global options that are applicable to documents of a particular type (e.g., a graph or a spreadsheet) will not change the current document. Instead, they will only be stored as program defaults that will affect the creation of the next (i.e., new) document of the respective type.

For example, if you change the ***Default Spreadsheet Layout*** in the ***Navigation / Defaults*** options pane of the ***Options*** dialog, you will see the new Spreadsheet Layout applied only when you create a new spreadsheet. However, these defaults will not affect any previously saved files because those spreadsheets are displayed with the specific appearance with which they were saved (use the options on the ***View*** tab to customize the existing objects).

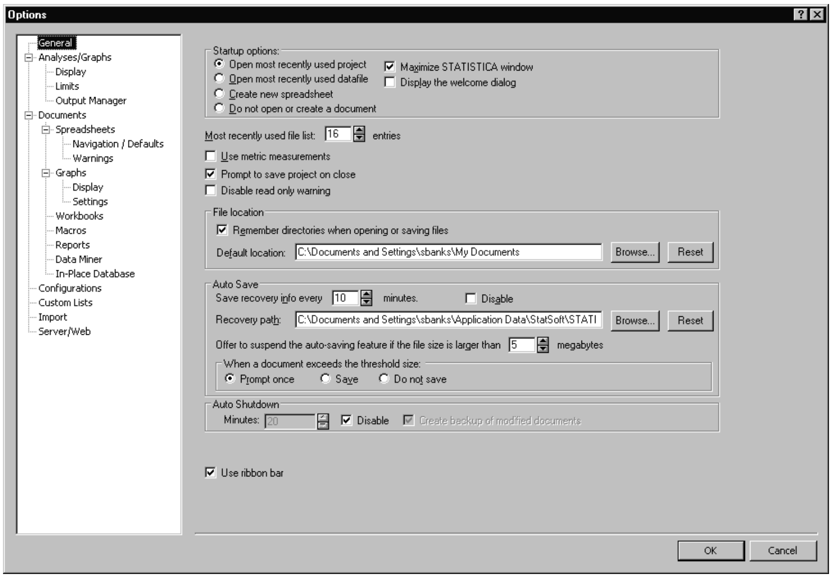
GENERAL DEFAULTS

Customization of the general system defaults. The general default settings of *STATISTICA* can be adjusted with the options in the ***Options*** dialog (select the ***Tools*** tab and click ***Options***). They control:

- The general aspects of the behavior of *STATISTICA* (such as maximizing *STATISTICA* on startup, workbook and report facilities, file locations, custom lists, etc.),

- The way in which the output is produced (e.g., in workbooks, reports, etc.),
- The general appearance of the application window (icons, toolbars, etc.), and
- The appearance of document windows.

The ***General*** options pane of the ***Options*** dialog is shown in the next illustration.



Startup options:
● Open most recently used project
○ Open most recently used datafile
○ Create new spreadsheet
○ Do not open or create a document
☑ Maximize STATISTICA window
☐ Display the welcome dialog

Most recently used file list: 16 entries
☐ Use metric measurements
☑ Prompt to save project on close
☐ Disable read only warning

File location
☑ Remember directories when opening or saving files
Default location: C:\Documents and Settings\sbanks\My Documents

Auto Save
Save recovery info every 10 minutes. ☐ Disable
Recovery path: C:\Documents and Settings\sbanks\Application Data\StatSoft\STATI
Offer to suspend the auto-saving feature if the file size is larger than 5 megabytes
When a document exceeds the threshold size:
● Prompt once ○ Save ○ Do not save

Auto Shutdown
Minutes: 20 ☑ Disable ☑ Create backup of modified documents

☑ Use ribbon bar

All these and other general settings are accessible regardless of the type of document that is currently active (e.g., a spreadsheet or a graph). For more information about a specific options pane, see the *Electronic Manual* (i.e., press F1 to view the *STATISTICA* Help topic describing the options currently displayed).

Switching between alternative sets of defaults (configurations). Options are provided in the ***Configurations*** options pane of the ***Options*** dialog that enable you to maintain “libraries” of settings and switch between them for different projects (or users). For further details, see *Maintaining Different Configurations of STATISTICA* on page 218 and in the *Electronic Manual*.

GRAPH CUSTOMIZATION

Interactive graph customization. The customization options in *STATISTICA* graphics include hundreds of features and tools that can be used to adjust every detail of the display and associated data processing. These options are arranged in a hierarchical manner, so those used most often are accessible directly via shortcuts by double-clicking or right-clicking on a specific element of the graph.

Permanent settings and automation options. The initial (default) settings of all graph features can be easily adjusted so that even the default appearance and behavior of *STATISTICA* Graphs will match your specific needs and/or will require very little intervention on your part. Various aspects of *STATISTICA* Graphs can be permanently adjusted by using:

1. the ***Options*** dialog (select the ***Tools*** tab and click ***Options***),
2. the comprehensive system of graph styles,
3. user-defined graphs, and
4. *STATISTICA* Visual Basic.

These facilities are briefly reviewed in Chapter 6 – *Graphs* (page 190). For more information, please refer to the *Electronic Manual*.

There are no limits to how "deeply customized" your *STATISTICA* custom graphs can be, because *STATISTICA* Visual Basic (with all its powerful custom drawing tools as well as the *STATISTICA*-based library of graphics procedures) can be used to produce virtually any graphics or multimedia output supported by contemporary computer hardware. Those custom developed displays or multimedia output can be assigned to *STATISTICA* toolbars, menus, or dialogs and become a permanent part of "your" *STATISTICA* application.

MAINTAINING DIFFERENT CONFIGURATIONS OF *STATISTICA*

STATISTICA stores all program settings when you exit the program, and restores them the next time you start the application. You can create different configurations of these settings by using the options in the ***Configurations*** options pane of the ***Options*** dialog (select the ***Tools*** tab and click ***Options***). With the configuration manager, you can save the current program state into a new or existing configuration, or you can restart *STATISTICA* using a different configuration. Other options include the ability to import or export configurations to a separate file so they can be shared among *STATISTICA* installations.

CUSTOMIZED CONFIGURATIONS FOR INDIVIDUAL USERS ON A NETWORK

The same principle described in the previous paragraph applies to network installations of *STATISTICA*. On a network, *STATISTICA* is installed in only one location (on a server), but each user can still configure *STATISTICA* differently because the setting configuration information is stored locally. Note that you need to choose ***Network Installation*** in the ***STATISTICA Setup*** program in order to install it properly on a non-local drive (network server). Note that a network version of *STATISTICA* is necessary to ensure its reliable operation when used by more than one user at a time or even one user if *STATISTICA* is not installed on the local system.

8

CHAPTER

STATISTICA VISUAL BASIC

8

CHAPTER

STATISTICA VISUAL BASIC

The *STATISTICA* Visual Basic (SVB) language (integrated into *STATISTICA*) is compatible with the industry standards and provides another user interface to the functionality of *STATISTICA*, and it offers incomparably more than just a “supplementary application programming language” that can be used to write custom extensions.

Note that *STATISTICA* Visual Basic is not Microsoft Visual Basic 6.0. StatSoft owns and maintains the code for *STATISTICA* Visual Basic. SVB is compatible with Microsoft’s VB.NET, Microsoft’s Visual Basic for Applications (VBA), and also with Microsoft’s Visual Basic 6.0 (VB6). SVB scripting language is unique in terms of its flexibility and compatibility, and it is also very powerful. It provides access to Visual Basic for Applications (used for scripting Microsoft Office products) and access to the .NET Framework within the same file (see Chapter 10 – Programming *STATISITCA* from .NET, page 247). Other APIs can also be accessed and leverage the flexibility of SVB such as, for example, Yahoo’s Stock Quote API or Google Analytics API. SVB offers a powerful 64-bit solution for system integration, expansion, and custom development.

SVB takes full advantage of the object model architecture of *STATISTICA* and is used to access programmatically every aspect and virtually every detail of the functionality of *STATISTICA*. Even the most complex analyses and graphs can be recorded into Visual Basic macros and later be run repeatedly or edited and used as building blocks of other applications. SVB adds an arsenal of more than 14,000 new functions to the standard comprehensive syntax of Visual Basic, thus comprising one of the largest and richest development environments available.

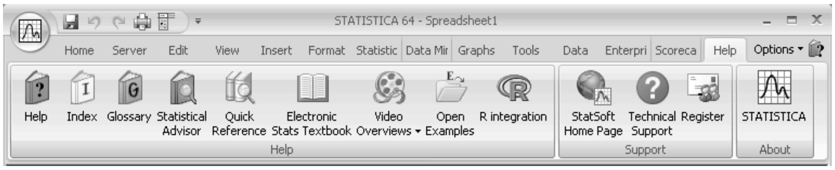
Applications for *STATISTICA* Visual Basic programs. *STATISTICA* Visual Basic programs can be used for a wide variety of applications, from simple macros recorded to automate a specific (repeatedly used) sequence of tasks, to elaborate custom analytic systems combining the power of optimized procedures of *STATISTICA* with custom developed extensions featuring their own user interface. When properly licensed, scripts for analyses developed this way can be integrated into larger computing environments or executed from within proprietary corporate software systems or Internet or intranet portals.

SVB programs can also be attached to virtually all important "events" in a *STATISTICA* analysis such as opening or closing files, clicking on cells in spreadsheets, etc.; in this manner, the basic user interface of *STATISTICA* can be highly customized for specific applications (e.g., for data entry operations, etc.).

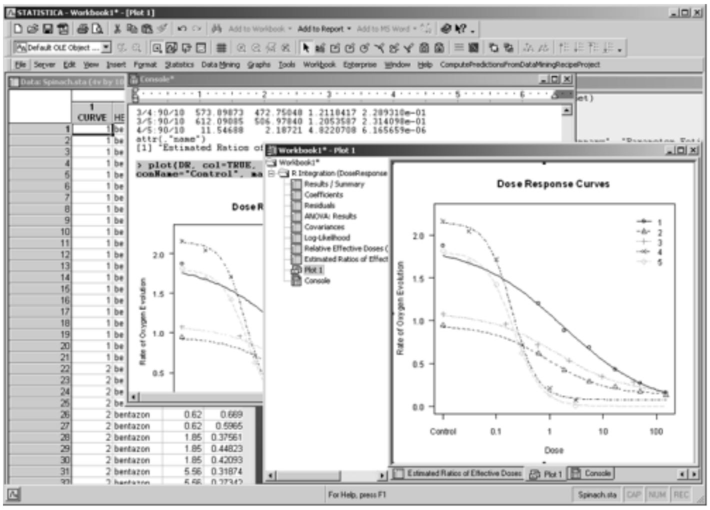
Several scripting languages are included in *STATISTICA*. You can select from SVB, Enhanced SVB, *STATISTICA* Visual Basic.NET, or R.

Enhanced *STATISTICA* Visual Basic is a superset of *STATISTICA* Visual Basic, and includes additional features. *STATISTICA* Visual Basic.NET features direct, native access to .NET Assemblies, i.e., not through COM Interop as would be required from standard SVB.

R is a programming language and environment for statistical computing. The R environment and its source code are freely available under the GNU GPL license. The R community maintains several centralized repositories that make hundreds of such packages readily available to all users over the Internet. Native R scripts can be run directly within *STATISTICA*, *STATISTICA Enterprise*, and *STATISTICA Enterprise Server*.



R output can be retrieved as native *STATISTICA* Spreadsheets and Graphs, and managed via highly flexible *STATISTICA* Workbook containers.



```
3/4:90/10   573.09873   472.75048 1.2118417 2.289310e-01
3/5:90/10   612.09085   506.97840 1.2053587 2.314098e-01
4/5:90/10    11.54688     2.18721 4.8220708 6.165659e-06
attr(,"name")
[1] "Estimated Ratios of

> plot(DR, col=TRUE,
conName="Control", ma
```

	1 CURVE	HE		
1	1	be		
2	1	be		
3	1	be		
4	1	be		
5	1	be		
6	1	be		
7	1	be		
8	1	be		
9	1	be		
10	1	be		
11	1	be		
12	1	be		
13	1	be		
14	1	be		
15	1	be		
16	1	be		
17	1	be		
18	1	be		
19	1	be		
20	1	be		
21	1	be		
22	2	be		
23	2	be		
24	2	be		
25	2	be		
26	2	bentazon	0.62	0.669
27	2	bentazon	0.62	0.5965
28	2	bentazon	1.85	0.37561
29	2	bentazon	1.85	0.44823
30	2	bentazon	1.85	0.42093
31	2	bentazon	5.56	0.31874
32	2	bentazon	5.56	0.27342

Using the R language requires that you have R installed on either the same computer running *STATISTICA* or a computer accessible from the *STATISTICA Enterprise Server* in order to use its specialized routines and capabilities to:

- Add new R-based “modules”
- Leverage *STATISTICA’s* superior graphics, flexible spreadsheets, and convenient workbook containers for various document types to handle output from R
- Integrate R into *STATISTICA Enterprise* to make specialized R functionality available as reusable analysis templates for users not familiar with the R language, in a secure, role-based enterprise analysis system
- Add R-based analytic nodes to *STATISTICA Data Miner*, thus leveraging all R capabilities inside *STATISTICA* and *Data Miner* workspaces
- Build scalable R servers using *STATISTICA Enterprise Server* to handle security and load balancing, and to take advantage of multiple processor servers to run R for demanding and/or validated enterprise applications

See the *Electronic Manual* for more information on these scripting languages.

RECORDING *STATISTICA* VISUAL BASIC (SVB) MACROS (PROGRAMS)

Analysis Macros, Master (Log) Macros, and Keyboard Macros

STATISTICA provides a comprehensive selection of facilities for recording macros, i.e., *STATISTICA* Visual Basic (SVB) programs, to automate repetitive work or to be used as a means to automatically generate programs for further editing and modification. The macro programs recorded by these facilities can be saved to be run "as is," or they can be used as the "building blocks" for more complex and highly customized Visual Basic application programs. Analysis Macros and Master Macros follow the identical syntax and can later be modified, but because of the different ways in which each of them is created, they offer distinctive advantages and disadvantages for specific applications.

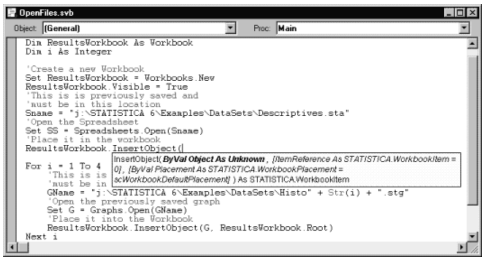
Analysis macros. Simple Analysis Macros automatically record the settings, selections, and chosen options for a specific analysis. Note that the term "analysis" in *STATISTICA* denotes one task selected either from the ***Statistics***, ***Data Mining***, or ***Graphs*** tabs and can be very small and simple (e.g., one scatterplot requested from the ***Graphs*** tab), or very elaborate (e.g., a complex structural equation modeling analysis selected by choosing that option from the ***Statistics*** tab, and involving hundreds of output documents). After selecting any of the statistical options from the ***Statistics*** or ***Data Mining*** tabs or graphics options from the ***Graphs*** tab, all actions such as variable selections, option settings, etc., are recorded "behind the scenes"; at any time you can transfer this recording (i.e., the Visual Basic code for that macro) to the Visual Basic Editor window. The ***Create Macro*** command is available from every analysis dialog via the drop-down menu displayed by clicking the ***Options*** button or the shortcut menu accessed by right-clicking the analysis button when the analysis is minimized.

Master macros (logs). You can record a Master Macro or Master Log of an entire session, which can consist of one or many analyses. This recording will "connect" analyses performed with various analysis options from the ***Statistics, Data Mining***, and/or ***Graphs*** tabs. However, unlike simple Analysis Macros, you can turn the recording of Master Macros on and off. The Master Macro recording will begin when you turn on the recording [select the ***Tools*** tab, click ***Macro***, and select ***Start Recording Log of Analyses (Master Macro)***], and it will end when you stop the recording (click ***Macro***, and select ***Stop Recording***). In between these actions, all file selections and data management operations are recorded, as are the analyses and selections for the analyses, in the sequence in which they were chosen.

Keyboard macros. This type of macro recording stores the sequences of keyboard input. When you select the ***Tools*** tab, click ***Macro***, and select ***Start Recording Keyboard Macro***, *STATISTICA* will record the actual keystrokes entered via the keyboard. When you ***Stop Recording***, a *STATISTICA* Visual Basic editor window opens with a simple program containing a single `SendKeys` command with symbols that represent all the different keystrokes performed during the recording session. Note that this type of macro is very simple in the sense that it will not record any context in which the recorded keystrokes are pressed and will not record their meaning (i.e., commands these keystrokes trigger), but this feature makes them useful for specific applications, e.g., to automate entering text, such as titles, selection conditions, etc.

***STATISTICA* Visual Basic editor and debugger.** Programs can be written from scratch using the *STATISTICA* Visual Basic professional development environment, which features a program editor with a powerful debugger (with breakpoints, etc.) and many facilities that aid in efficient code building. These facilities are described in detail in the *STATISTICA Electronic Manual*.

When editing macro programs by typing in Visual Basic commands or program commands specific to SVB, the editor displays type-ahead help to illustrate the appropriate syntax. Help on the members and functions for each class (object) is also provided in-line.



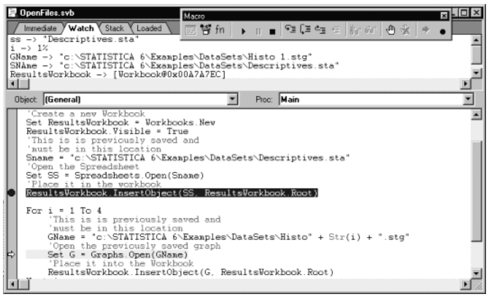
```
Dim ResultsWorkbook As Workbook
Dim i As Integer

'Create a new Workbook
Set ResultsWorkbook = Workbooks.New
ResultsWorkbook.Visible = True
'This is is previously saved and
'must be in this location
Sname = "j:\STATISTICA 6\Examples\DataSets\Descriptives.sta"
'Open the Spreadsheet
Set SS = Spreadsheets.Open(Sname)
'Place it in the workbook
ResultsWorkbook.InsertObject(

For i = 1 To 4
    'This is is
    'must be in
    GName = "j:\STATISTICA 6\Examples\DataSets\Histo" + Str(i) + ".stg"
    'Open the previously saved graph
    Set G = Graphs.Open(GName)
    'Place it into the Workbook
    ResultsWorkbook.InsertObject(G, ResultsWorkbook.Root)
Next i
```

InsertObject(**ByVal Object As Unknown** , *[ItemReference As STATISTICA.WorkbookItem = 0]* , *[ByVal Placement As STATISTICA.WorkbookPlacement = scWorkbookDefaultPlacement]*) As STATISTICA.WorkbookItem

When executing a program, you can set breakpoints in the program, step through it line by line, and observe and change the values of variables in the macro program as it is running.



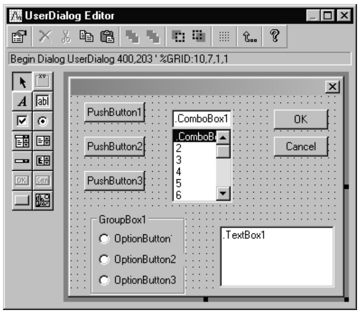
```
ss -> "Descriptives.sta"
i -> 1%
GName -> "c:\STATISTICA 6\Examples\DataSets\Histo 1.stg"
SNAme -> "c:\STATISTICA 6\Examples\DataSets\Descriptives.sta"
ResultsWorkbook -> [Workbook@0x00A7A7EC]
```

Object: (General) Proc: Main

```
'Create a new Workbook
Set ResultsWorkbook = Workbooks.New
ResultsWorkbook.Visible = True
'This is is previously saved and
'must be in this location
Sname = "c:\STATISTICA 6\Examples\DataSets\Descriptives.sta"
'Open the Spreadsheet
Set SS = Spreadsheets.Open(Sname)
'Place it in the workbook
ResultsWorkbook.InsertObject(SS, ResultsWorkbook.Root)

For i = 1 To 4
    'This is is previously saved and
    'must be in this location
    GName = "c:\STATISTICA 6\Examples\DataSets\Histo" + Str(i) + ".stg"
    'Open the previously saved graph
    Set G = Graphs.Open(GName)
    'Place it into the Workbook
    ResultsWorkbook.InsertObject(G, ResultsWorkbook.Root)
```

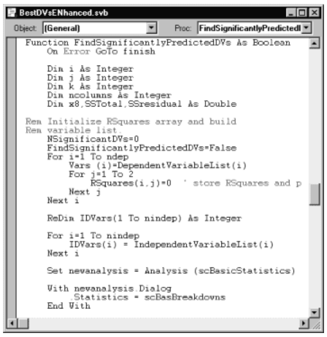
Also available is an interactive dialog editor that enables you to build dialog boxes.



Begin Dialog UserDialog 400,203 ' %GRID:10,7,1,1

To summarize, *STATISTICA* Visual Basic is not only a powerful programming language, but it represents a very powerful, professional programming environment for developing simple macros as well as complex custom applications.

Visual Basic from other applications. SVB programs can also be developed by enhancing Visual Basic programs created in other applications (e.g., Excel) by calling *STATISTICA* functions and procedures.



```
Function FindSignificantlyPredictedDVs As Boolean
    On Error GoTo finish

    Dim i As Integer
    Dim j As Integer
    Dim k As Integer
    Dim ncolumns As Integer
    Dim x8,SSTotal,SSresidual As Double

Rem Initialize RSquares array and build
Rem variable list.
    NSignificantDVs=0
    FindSignificantlyPredictedDVs=False
    For i=1 To ndep
        Vars (i)=DependentVariableList(i)
        For j=1 To 2
            RSquares(i,j)=0  ' store RSquares and p
        Next j
    Next i

    ReDim IDVars(1 To nindep) As Integer

    For i=1 To nindep
        IDVars(i) = IndependentVariableList(i)
    Next i

    Set newanalysis = Analysis (scBasicStatistics)

    With newanalysis.Dialog
        .Statistics = scBasBreakdowns
    End With
```

Executing *STATISTICA* Visual Basic Programs

STATISTICA Visual Basic programs can be executed from within *STATISTICA,* but because of the industry standard compatibility of SVB, you can also execute its programs from any other compatible environment (e.g., Excel, Word, or a stand-alone Visual Basic language). In practice, you would typically call *STATISTICA* functions from Visual Basic in another application. Note, however, that when you run an SVB program or attempt to call *STATISTICA* functions from any other application, all calls to the *STATISTICA* specific functions (as opposed to the generic functions of MS Visual Basic) will be executed only if the respective *STATISTICA* libraries are present on the computer where the execution takes place. That is, you must be a licensed user of the respective *STATISTICA* libraries of procedures. Note that this large library of *STATISTICA* functions (more than 14,000 procedures) is transparently accessible not only to Visual Basic, but also to calls from any other compatible programming language or environment, such as C/C++, C#, or Delphi.

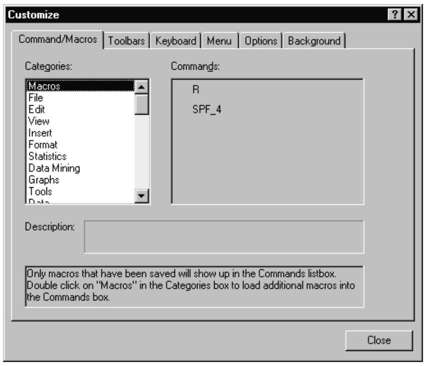
Performance of *STATISTICA* Visual Basic programs. While the obvious advantages of Visual Basic (compared to other languages) are its ease of use and familiarity to a very large number of computer users, the possible drawback of Visual Basic programs is that they do not perform as fast as applications developed in lower-level programming languages (such as C). However, that potential problem does not apply to SVB applications, especially those that rely mostly on executing calls to *STATISTICA's* analytic, graphics, and data management procedures. These procedures fully employ *STATISTICA* technology and perform at a speed comparable to running the respective procedures in *STATISTICA* directly.

Structure of *STATISTICA* Visual Basic. *STATISTICA* Visual Basic consists of two major components: 1) The general Visual Basic programming environment with facilities and extensions for designing user interfaces (dialogs) and file handling, and 2) the *STATISTICA* libraries with thousands of functions that provide access to practically all functionality of *STATISTICA*.

The Visual Basic programming environment follows the industry standard syntax conventions of the Microsoft Visual Basic Language; the few differences pertain mostly to the manner in which dialogs are created (see *Custom Dialogs* and *Custom User Interfaces* in the *STATISTICA Electronic Manual*), and are designed to offer programmers/developers more flexibility in the way user interfaces are handled in complex programs. In the SVB programming environment, dialogs can be entirely handled inside separate subroutines, which can be flexibly combined into larger multiple-dialog programs; MS Visual Basic is form based, where the forms or dialogs, and all events that occur in the dialogs, are handled in separate program units.

Attaching Macros to Toolbars and Menus

A *STATISTICA* Visual Basic program can be saved and then attached to a custom classic menu/toolbar or to the Quick Access toolbar on the ribbon bar. This enables you to easily customize and extend the operation and appearance of *STATISTICA* with your own custom macros. To utilize these facilities, save the macro by selecting ***Save As Global Macro*** from the ***File*** menu. Then, to customize the menus and/or toolbars, select ***Customize*** from the ***Tools*** menu to display the ***Customize*** dialog. To add the macro to a menu or toolbar, choose the ***Command/Macros*** tab, and select ***Macros*** from the ***Categories*** list. All your global macros will be listed in the ***Commands*** section of the tab.



Only macros that have been saved will show up in the Commands listbox. Double click on "Macros" in the Categories box to load additional macros into the Commands box.

You can then select and drag the specific item from the ***Commands*** list onto any menu or toolbar. Note that as your mouse pointer hovers over a menu, the menu will expand, enabling you to insert the item in any submenu as well. Once the macro is placed on the menu or toolbar while the ***Customize*** dialog is displayed, you can right-click the macro and change the appearance and text of the item, as well as add icons.

Running Macros from a command line. With *STATISTICA*, you can execute SVB programs from the command line by using the /RunMacro= command line parameter. The syntax is:

```
statist.exe /RunMacro=macroname
```

where “macroname” is the file name of the macro. If a full path is not specified, *STATISTICA* will attempt to run the macro from the application’s currently selected directory (which is Windows default behavior).

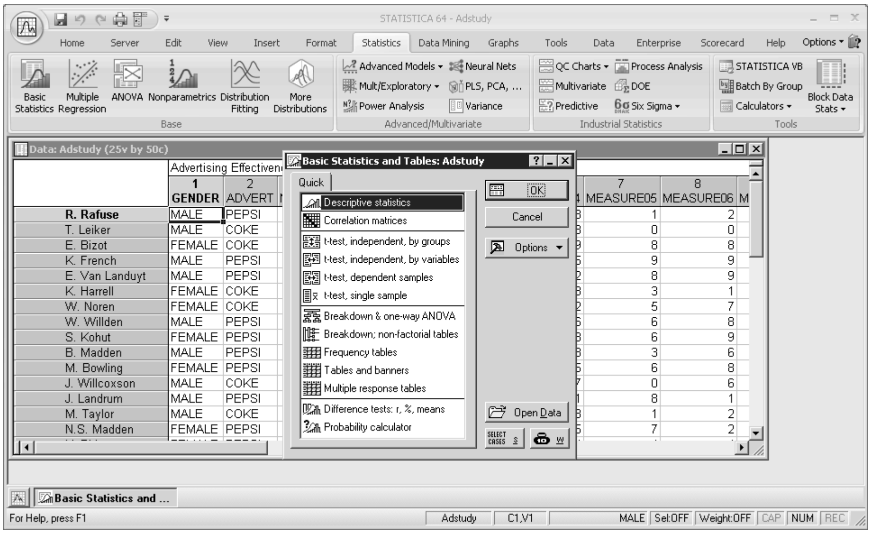
If the macro does not make the application or any document visible (through the `Application.Visible = True`, or similar document properties), the *STATISITCA* instance will automatically shut down when complete. If the application is made visible, the application will remain visible after the macro completes, and you will need to shut down the program.

EXAMPLE: RECORDING AN ANALYSIS

This example illustrates how to record an analysis into a script that can be executed to re-run the analysis. Then the script will be edited and combined with another script to create a customized script that can run analyses on demand. Additionally, this example shows how you can use attached scripts to auto-update and re-run analyses from results workbooks.

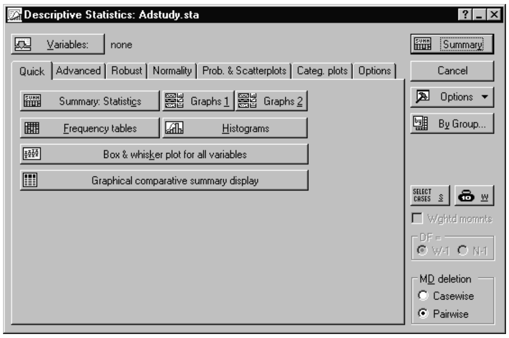
Start by opening the example *Adstudy* data set. Select the ***Home*** tab, click the ***Open*** arrow, and select ***Open Examples*** to display the ***Open a STATISTICA Data File*** dialog. Double-click on the *Datasets* file, and then open the *STATISTICA* data set *Adstudy.sta*.

Then, select the ***Statistics*** tab. In the ***Base*** group, click ***Basic Statistics*** to display the ***Basic Statistics and Tables*** Startup Panel. Select ***Descriptive statistics***.



	1 GENDER	2 ADVERT	7 MEASURE05	8 MEASURE06
R. Rafuse	MALE	PEPSI	1	2
T. Leiker	MALE	COKE	0	0
E. Bizot	FEMALE	COKE	8	8
K. French	MALE	PEPSI	9	9
E. Van Landuyt	MALE	PEPSI	8	9
K. Harrell	FEMALE	COKE	3	1
W. Noren	FEMALE	COKE	5	7
W. Willden	MALE	PEPSI	6	8
S. Kohut	FEMALE	PEPSI	6	9
B. Madden	MALE	PEPSI	3	6
M. Bowling	FEMALE	PEPSI	6	8
J. Willcoxson	MALE	COKE	0	6
J. Landrum	MALE	PEPSI	8	1
M. Taylor	MALE	COKE	1	2
N.S. Madden	FEMALE	PEPSI	7	2

Click the ***OK*** button to display the ***Descriptive Statistics*** dialog.



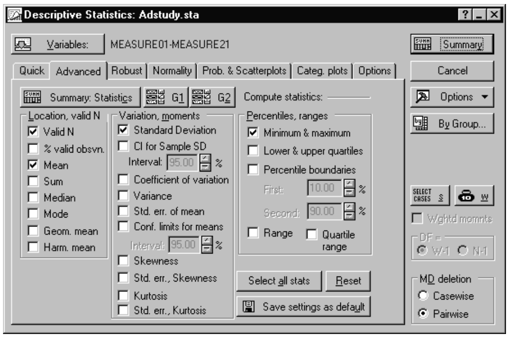
Descriptive Statistics: Adstudy.sta

Variables: none

MD deletion ○ Casewise ● Pairwise

Click the ***Variables*** button to display the ***Select the variables for the analysis*** dialog. Select variables *MEASURE01* through *MEASURE23* by clicking *MEASURE01* and dragging to *MEASURE23*, and then click ***OK***.

In the ***Descriptive Statistics*** dialog, select the ***Advanced*** tab, and note the numerous options available.



Descriptive Statistics: Adstudy.sta

Variables: MEASURE01-MEASURE21

Compute statistics:

Location, valid N
☑ Valid N
□ % valid obsvn.
☑ Mean
□ Sum
□ Median
□ Mode
□ Geom. mean
□ Harm. mean

Variation, moments
☑ Standard Deviation
□ CI for Sample SD
Interval: 95.00 %
□ Coefficient of variation
□ Variance
□ Std. err. of mean
□ Conf. limits for means
Interval: 95.00 %
□ Skewness
□ Std. err., Skewness
□ Kurtosis
□ Std. err., Kurtosis

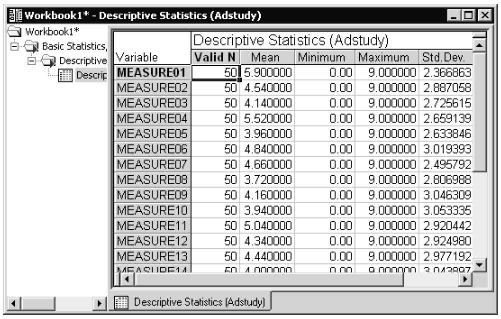
Percentiles, ranges
☑ Minimum & maximum
□ Lower & upper quartiles
□ Percentile boundaries
First: 10.00 %
Second: 90.00 %
□ Range
□ Quartile range

□ Wghtd momnts

DF = ● W-1 ○ N-1

MD deletion ○ Casewise ● Pairwise

For this example, we will leave all options at their default. Click the ***Summary*** button to display the descriptive statistics for the selected variables.



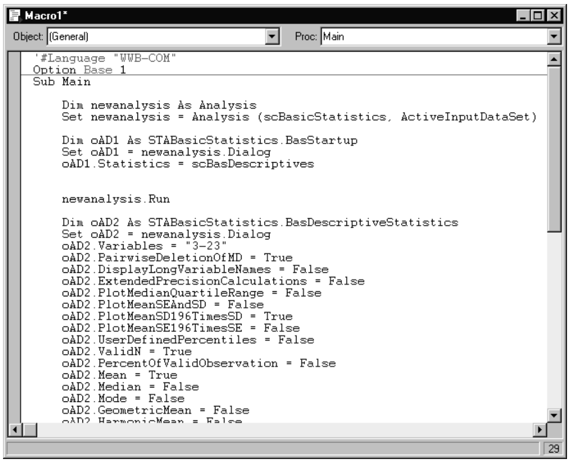
Workbook1* - Descriptive Statistics (Adstudy)

Descriptive Statistics (Adstudy)

Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
MEASURE01	50	5.900000	0.00	9.000000	2.366863
MEASURE02	50	4.540000	0.00	9.000000	2.887058
MEASURE03	50	4.140000	0.00	9.000000	2.725615
MEASURE04	50	5.520000	0.00	9.000000	2.659139
MEASURE05	50	3.960000	0.00	9.000000	2.633846
MEASURE06	50	4.840000	0.00	9.000000	3.019393
MEASURE07	50	4.660000	0.00	9.000000	2.495792
MEASURE08	50	3.720000	0.00	9.000000	2.806988
MEASURE09	50	4.160000	0.00	9.000000	3.046309
MEASURE10	50	3.940000	0.00	9.000000	3.053335
MEASURE11	50	5.040000	0.00	9.000000	2.920442
MEASURE12	50	4.340000	0.00	9.000000	2.924980
MEASURE13	50	4.440000	0.00	9.000000	2.977192
MEASURE14	50	4.000000	0.00	9.000000	3.043897

When you produce the results workbook, the ***Descriptive Statistics*** dialog is automatically minimized so you can see the results. To restore the dialog, click the ***Descriptive Statistics*** button on the Analysis Bar in the lower-left of the screen.

While you are running this analysis, *STATISTICA* automatically records all the analysis steps behind the scenes. You can now produce a *STATISTICA* Visual Basic (SVB) macro to re-create this analysis. In the ***Descriptive Statistics*** dialog, click the Options ▾ button, and select ***Create Macro*** from the drop-down menu. The ***New Macro*** dialog will be displayed, where you can name the macro and enter a description. Leave all the entries at their defaults, and click ***OK***. An SVB macro window will be displayed, containing the recorded *Descriptive Statistics* session.



```
'#Language "WWB-COM"
Option Base 1
Sub Main

    Dim newanalysis As Analysis
    Set newanalysis = Analysis (scBasicStatistics, ActiveInputDataSet)

    Dim oAD1 As STABasicStatistics.BasStartup
    Set oAD1 = newanalysis.Dialog
    oAD1.Statistics = scBasDescriptives


    newanalysis.Run

    Dim oAD2 As STABasicStatistics.BasDescriptiveStatistics
    Set oAD2 = newanalysis.Dialog
    oAD2.Variables = "3-23"
    oAD2.PairwiseDeletionOfMD = True
    oAD2.DisplayLongVariableNames = False
    oAD2.ExtendedPrecisionCalculations = False
    oAD2.PlotMedianQuartileRange = False
    oAD2.PlotMeanSEAndSD = False
    oAD2.PlotMeanSD196TimesSD = True
    oAD2.PlotMeanSE196TimesSE = False
    oAD2.UserDefinedPercentiles = False
    oAD2.ValidN = True
    oAD2.PercentOfValidObservation = False
    oAD2.Mean = True
    oAD2.Median = False
    oAD2.Mode = False
    oAD2.GeometricMean = False
    oAD2.HarmonicMean = False
```

To run this macro, select the ***Debug*** tab, and in the ***Run*** group, click ***Run*** (or press F5 on your keyboard). The exact *Descriptive Statistics* results that were generated in the initial analysis will be reproduced.

Look at the SVB macro for a moment. Toward the top, one of the lines is:

```
Set newanalysis = Analysis (scBasicStatistics, ActiveInputDataSet)
```

This is telling the macro that it is going to run the *Basic Statistics* analysis, and that it will be using the “active” data set, that is, the spreadsheet that is currently selected when the macro runs.

A few lines further down is a section that starts with:

```
Dim oAD2 As STABasicStatistics.BasDescriptiveStatistics
```

and under that are properties such as:

```
.PairwiseDeletionOfMD = True
```

These properties correspond to all the options that were available on the different tabs of the ***Descriptive Statistics*** dialog. Every option in the dialog is represented by a property, and all the current settings are recorded. If you decide to include a *Median* and the *Sum* of each of the variables, it is easy to add this to the SVB macro; just find the lines that read:

```
.Median = False
```

and

```
.Sum = False
```

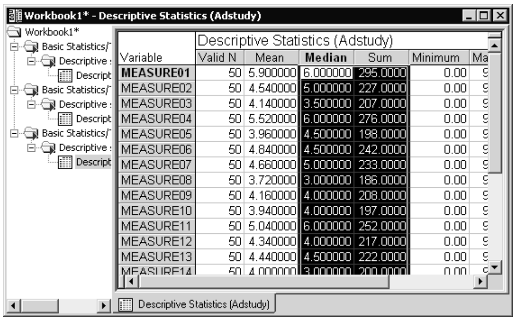
and change these to:

```
.Median = True
```

and

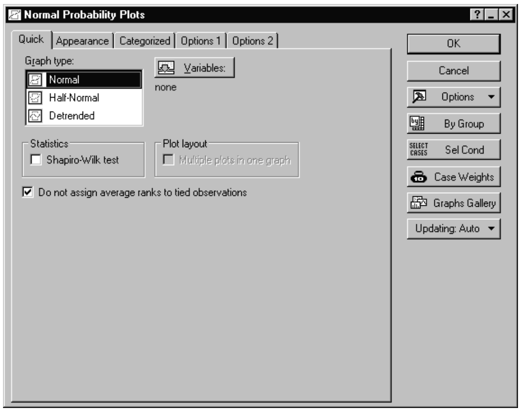
```
.Sum = True
```

Now, run the macro again by pressing F5. A new results spreadsheet will be added to the workbook, this time with new columns of *Median* and *Sum*:



	Descriptive Statistics (Adstudy)					
Variable	Valid N	Mean	Median	Sum	Minimum	Ma
MEASURE01	50	5.900000	6.000000	295.0000	0.00	9
MEASURE02	50	4.540000	5.000000	227.0000	0.00	9
MEASURE03	50	4.140000	3.500000	207.0000	0.00	9
MEASURE04	50	5.520000	6.000000	276.0000	0.00	9
MEASURE05	50	3.960000	4.500000	198.0000	0.00	9
MEASURE06	50	4.840000	4.500000	242.0000	0.00	9
MEASURE07	50	4.660000	5.000000	233.0000	0.00	9
MEASURE08	50	3.720000	3.000000	186.0000	0.00	9
MEASURE09	50	4.160000	4.000000	208.0000	0.00	9
MEASURE10	50	3.940000	4.000000	197.0000	0.00	9
MEASURE11	50	5.040000	6.000000	252.0000	0.00	9
MEASURE12	50	4.340000	4.000000	217.0000	0.00	9
MEASURE13	50	4.440000	4.500000	222.0000	0.00	9
MEASURE14	50	4.000000	3.000000	200.0000	0.00	9

Let's keep the macro window open and start a new analysis on the same sample data set. Select the *Adstudy* spreadsheet to bring it to the front. Select the ***Graphs*** tab, and in the ***More*** group, click ***2D***. Select ***Normal Probability Plots*** to display the ***Normal Probability Plots*** dialog.



Graph type: Normal | Half-Normal | Detrended

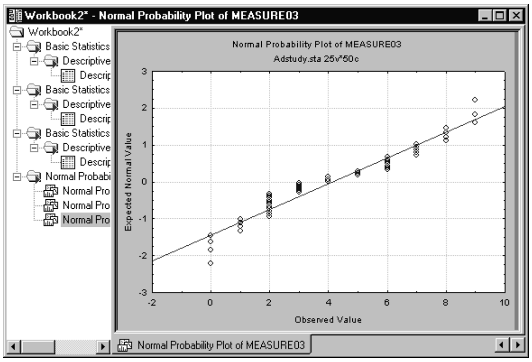
Variables: none

Statistics

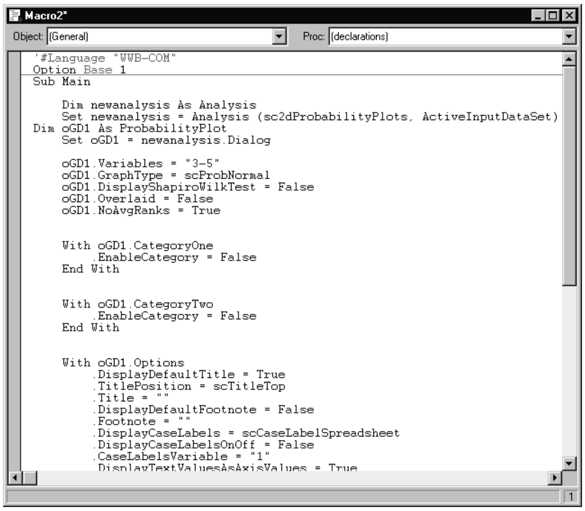
☐ Shapiro-Wilk test

☑ Do not assign average ranks to tied observations

Click the ***Variables*** button, and in the ***Select Variables for Probability Plot*** dialog, select variables *MEASURE01* through *MEASURE03*. Click ***OK*** to close this dialog, and click ***OK*** in the ***Normal Probability Plots*** dialog. Three *Probability Plot* graphs will be placed in the results workbook, one for each of the three variables that were selected.



The steps of the *Probability Plot* analysis were recorded just as they were for the *Descriptive Statistics* analysis. To create a new macro with these steps, bring the ***Normal Probability Plot*** dialog to the front by clicking that button on the Analysis Bar in the lower-left of the screen, click the Options ▾ button, and select ***Create Macro*** from the drop-down menu. In the ***New Macro*** dialog, click ***OK***, and a new SVB Macro window is opened with the recorded *Probability Plot* script.



```
'#Language "WWB-COM"
Option Base 1
Sub Main

    Dim newanalysis As Analysis
    Set newanalysis = Analysis (sc2dProbabilityPlots, ActiveInputDataSet)
Dim oGD1 As ProbabilityPlot
    Set oGD1 = newanalysis.Dialog

    oGD1.Variables = "3-5"
    oGD1.GraphType = scProbNormal
    oGD1.DisplayShapiroWilkTest = False
    oGD1.Overlaid = False
    oGD1.NoAvgRanks = True


    With oGD1.CategoryOne
        .EnableCategory = False
    End With


    With oGD1.CategoryTwo
        .EnableCategory = False
    End With


    With oGD1.Options
        .DisplayDefaultTitle = True
        .TitlePosition = scTitleTop
        .Title = ""
        .DisplayDefaultFootnote = False
        .Footnote = ""
        .DisplayCaseLabels = scCaseLabelSpreadsheet
        .DisplayCaseLabelsOnOff = False
        .CaseLabelsVariable = "1"
        .DisplayTextValuesAsAxisValues = True
```

As with the *Descriptive Statistics* analysis, all the options selected in the ***Probability Plot*** dialog are specified as properties within the macro. For instance, to change this from a *Normal Probability Plot* to a *Half Normal Probability Plot*, locate the following line:

```
    .GraphType = scProbNormal
```

and change it to:

```
    .GraphType = scProbHalfNormal
```

Also, let's expand the variables to include variable *MEASURE04*. To do this, find the following line:

```
    .Variables = "3-5"
```

This line corresponds to the variables selected for the plots. Since we selected *MEASURE01* through *MEASURE03*, and these are variable numbers *3* through *5* from the data set, this string was recorded. To add *MESURE04* (variable number *6*), change this line to:

```
    .Variables = 3-6
```

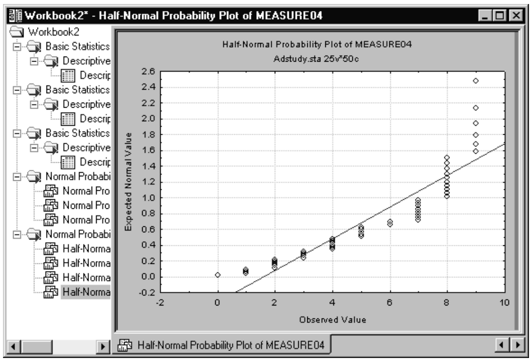
Now run the macro by pressing F5. Four new graphs are produced as *Half-Normal Probability Plots* for variables *MEASURE01* through *MEASURE04*.

This example has demonstrated how you can run any analysis, and then create a macro of the analysis that can be edited and rerun. Additionally, this example has shown how these macros can be combined to develop macros that are more complex. This is the building block of creating your own powerful customized analyses using the SVB language.

Rerunning Analyses from Results Workbooks

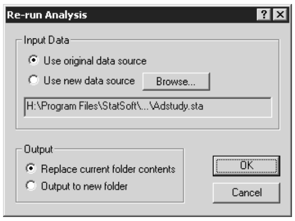
In the previous example, you learned that all analyses in *STATISTICA* will record the steps used to produce them, and these can be loaded into a macro that you can edit and run. When an analysis produces results that are placed in a workbook, *STATISTICA* automatically associates the recorded script's steps to the workbook folder that contains the results. This enables you to either rerun the analysis or to resume an analysis.

Thus far, we have produced several instances of running both *Descriptive Statistics* and *Probability Plots*. The results workbook looks similar to the following illustration.



Notice that there is a red arrow on each workbook folder. This is an indicator that the script that produced the results in that folder has been attached to the folder. This enables *STATISTICA* to rerun or resume the analysis.

To rerun an analysis, right-click on one of the folders labeled *Descriptive statistics dialog*, and from the shortcut menu, select ***Re-run Analysis***. The ***Re-run Analysis*** dialog will be displayed.

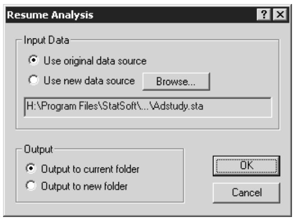


● Use original data source
○ Use new data source
H:\Program Files\StatSoft\...\Adstudy.sta
● Replace current folder contents
○ Output to new folder

Here you can choose to ***Use original data source*** or ***Use new data source***. The latter option gives you the powerful ability to create "templates" that can then be applied to new data sources. In addition to specifying the data source, you can choose ***to Replace current folder contents*** or ***Output to new folder***. In this example, leave the defaults, and click ***OK***. You will see that the contents of the folder are briefly deleted and then added again as the analysis is rerun.

One purpose for this feature is the ability to update/rerun results from complex analyses if new data is entered into the spreadsheet. For instance, if the data in the open data file *Adstudy.sta* has been changed and the analysis is rerun, the new results will be calculated with the new data.

The resume analysis functionality enables you to bring an analysis back to the point before the results were generated, allowing you to select different options or continue an analysis in progress. Right-click the same *Descriptive statistics dialog* folder, and from the shortcut menu, select ***Resume Analysis***. The ***Resume Analysis*** dialog will be displayed. This dialog also contains options to specify the input data source (original or new). The ***Output*** options for the new results are to ***Output to current folder*** (as if this is just an extension of the previous analysis) ***or Output to new folder*** (as if this is a brand new analysis).



● Use original data source
○ Use new data source
H:\Program Files\StatSoft\...\Adstudy.sta
● Output to current folder
○ Output to new folder

Leave the defaults as they are, and click ***OK***. The ***Descriptive Statistics*** dialog will be displayed, with all the options set to what was used when the selected output was created. Since the default was to ***Output to current folder***, clicking the ***Summary*** button will generate new output to the same folder.

ActiveX OBJECTS AND DOCUMENTS
(A TECHNICAL NOTE)

The term ActiveX is used in different contexts, and its definitions stress different aspects of that concept. Its use within *STATISTICA*, however, can be grouped into two general categories: ActiveX objects and ActiveX documents.

ActiveX objects. An ActiveX object is what was once referred to as an OLE (Object Linking and Embedding) object. At its heart is the Microsoft COM (Component Object Model) technology that makes it possible for objects to be accessed in a uniform manner. Through the use of standard protocols, objects created in one application can be stored and edited in a different application. To support this functionality, the containing object needs to be an ActiveX object client, and the application that initially created the object needs to be an ActiveX object server. *STATISTICA* is both. As an ActiveX object client, *STATISTICA* allows you to embed and link objects from other applications in spreadsheets, graphs, and reports. As an ActiveX object server, it allows you to embed and link spreadsheets and graphs into other applications.

ActiveX documents. ActiveX documents take the ActiveX controls one step further, in that they allow entire documents to be embedded into other applications. An ActiveX document container allows other application documents to be used within it, and an ActiveX document server allows its documents to be used within any ActiveX document container. Again, *STATISTICA* does both. *STATISTICA* Workbooks are ActiveX document containers, and allow documents from other ActiveX servers to be displayed within the workbook. Examples of this are Word and Excel; these documents can be used directly from within a *STATISTICA* Workbook. Similarly, *STATISTICA* Spreadsheets, Graphs, and Reports are ActiveX document servers, and they also can be placed within any ActiveX document container such as Microsoft Internet Explorer.

Office integration and ActiveX documents. The ActiveX document technology has special application with Word and Excel documents. *STATISTICA* can open these particular documents natively in their own windows within the *STATISTICA* workspace. This “Office integration” enables you to use Excel documents as data sources and Word documents as reports for analyses. When the documents are open in the *STATISTICA* window, the appropriate menus and toolbars for Excel/Word are available for use.

9 CHAPTER

STATISTICA QUERY

9

CHAPTER

STATISTICA *QUERY*

Note: For an explanation of all technical terms used in this overview (e.g., ODBC, SQL, OLAP, etc.), please refer to the glossary in the *STATISTICA Electronic Manual*, accessible by clicking ***Help*** on the ***Help*** tab in the ***Help*** group.

This chapter includes a brief overview of *STATISTICA Query*, a flexible tool for accessing data from external databases. It also includes information on retrieving data from OLAP Cube providers such as MS OLE DB Provider for Analysis Services or SAP Business Warehouse MDX.

OVERVIEW

STATISTICA Query is used to access data easily from a wide variety of databases (including many large system databases such as Oracle, MS SQL Server, Sybase, etc.) using Microsoft's OLE DB conventions. OLE DB is a powerful database technology that provides universal data integration over an enterprise's network, from mainframe to desktop, regardless of the data type. OLE DB offers a more generalized and more efficient strategy for data access than the older ODBC conventions because it allows access to more types of data and is based on the Component Object Model (COM).

STATISTICA Query supports multiple database tables; specific records (rows of tables) can be selected by entering SQL statements, which *STATISTICA Query* automatically builds for you as you select the components of the query via a simple graphical interface and/or intuitive menu options and dialogs. Therefore, an extensive knowledge of SQL is not necessary in order for you to create advanced and powerful queries of data in a quick and straightforward manner. Multiple queries based on one or many different databases can also be created to return data to an individual spreadsheet, and you can maintain connections to multiple external databases simultaneously.

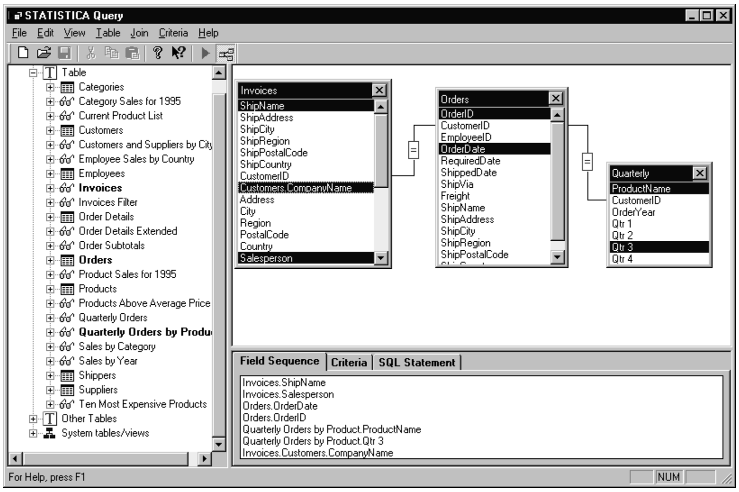
STATISTICA QUERY: QUICK STEP-BY-STEP INSTRUCTIONS

The steps necessary to retrieve external data via *STATISTICA Query* are outlined below:

1. Select the ***Home*** tab. In the ***File*** group, click the ***Open*** arrow. Select ***Open External Data - Create Query*** to display the ***Database Connection*** dialog. (You can also select the ***Data*** tab. In the ***Manage*** group, click ***External Data*** and select ***Create Query*** to display the ***Database Connection*** dialog.) In this dialog, select a predefined database connection (the provider, data source location, and advanced settings of the server or directory on which the data resides).

   Note that if you have not already created the database connection, you can do so by clicking the ***New*** button in the ***Database Connection*** dialog. The ***Data Link Properties*** dialog will be displayed, which will guide you through a step-by-step wizard to create a database connection. For specific documentation when you are using the ***Data Link Properties*** dialog, press the F1 key on your keyboard to display the Microsoft Data Link Help®.

2. After you have selected a database connection and clicked the ***OK*** button in the ***Data Link Properties*** dialog, you will have access to *STATISTICA Query* in which you can create a SQL statement by specifying the desired tables, fields, joins, criteria, etc. (via the ***Table***, ***Join,*** and ***Criteria*** menus) to be included in your query.



Invoices.ShipName
Invoices.Salesperson
Orders.OrderDate
Orders.OrderID
Quarterly Orders by Product.ProductName
Quarterly Orders by Product.Qtr 3
Invoices.Customers.CompanyName

3. Once you have specified a query, select ***Return Data to STATISTICA*** from the ***File*** menu. The ***Returning External Data to Spreadsheet*** dialog will be displayed, in which you can specify the name of the query, where you want *STATISTICA Query* to put the data that the query returns, and additional options.

   See the *Electronic Manual* for further details.

IN-PLACE PROCESSING OF DATA ON REMOTE SERVERS
(THE IDP TECHNOLOGY OPTION)

The query facilities (described in the previous sections), when offered as part of the enterprise versions of *STATISTICA* (see *STATISTICA Enterprise Systems*, page 278), are additionally enhanced by options to process data from remote servers “in-place,” that is, without having to import them and create a local data file. This *In-Place Database Processing* (*IDP*) technology is particularly useful for processing extremely large data files where it can produce significant performance gains and enable *STATISTICA* users to process data files that exceed the storage capacity of the local device or even the *STATISTICA Enterprise Server*.

Technical note. The *IDP* technology is based on distributed processing architecture, where the queries are performed on the server side (using the server

CPU resources) and the respective records sent to the *STATISTICA* computer where they are simultaneously (asynchronously) processed as they become available.

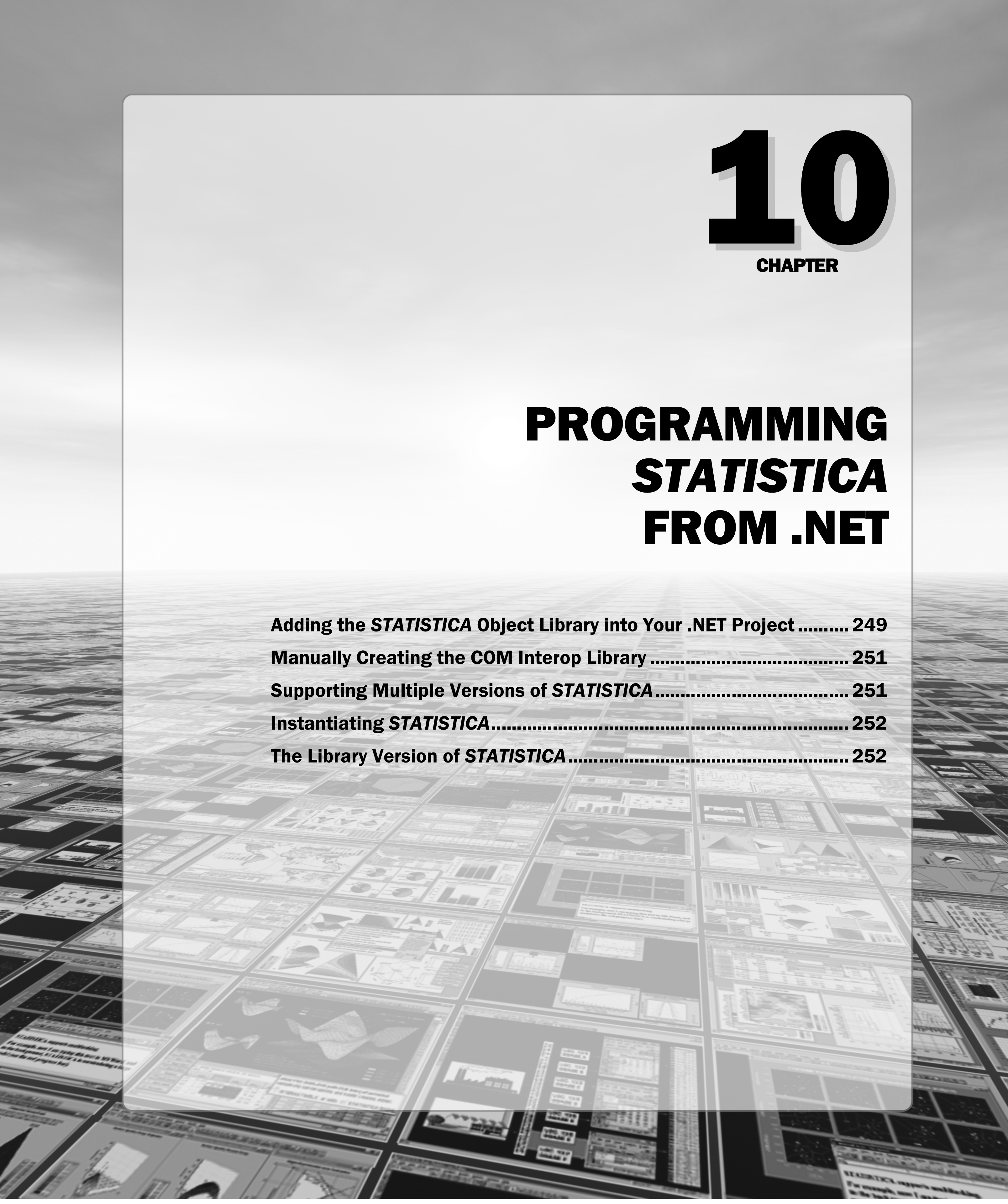
OLAP CUBES

OLAP (On-Line Analytic Processing) is a generic term for a system that provides efficient access to summary data about very large databases. Unlike ordinary relational databases, which organize data as a set of well-defined, two-dimensional tables, an OLAP data warehouse represents data at many levels of detail in multi-dimensional data sets known as "cubes." When a *STATISTICA* user wants to perform an analysis against data from an OLAP Cube, the data must be reduced to a two-dimensional form – cases and variables – that can be represented in a *STATISTICA* spreadsheet. The *STATISTICA Query* tool provides a graphical, drag-and-drop interface for specifying the dimensions and levels of detail that will be extracted from the cube to feed into the query. The "MDX" (Multi-Dimensional eXpressions) mode is triggered automatically when an OLAP data source is selected.

Customers who require OLAP integration will usually have sophisticated database support technicians through their in-house information technology department who can help design these queries. Because the configuration of the dimensions in an OLAP cube is determined by the customer's database administrators, StatSoft can provide only limited assistance in this area.

LARGE DATABASE FILES

STATISTICA products are designed for large-scale analytics; consequently, they integrate well with database systems designed for managing very large amounts of data, such as Teradata and others. For example, *STATISTICA* can both extract data for analysis from Teradata, and it can also score results directly inside Teradata through deployment code created by *STATISTICA Data Miner* and applied to the Teradata as a user-defined function, which significantly accelerates processing of large amounts of data.



10

CHAPTER

PROGRAMMING *STATISTICA* FROM .NET

10

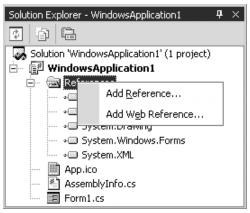
CHAPTER

PROGRAMMING *STATISTICA* FROM .NET

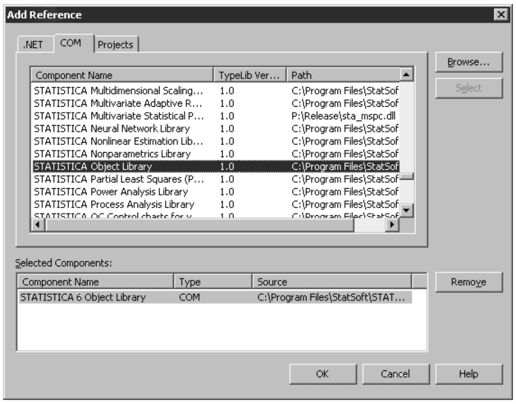
Virtually every aspect of *STATISTICA* is exposed as a set of COM interfaces that are registered on a machine when *STATISTICA* is installed. Since .NET-based languages cannot communicate with COM directly, a wrapper class called the COM Interop can be utilized to integrate the *STATISTICA* libraries into your .NET project. The COM Interop layer is created automatically by the Visual Studio .NET IDE when you import a COM interface. The COM Interop layer handles all of the details regarding interacting with the COM libraries in .NET. With the COM Interop layer in place, the *STATISTICA* COM interfaces behave like any other .NET object.

Adding the *STATISTICA* Object Library into Your .NET Project

The .NET Interop layer is created automatically by adding the desired *STATISTICA* COM interfaces into your .NET project. *STATISTICA* Object Library is the base *STATISTICA* COM library. To add the *STATISTICA* Object Library to a .NET project, first select the desired .NET project in Solution Explorer, and then select ***Add Reference*** from the shortcut menu (accessed by right-clicking on the .NET project).



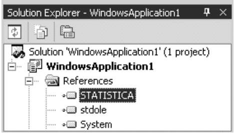
The ***Add Reference*** dialog will be displayed. Select the ***COM*** tab. From the ***Component Name*** list, select ***STATISTICA Object Library***, and click ***OK***.



Component Name	TypeLib Ver...	Path
STATISTICA Multidimensional Scaling...	1.0	C:\Program Files\StatSof
STATISTICA Multivariate Adaptive R...	1.0	C:\Program Files\StatSof
STATISTICA Multivariate Statistical P...	1.0	P:\Release\sta_mspc.dll
STATISTICA Neural Network Library	1.0	C:\Program Files\StatSof
STATISTICA Nonlinear Estimation Lib...	1.0	C:\Program Files\StatSof
STATISTICA Nonparametrics Library	1.0	C:\Program Files\StatSof
STATISTICA Object Library	1.0	C:\Program Files\StatSof
STATISTICA Partial Least Squares (P...	1.0	C:\Program Files\StatSof
STATISTICA Power Analysis Library	1.0	C:\Program Files\StatSof
STATISTICA Process Analysis Library	1.0	C:\Program Files\StatSof
STATISTICA QC Control charts for v...	1.0	C:\Program Files\StatSof

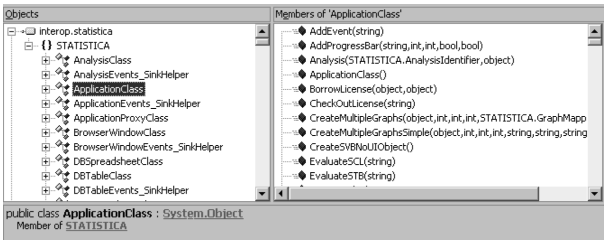
Component Name	Type	Source
STATISTICA 6 Object Library	COM	C:\Program Files\StatSoft\STAT...

At this point, the necessary COM Interop library is created automatically. Under the project *References* node, you will now see the entry *STATISTICA*.



```
Solution 'WindowsApplication1' (1 project)
  WindowsApplication1
    References
      STATISTICA
      stdole
      System
```

The file Interop.*STATISTICA.dll* is also added to the project output directory. The *STATISTICA* COM Interop library is stored in this file. To view the *STATISTICA* object library from your .NET project, right-click on the *STATISTICA* reference, and from the shortcut menu, select ***View in Object Browser***.



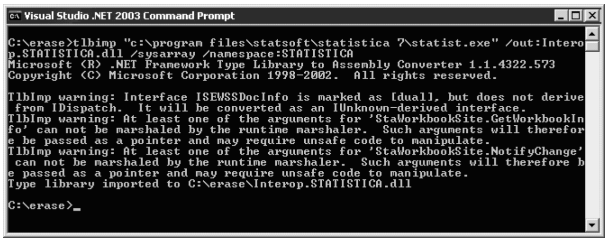
```
interop.statistica
  {} STATISTICA
    AnalysisClass
    AnalysisEvents_SinkHelper
    ApplicationClass
    ApplicationEvents_SinkHelper
    ApplicationProxyClass
    BrowserWindowClass
    BrowserWindowEvents_SinkHelper
    DBSpreadsheetClass
    DBTableClass
    DBTableEvents_SinkHelper
```

```
AddEvent(string)
AddProgressBar(string,int,int,bool,bool)
Analysis(STATISTICA.AnalysisIdentifier,object)
ApplicationClass()
BorrowLicense(object,object)
CheckOutLicense(string)
CreateMultipleGraphs(object,int,int,int,STATISTICA.GraphMapp
CreateMultipleGraphsSimple(object,int,int,int,string,string,string
CreateSVBNoUIObject()
EvaluateSCL(string)
EvaluateSTB(string)
```

```
public class ApplicationClass : System.Object
   Member of STATISTICA
```

Manually Creating the COM Interop Library

It is also possible to create the COM Interop library manually and import it into your .NET project. This gives you the ability to specify a different name for the Interop DLL as well as define a custom namespace. The program that enables you to create an Interop is *TLBIMP.EXE*. From a Visual Studio command prompt, execute *TLBIMP* with an initial parameter of the type library source. In the example below, the output DLL name and namespace are also specified.



```
C:\erase>tlbimp "c:\program files\statsoft\statistica 7\statist.exe" /out:Intero
p.STATISTICA.dll /sysarray /namespace:STATISTICA
Microsoft (R) .NET Framework Type Library to Assembly Converter 1.1.4322.573
Copyright (C) Microsoft Corporation 1998-2002.  All rights reserved.

TlbImp warning: Interface ISEWSSDocInfo is marked as [dual], but does not derive
 from IDispatch.  It will be converted as an IUnknown-derived interface.
TlbImp warning: At least one of the arguments for 'StaWorkbookSite.GetWorkbookIn
fo' can not be marshaled by the runtime marshaler.  Such arguments will therefor
e be passed as a pointer and may require unsafe code to manipulate.
TlbImp warning: At least one of the arguments for 'StaWorkbookSite.NotifyChange'
 can not be marshaled by the runtime marshaler.  Such arguments will therefore b
e passed as a pointer and may require unsafe code to manipulate.
Type library imported to C:\erase\Interop.STATISTICA.dll

C:\erase>_
```

In this example, we reference the file *STATIST.EXE* since that executable contains the *STATISTICA Object Library* type library. Once the Interop DLL is generated, you can add it to your .NET project by selecting ***Add Reference*** from the *Solution Explorer* as before, but this time click the ***Browse*** button to select the newly created Interop DLL.

Supporting Multiple Versions of *STATISTICA*

To support multiple versions of *STATISTICA*, it is necessary to maintain separate *STATISTICA* Object Library Interop DLLs for each version of *STATISTICA* you want to support. You can use the *TLBIMP* command to generate Interop DLLs against specific versions of *STATIST.EXE* and other DLLs. When distributing the application, ensure that the correct version of the *STATISTICA* Interop DLL is deployed with your .NET application.

Instantiating *STATISTICA*

Because of its COM architecture, *STATISTICA* can be incorporated into many different development environments. When using *STATISTICA* from an external development environment, it is necessary to have a top-level object called the application object. The application object is the application itself and will contain other objects (for example, spreadsheets and graphs), but access to these other objects is restricted unless the application object is running.

Assuming you are using the default namespace *STATISTICA*, the interface you should declare your variable as is *STATISTICA.Application*. To create an instance of *STATISTICA*, set your variable equal to `new STATISTICA.ApplicationClass().`

```
STATISTICA.Application pApp = (STATISTICA.Application)
      new STATISTICA.ApplicationClass();
pApp.Visible = true;
```

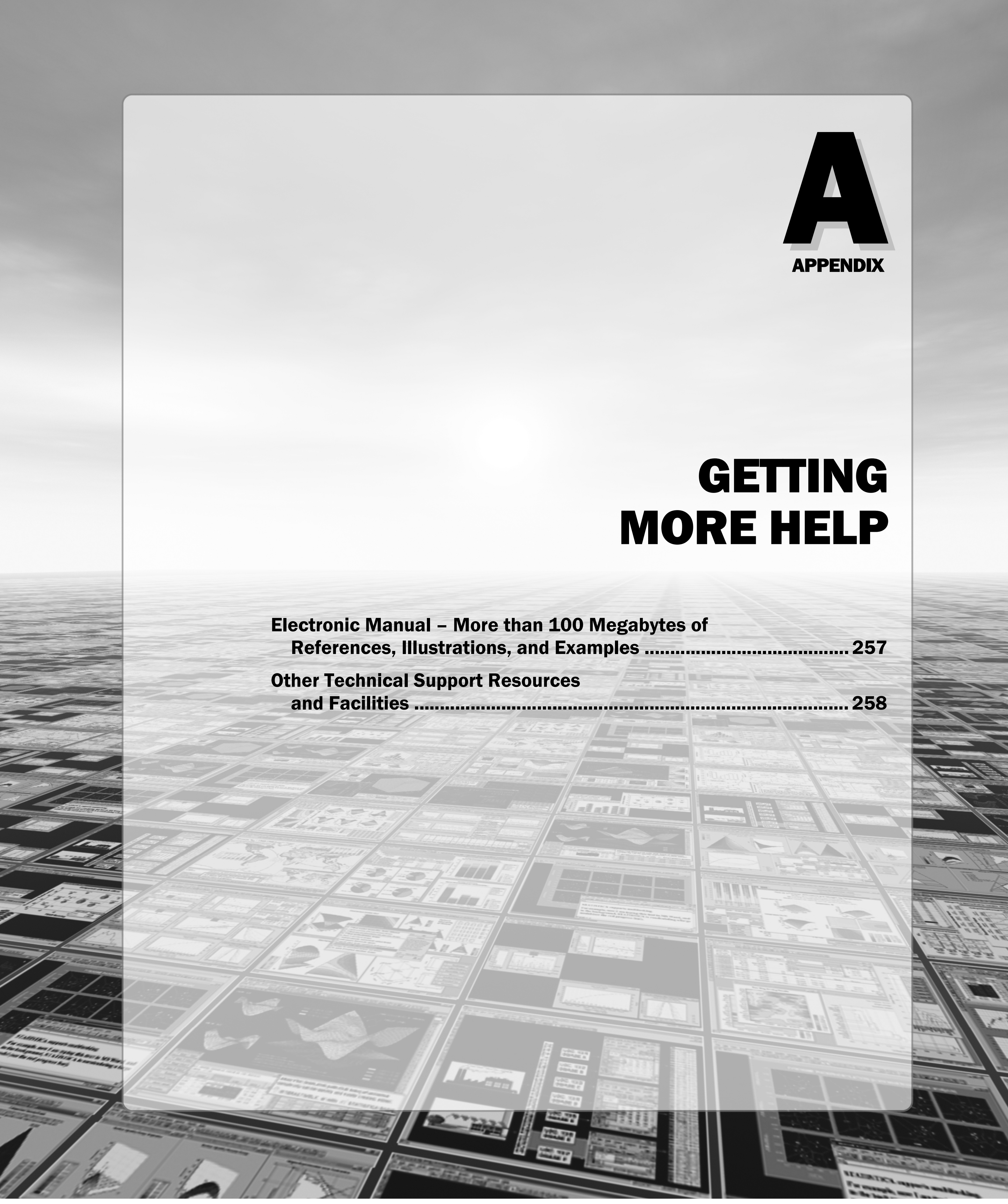
When an instance of the STATISTICA.ApplicationClass is created, a *STATIST.EXE* process will be launched. This is equivalent to launching *STATISTICA* from the ***Start*** menu. The *STATISTICA* instance is initially hidden but can be made visible. Since it is a separate process, all calls to this instance are made out of process.

The Library Version of *STATISTICA*

In addition to the *STATISTICA.Application* object, there is also a lighter-weight, higher-performance version of the object called *STATISTICA.Library*. The *Library* version is licensed separately and therefore may not be available with your installation. It contains identical interfaces as the *STATISTICA.Application* library. Any existing code that uses the *Application* object can be replaced with the *Library* object.

The main restriction is that the *STATISTICA* user interface features are not available from the *Library* version. Therefore, in the example above, if the *Application* object was instantiated as a new *STATISTICA.LibraryClass*, it would not be possible to make the object visible (and show the *STATISTICA* interface).

The *Library* version of *STATISTICA* is loaded in-process, which means accessing its COM interfaces is more efficient than using the *Application* version of the object (which is loaded out of process). Since it is loaded in-process, multiple versions of the library cannot be instantiated. Normally, you would only instantiate one *Library* object or one *Application* object in your program.



A

APPENDIX

GETTING MORE HELP

A

APPENDIX

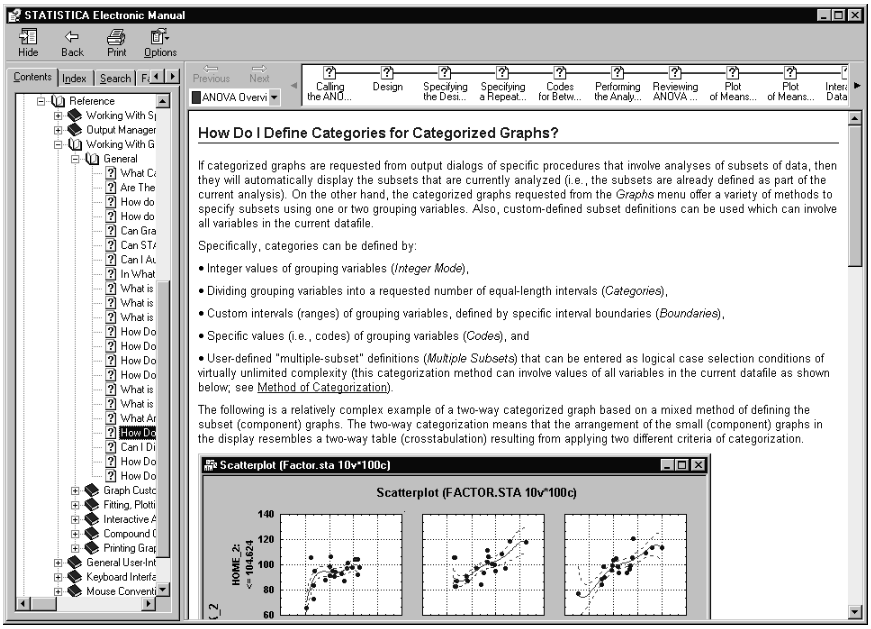
GETTING MORE HELP

Electronic Manual

The most convenient place to get assistance and access a vast repository of information about *STATISTICA* is the *Electronic Manual* (Help), which contains more than 100 Megabytes of references, illustrations, and examples.

To access Help in *STATISTICA*, select the ***Help*** tab. In the ***Help*** group, click ***Help***. You can also click the Help ? button in the upper-right corner of any dialog to access Help topics describing all the options in that dialog.

This hypertext document offers much more than just an explanation of the options in *STATISTICA*. It includes numerous examples, overviews, and illustrations, as well as thousands of tips on how to optimize your work.



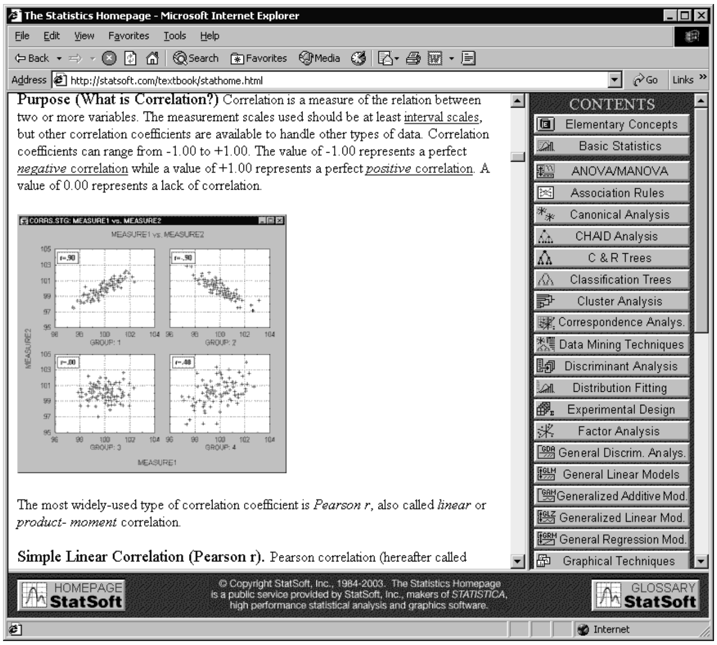
How Do I Define Categories for Categorized Graphs?

If categorized graphs are requested from output dialogs of specific procedures that involve analyses of subsets of data, then they will automatically display the subsets that are currently analyzed (i.e., the subsets are already defined as part of the current analysis). On the other hand, the categorized graphs requested from the *Graphs* menu offer a variety of methods to specify subsets using one or two grouping variables. Also, custom-defined subset definitions can be used which can involve all variables in the current datafile.

Specifically, categories can be defined by:

- Integer values of grouping variables (*Integer Mode*),
- Dividing grouping variables into a requested number of equal-length intervals (*Categories*),
- Custom intervals (ranges) of grouping variables, defined by specific interval boundaries (*Boundaries*),
- Specific values (i.e., codes) of grouping variables (*Codes*), and
- User-defined "multiple-subset" definitions (*Multiple Subsets*) that can be entered as logical case selection conditions of virtually unlimited complexity (this categorization method can involve values of all variables in the current datafile as shown below; see Method of Categorization).

The following is a relatively complex example of a two-way categorized graph based on a mixed method of defining the subset (component) graphs. The two-way categorization means that the arrangement of the small (component) graphs in the display resembles a two-way table (crosstabulation) resulting from applying two different criteria of categorization.



http://statsoft.com/textbook/stathome.html

Purpose (What is Correlation?) Correlation is a measure of the relation between two or more variables. The measurement scales used should be at least interval scales, but other correlation coefficients are available to handle other types of data. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect *negative* correlation while a value of +1.00 represents a perfect *positive* correlation. A value of 0.00 represents a lack of correlation.

The most widely-used type of correlation coefficient is *Pearson r*, also called *linear* or *product- moment* correlation.

Simple Linear Correlation (Pearson r). Pearson correlation (hereafter called

© Copyright StatSoft, Inc., 1984-2003. The Statistics Homepage is a public service provided by StatSoft, Inc., makers of *STATISTICA*, high performance statistical analysis and graphics software.

The *STATISTICA Electronic Manual* is extremely comprehensive. It offers a built-in ***Statistical Advisor*** (see page 33) supplemented with the complete contents of StatSoft's award-winning *Electronic Statistics Textbook* and *Glossary*.

StatSoft's *Electronic Statistics Textbook*, located on the company Web site (***StatSoft.com***), has been recommended by *Encyclopedia Britannica* for its "Quality, Accuracy, Presentation, and Usability."

StatSoft web site praised for Quality, Accuracy, Presentation and Usability.

This unique textbook has been used for many years in educational and research activities at universities and research organizations worldwide.

Other Technical Support Resources and Facilities

Web site resources. StatSoft's Web site, one of the most visited Internet addresses related to data analysis, offers not only access to many resources that are useful for data analysis professionals in general, but it also includes:

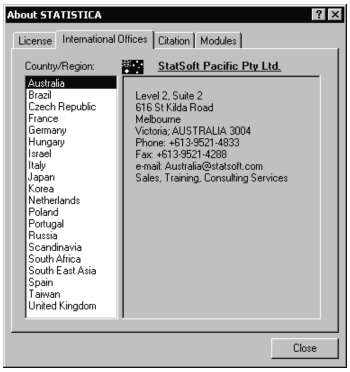
- A continuously updated *Frequently Asked Questions* section, and
- A download area where users of the current version of *STATISTICA* products can receive downloadable updates of their software. We are constantly working on increasing the compatibility of *STATISTICA* software even with those applications that violate standard conventions. Therefore, in many circumstances, downloading an update can help when the problem that you are experiencing is caused by nonstandard system configurations or conflicts with other applications.

E-mail technical support. If your question is not answered in the locations mentioned, you may send e-mail to us. Please include your serial number (in *STATISTICA*, select the ***Help*** tab, and in the ***About*** group, click ***STATISTICA*** to view your serial number) and information about your hardware [the type of processor (CPU) and the amount of memory (RAM) and disk space] and the version of the operating system that you are using.

If you live in North America, send your e-mail to info@StatSoft.com; otherwise, e-mail your local StatSoft office (see below).

Phone technical support. You can also call your local StatSoft office to talk to a technician. If you live in North America, call (918) 749-1119 (the North American technical support office hours are 9:00 AM to 5:00 PM Central Time, Monday through Friday).

If you live in another location, please contact the office that serves your specific area. To locate that office, select the ***Help*** tab. In the ***About*** group, click ***STATISTICA*** to display the ***About STATISTICA*** dialog, and then select the ***International Offices*** tab.



Australia
Brazil
Czech Republic
France
Germany
Hungary
Israel
Italy
Japan
Korea
Netherlands
Poland
Portugal
Russia
Scandinavia
South Africa
South East Asia
Spain
Taiwan
United Kingdom

StatSoft Pacific Pty Ltd.

Level 2, Suite 2
616 St Kilda Road
Melbourne
Victoria; AUSTRALIA 3004
Phone: +613-9521-4833
Fax: +613-9521-4288
e-mail: Australia@statsoft.com
Sales, Training, Consulting Services

Please know your serial number (in *STATISTICA*, select the ***Help*** tab, and in the ***About*** group, click ***STATISTICA*** to access your serial number), information about your hardware [the type of processor (CPU) and the amount of memory (RAM) and disk space], and the version of the operating system that you are using before you contact StatSoft technical support offices.

APPENDIX B

STATISTICA ENTERPRISE SERVER

B

APPENDIX

STATISTICA ENTERPRISE SERVER

STATISTICA Enterprise Server is a highly scalable, enterprise-level, fully Web-enabled data analysis and database gateway application system that is built on distributed processing technology and fully supports multi-tier Client-Server architecture configurations. *STATISTICA Enterprise Server* exposes the analytic, query, reporting, and graphics functionality of *STATISTICA* through easy-to-use, interactive, standard Web interfaces. Alternatively, it enables users of the desktop version ("thick client") to offload computationally intensive analytics and database operations to the Server. It is offered as a complete, ready-to-install application with an interactive, Internet browser-based ("point-and-click") user interface ("thin client") that makes it possible for users to interactively create data sets, run analyses, and review output. However, *STATISTICA Enterprise Server* is built using open architecture and includes .NET-compatible development kit tools (based entirely on industry standard syntax conventions such as VB Script, C++/C#, HTML, Java, and XML) that enables IT department personnel to customize all main components of the system or expand it by building on its foundations, for example, by adding new components and/or company-specific analytic or database facilities.

As mentioned, *STATISTICA Enterprise Server* is provided with an Internet browser-based user interface (in the form of simple-to-navigate and easy-to-use dialogs) enabling you to specify analyses and review results. However, tools are provided to customize these dialogs and easily set up new user interfaces or to add new functions. For example, a simple dialog with only three buttons can be created in the browser, and clicking each button will run a series of analyses and generate a detailed report. *STATISTICA Enterprise Server* applications add a new dimension and an endless array of possibilities to the entire line of *STATISTICA* Data Analysis, Data Mining, and Quality Control/Six Sigma software.

The system is compatible with all major Web server software platforms (e.g., UNIX Apache, and Microsoft IIS), works in both Microsoft .NET and Sun/Java environments, and does not require any changes to the existing firewall and Internet/Intranet security systems.

A Broad Choice of Analytic Facilities and Configurations

The *STATISTICA Enterprise Server* system is offered as a complete solution that includes the analytic functionality of any *STATISTICA* product or any combination of products, from *STATISTICA Base* to *Data Miner* applications.

The minimum installation of *STATISTICA Enterprise Server* software includes the analytic functionality of *STATISTICA Base* and a license for 5 concurrent users (minimum).

Customers can either order a specific version of *STATISTICA Enterprise Server* including the analytic functionality that they require (e.g., *STATISTICA Base* for 10 users), or they can add the *Enterprise Server* functionality (as described in this section) to some or all of the seats of the currently licensed *STATISTICA* product (e.g., add the *Enterprise Server* functionality to 20 out of 50 existing licenses of *STATISTICA Enterprise*).

Functionality and Applications: The Advantages of *STATISTICA Enterprise Server*

A powerful, enterprise-wide collaborative-intelligence system. Another important way to take advantage of the *STATISTICA Enterprise Server* functionality is to use it as the core and natural extension of any of the *STATISTICA* enterprise systems (e.g., *STATISTICA Data Miner* applications).

Specifically, *STATISTICA Enterprise Server* can act as the core of an enterprise-wide network system that enables the participants to work collaboratively and quickly share results (reports), as well as scripts of analyses or queries. User or group permissions (see the Technical Note on page 270) can be used by the administrators to manage access of specific groups of users to specific data or reports. The accessibility of its tools makes *STATISTICA Enterprise Server* a perfect system to facilitate collaborative projects of employees who are telecommuting or traveling.

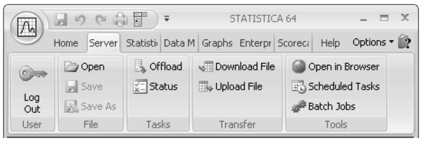
Advantages of distributed processing, and multi-tier Client-Server architecture. Users will benefit not only from the collaborative work tools but also the options to offload the computationally intensive or time-consuming tasks to the server computers. Specifically, because the most powerful multiprocessor CPUs (and/or clusters of computers) are usually used as servers, users can offload computationally intensive tasks, and, for example, run “in the background” queries that will scan terabytes of data on remote servers and perform time-consuming, long sequences of analyses or reports, while keeping the end users’ computers completely free to do other tasks. Because of its distributed processing architecture, *STATISTICA Enterprise Server* scales in a highly efficient manner to take advantage of multi-processor CPUs and/or multiple computers and, therefore, users can take full advantage of multi-tier Client-Server architecture, where:

- Tier 1 is the user interface on the client computer (a plain browser or *STATISTICA* thick client, see *STATISTICA Client*, page 265),
- Tier 2 is the *STATISTICA Enterprise Server* software and the implementation of the “business intelligence” that it may contain (specific queries, scripts of custom/proprietary analyses, etc.), and
- Tier 3 is *STATISTICA* databases (e.g., *STATISTICA Data Warehouse*) or other corporate repositories of data.

In the desktop version of *STATISTICA,* all computations are performed on the local computer, and resources of other computers are used only in the case when the In-Place Database Processing (IDP), see page 245, interface to external databases is established. IDP is a technology that reads data asynchronously directly from remote database servers (using distributed processing if supported by the server), and bypasses the need to “import” data and create a local copy of the data set. Records of data are retrieved and sent to the *STATISTICA* computer asynchronously by the CPU of the database server, while *STATISTICA* simultaneously processes them using the CPU of the local computer.

When a Client-Server version of *STATISTICA* is used, the local computer drives only the user interface of *STATISTICA*, and all calculations are performed on the server. The Client-Server architecture offers obvious advantages when your projects are large (e.g., computationally intensive or involving processing of extremely large data sets) and, thus, when they can be offloaded to the servers, freeing your local computer to perform other jobs.

***STATISTICA* Client.** While no components of the *STATISTICA* system are necessary on the client computer (only a browser), having a copy of *STATISTICA* installed on the client side adds new possibilities. One could ask, *Why would I want to use STATISTICA Enterprise Server if I have a copy of STATISTICA installed on my laptop?* The answer is that having *STATISTICA* installed on the client computer enables you to take additional advantage of the *multi-tier Client-Server* architecture (see page 264) and work interactively with *STATISTICA* installed locally while offloading certain time-consuming tasks to the server machine(s) and/or exchange data and output between all the three tiers. You can run *STATISTICA Enterprise Server* from within desktop *STATISTICA* and flexibly control the interaction between the two. A variety of options are available to share tasks between the desktop and server computer.



Also, when you review your *STATISTICA Enterprise Server* output in the browser, you have options to bring any or all output objects to your desktop computer for further processing. For example, a click on a small button placed optionally (depending on the user configuration) next to every output object (table or graph) sent to your browser by the *STATISTICA Enterprise Server* system will offer you the option to download that object (a *STATISTICA* table or a graph) to the client computer in its native *STATISTICA* format (in *.sta* or *.stg* file format) so you can work with it offline using the locally installed *STATISTICA* tools.

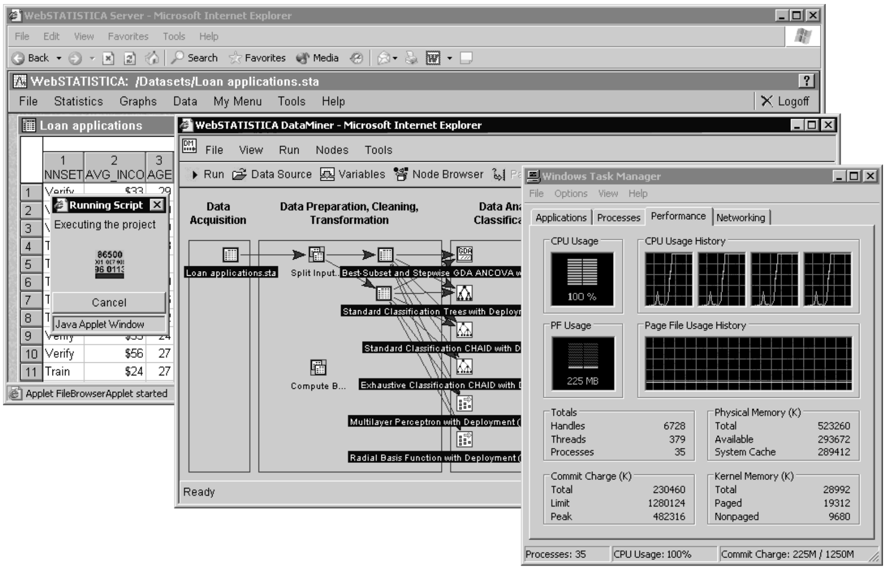
Advantages of Multithreading Technology

The *STATISTICA Enterprise Server* platform is built on advanced distributed processing and multithreading technology to support optimal management of large computational loads. This technology enables rapid processing of even very large and computationally intensive projects, taking full advantage of the multiple CPUs on the server, or even multiple servers working in parallel.

The illustration on the next page shows a project running on a quad processor server, along with the server performance monitor demonstrating the full utilization of the resources of all four CPUs executing in the multithreading mode a single, computationally intensive *STATISTICA Data Miner* project.

In addition, the *STATISTICA Enterprise Server* architecture delivers a platform-independent, Web browser-based user interface, and provides an ultimate, large enterprise-level ability to manage projects or groups of users.

Ultimate scalability (parallel processing technology). One of the unique features of the *STATISTICA* distributed processing technology is that it flexibly scales not only to take advantage of all CPUs on the current server computer (to support both multiple jobs/users and also individual, computationally intensive projects), but it also scales to multiple server computers (clusters). This unique feature is important, since it delivers significant performance gains. *STATISTICA* uses the parallel processing technology across separate hardware units (as some super-computers do) and, therefore, if you have, for example, three servers with four processors each, *STATISTICA* can run an individual project on all 12 processors (if the scale of that project warrants that mode of processing).


WebSTATISTICA Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
WebSTATISTICA: /Datasets/Loan applications.sta
File Statistics Graphs Data My Menu Tools Help
Logoff
Loan applications
1 NNSET
2 AVG_INCO
3 AGE
Running Script
Executing the project
Cancel
Java Applet Window
Verify $56 27
Train $24 27
Applet FileBrowserApplet started
WebSTATISTICA DataMiner - Microsoft Internet Explorer
File View Run Nodes Tools
Run Data Source Variables Node Browser
Data Acquisition
Data Preparation, Cleaning, Transformation
Data Ana Classifica
Loan applications.sta
Split Input...
Best Subset and Stepwise GDA ANCOVA w
Standard Classification Trees with Deploym
Standard Classification CHAID with D
Compute B...
Exhaustive Classification CHAID with
Multilayer Perceptron with Deployment
Radial Basis Function with Deployment
Ready
Windows Task Manager
File Options View Help
Applications Processes Performance Networking
CPU Usage
100 %
CPU Usage History
PF Usage
225 MB
Page File Usage History

Totals
Handles 6728
Threads 379
Processes 35

Physical Memory (K)
Total 523260
Available 293672
System Cache 289412

Commit Charge (K)
Total 230460
Limit 1280124
Peak 482316

Kernel Memory (K)
Total 28992
Paged 19312
Nonpaged 9680

Processes: 35 | CPU Usage: 100% | Commit Charge: 225M / 1250M

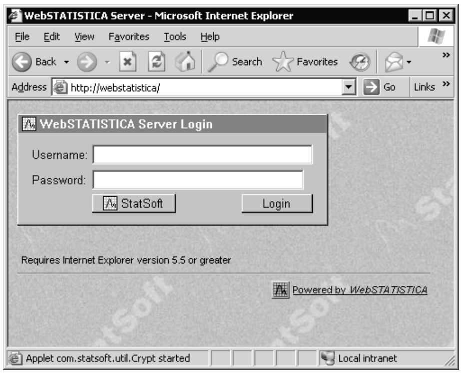
STATISTICA Enterprise Server User Interface

With the *STATISTICA Enterprise Server* implementation of *STATISTICA*, users can interactively run the program from the client machine in a Web browser interface that is similar to that available for the desktop installation. Therefore, the client side of the application (the "front end") can be run on any computer using only a browser. However, the actual computations and other operations performed on the data will remain on the (remote) server with its usually more powerful processors and storage resources (and they will be managed using the optimized, multithreading and distributed processing architecture of the system for maximum performance).

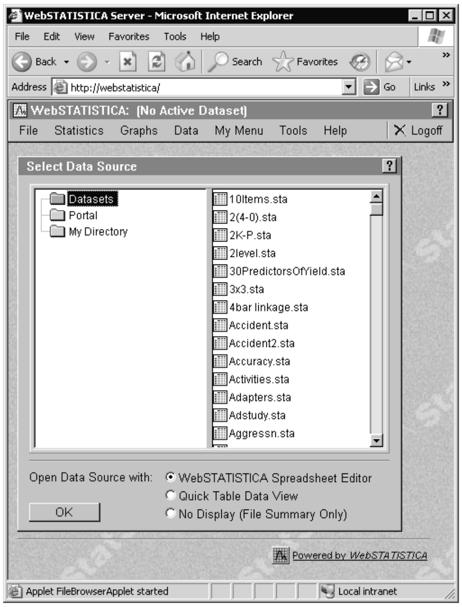
In essence, the user interface aspects of *STATISTICA* can be run by one or multiple users, while the server performs all computations and data operations, enforcing the proper security and access privileges applicable to the respective projects and classes of users, as designed by the network administrator.

STATISTICA Enterprise Server offers a straightforward user interface supporting a selection of interactive data analysis, data mining, quality control, database management, database query, and graph customization operations.

After logging on to the *STATISTICA Enterprise Server* system,

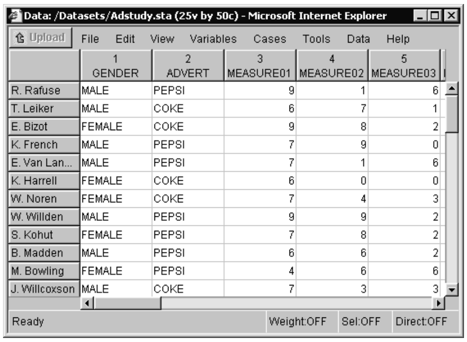

WebSTATISTICA Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://webstatistica/ Go Links
WebSTATISTICA Server Login
Username:
Password:
StatSoft
Login
Requires Internet Explorer version 5.5 or greater
Powered by WebSTATISTICA
Applet com.statsoft.util.Crypt started
Local intranet

you can select a data source (a data set or a live database connection),



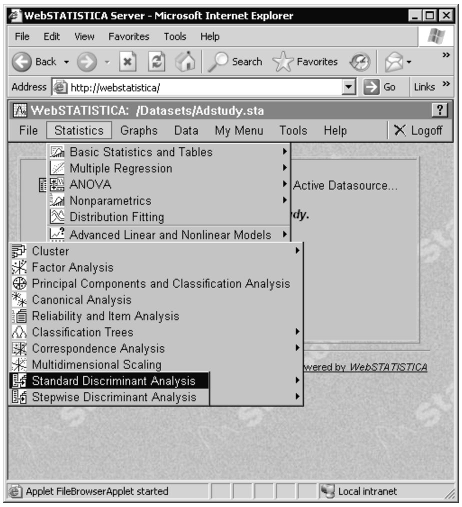
Open Data Source with: ● WebSTATISTICA Spreadsheet Editor ○ Quick Table Data View ○ No Display (File Summary Only)

review and edit the data in the interactive Spreadsheet Editor,

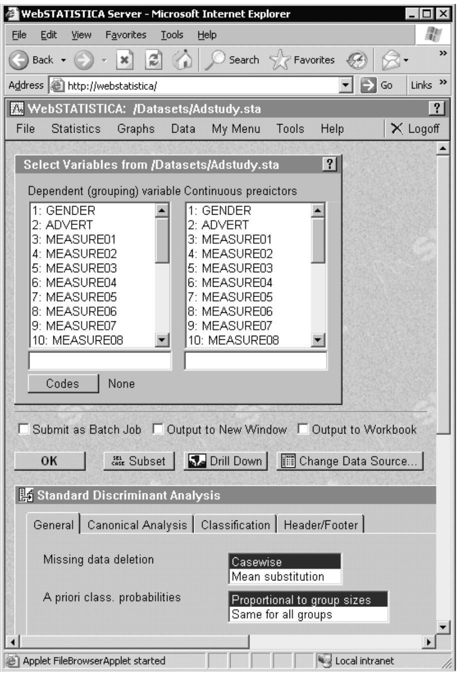

Data: /Datasets/Adstudy.sta (25v by 50c) - Microsoft Internet Explorer
Upload File Edit View Variables Cases Tools Data Help
Ready Weight:OFF Sel:OFF Direct:OFF

	1 GENDER	2 ADVERT	3 MEASURE01	4 MEASURE02	5 MEASURE03
R. Rafuse	MALE	PEPSI	9	1	6
T. Leiker	MALE	COKE	6	7	1
E. Bizot	FEMALE	COKE	9	8	2
K. French	MALE	PEPSI	7	9	0
E. Van Lan...	MALE	PEPSI	7	1	6
K. Harrell	FEMALE	COKE	6	0	0
W. Noren	FEMALE	COKE	7	4	3
W. Willden	MALE	PEPSI	9	9	2
S. Kohut	FEMALE	PEPSI	7	8	2
B. Madden	MALE	PEPSI	6	6	2
M. Bowling	FEMALE	PEPSI	4	6	6
J. Willcoxson	MALE	COKE	7	3	3

select the analysis to be performed using the standard menu system (or a shortcut in the user-defined ***My Menu***),


WebSTATISTICA Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://webstatistica/ Go Links
WebSTATISTICA: /Datasets/Adstudy.sta
File Statistics Graphs Data My Menu Tools Help Logoff
Basic Statistics and Tables
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
Advanced Linear and Nonlinear Models
Cluster
Factor Analysis
Principal Components and Classification Analysis
Canonical Analysis
Reliability and Item Analysis
Classification Trees
Correspondence Analysis
Multidimensional Scaling
Standard Discriminant Analysis
Stepwise Discriminant Analysis
Active Datasource...
Powered by WebSTATISTICA
Applet FileBrowserApplet started
Local intranet

select variables and specify optional analysis parameters,


WebSTATISTICA Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://webstatistica/
Go Links
WebSTATISTICA: /Datasets/Adstudy.sta
File Statistics Graphs Data My Menu Tools Help Logoff

Select Variables from /Datasets/Adstudy.sta

Dependent (grouping) variable	Continuous preaictors
1: GENDER 2: ADVERT 3: MEASURE01 4: MEASURE02 5: MEASURE03 6: MEASURE04 7: MEASURE05 8: MEASURE06 9: MEASURE07 10: MEASURE08	1: GENDER 2: ADVERT 3: MEASURE01 4: MEASURE02 5: MEASURE03 6: MEASURE04 7: MEASURE05 8: MEASURE06 9: MEASURE07 10: MEASURE08

Codes None

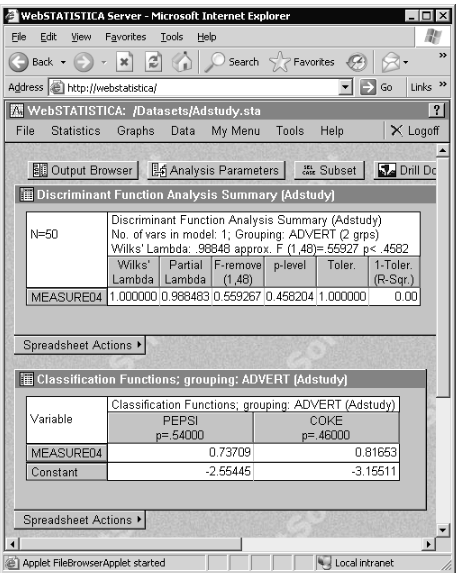
□ Submit as Batch Job □ Output to New Window □ Output to Workbook

Standard Discriminant Analysis

Missing data deletion: Casewise (Mean substitution)

A priori class. probabilities: Proportional to group sizes (Same for all groups)

and interactively review the output.


WebSTATISTICA Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://webstatistica/
Go Links
WebSTATISTICA: /Datasets/Adstudy.sta
File Statistics Graphs Data My Menu Tools Help Logoff
Output Browser
Analysis Parameters
Subset
Drill Dc

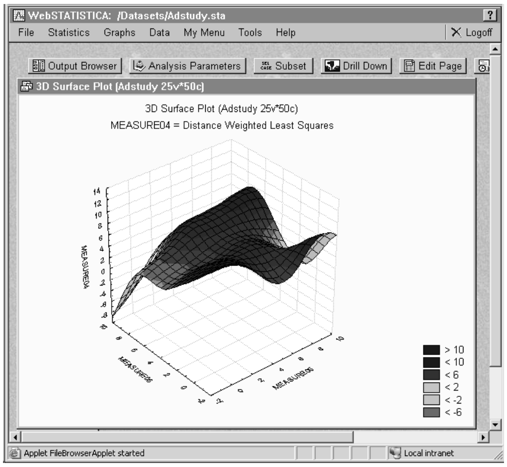
Discriminant Function Analysis Summary (Adstudy)

N=50	Discriminant Function Analysis Summary (Adstudy) No. of vars in model: 1; Grouping: ADVERT (2 grps) Wilks' Lambda: .98848 approx. F (1,48)=.55927 p< .4582					
	Wilks' Lambda	Partial Lambda	F-remove (1,48)	p-level	Toler.	1-Toler. (R-Sqr.)
MEASURE04	1.000000	0.988483	0.559267	0.458204	1.000000	0.00

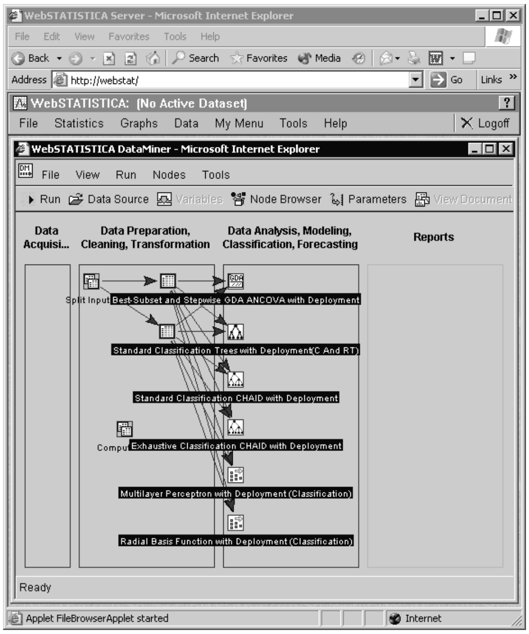
Classification Functions; grouping: ADVERT (Adstudy)

Variable	Classification Functions; grouping: ADVERT (Adstudy) PEPSI p=.54000	COKE p=.46000
MEASURE04	0.73709	0.81653
Constant	-2.55445	-3.15511

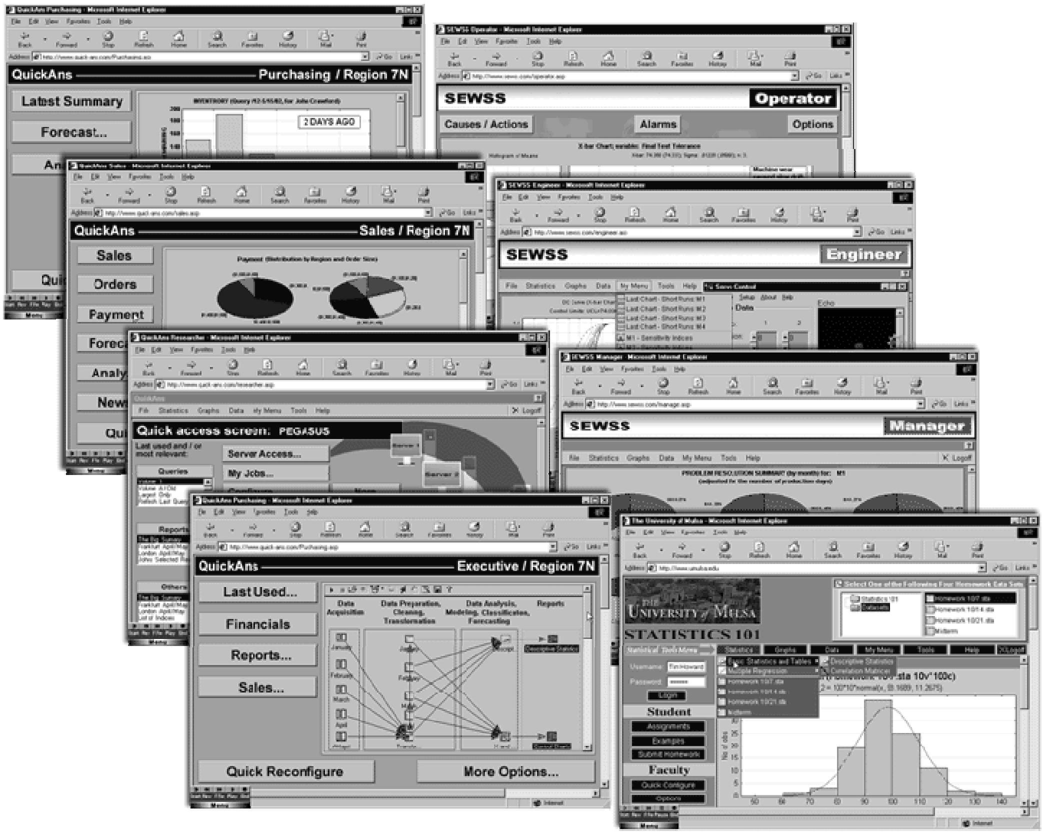
Spreadsheet Actions
Applet FileBrowserApplet started
Local intranet


WebSTATISTICA: /Datasets/Adstudy.sta
File Statistics Graphs Data My Menu Tools Help Logoff
Output Browser
Analysis Parameters
Subset
Drill Down
Edit Page
3D Surface Plot (Adstudy 25v*50c)
3D Surface Plot (Adstudy 25v*50c)
MEASURE04 = Distance Weighted Least Squares
MEASURE04
MEASURE05
MEASURE06
> 10
< 10
< 6
< 2
< -2
< -6
Applet FileBrowserApplet started
Local intranet

A variety of interactive facilities to perform special database, quality control, or data mining operations (including interactively building data mining models by dragging arrows in the model workspace; see below) are provided, and are accessible from the standard browser.


WebSTATISTICA Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://webstat/
Go Links
WebSTATISTICA: (No Active Dataset)
File Statistics Graphs Data My Menu Tools Help Logoff
WebSTATISTICA DataMiner - Microsoft Internet Explorer
File View Run Nodes Tools
Run
Data Source
Variables
Node Browser
Parameters
View Document
Data Acquisi...
Data Preparation, Cleaning, Transformation
Data Analysis, Modeling, Classification, Forecasting
Reports
Split Input
Best Subset and Stepwise GDA ANCOVA with Deployment
Standard Classification Trees with Deployment (C And RT)
Standard Classification CHAID with Deployment
Compu
Exhaustive Classification CHAID with Deployment
Multilayer Perceptron with Deployment (Classification)
Radial Basis Function with Deployment (Classification)
Ready
Applet FileBrowserApplet started
Internet

In addition to these built-in, straightforward user-interface facilities, *STATISTICA Enterprise Server* also includes a toolkit that enables users to customize the user interface and develop custom applications with specifically predefined functionality, packaged in a way that matches the requirements of their specific applications.



Compatibility with Industry Standards

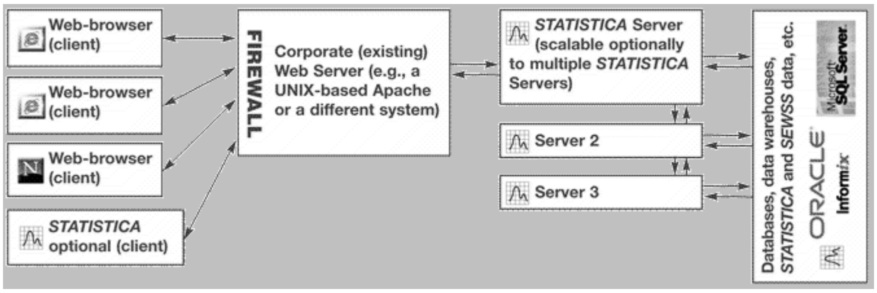
The unsurpassed compatibility with industry standards is another in the long list of unique advantages of *STATISTICA Enterprise Server*.

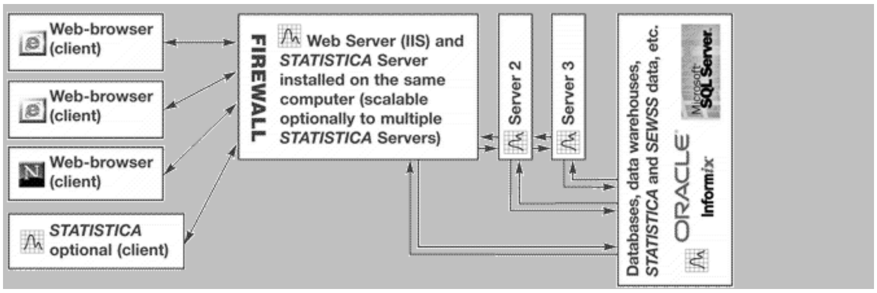
STATISTICA Enterprise Server can be deployed on any of the popular Web server platforms (e.g., a UNIX-based Apache or IIS), and therefore, it will conform to the existing local security protocols (fire walls) as required by the corporate client.

STATISTICA Enterprise Server uses advanced proprietary technology developed at StatSoft to ensure its high performance and scalability (e.g., multiple, multiprocessor *STATISTICA Enterprise Server* computers working in a distributed processing environment). This technology is built on StatSoft's years of experience providing high performance, scalable enterprise systems to major corporations in the United States and around the world. However, *STATISTICA Enterprise Server* is still based on the industry standard communication protocols (e.g., XML) to ensure 1) its platform independence, 2) smooth transition to future technologies, and 3) ease of customization by the client. Note that the ease of customization is additionally boosted by the fact that only the industry standard syntax conventions (such as VB script, C++, HTML, and XML) are used to customize, configure, and define all the specific analytic operations and all output in *STATISTICA Enterprise Server*.

Architecture of the System (A Technical Note)

Although the general design uses two computers in a typical configuration, the Web server (e.g., a UNIX-based Apache system) and at least one *STATISTICA Enterprise Server* (optionally scalable to multiple *STATISTICA Enterprise Servers*), in many cases, *STATISTICA Enterprise Server* could be installed on the same machine if desired (when IIS is used as the Web host):


Web-browser (client)
Web-browser (client)
Web-browser (client)
STATISTICA optional (client)
FIREWALL
Corporate (existing) Web Server (e.g., a UNIX-based Apache or a different system)
STATISTICA Server (scalable optionally to multiple STATISTICA Servers)
Server 2
Server 3
Databases, data warehouses, STATISTICA and SEWSS data, etc.
Microsoft SQL Server
ORACLE
Informix


Web-browser (client)
Web-browser (client)
Web-browser (client)
STATISTICA optional (client)
FIREWALL
Web Server (IIS) and STATISTICA Server installed on the same computer (scalable optionally to multiple STATISTICA Servers)
Server 2
Server 3
Databases, data warehouses, STATISTICA and SEWSS data, etc.
Microsoft SQL Server
ORACLE
Informix

The design allows for a flexible, generic Web server implementation by using a standard scripting language on the Web server. The purpose of the Web server is to package requests from the user (received from a browser), send these to the *STATISTICA Enterprise Server*, and then process responses from the *STATISTICA Enterprise Server* for display to the users (on their browsers).

Communication between the Web server and the *STATISTICA Enterprise Server* is accomplished through technology based on the industry standard XML conventions. The system is fully customizable, and for customers who want to develop their own modifications or extensions of this (ready-to-deploy) system, it provides development tool kit facilities allowing modification of all aspects of both the scripts that are being executed by *STATISTICA* (on the *STATISTICA Enterprise Server* side) and the appearance of the user interface exposed to the end users on the (browser-based) thin client side. Only the most standard, commonly known tools (such as VB or XML/HTML) are used to customize or expand the system.

The actual Web page definitions and *STATISTICA* scripts to be executed are stored in a designated Repository Facility on the *STATISTICA*

Enterprise Server, and they are managed in a queue-like fashion. The system also includes a highly optimized Distributed Processing Manager that handles the incoming processing load and distributes it optimally over multiple threads of *STATISTICA* and multiple *STATISTICA Enterprise Server* computers.

The *STATISTICA Enterprise Server* software system also includes the *STATISTICA* Visual Basic Web Extensions. These extensions to the SVB language enable the script writer to either let the system display the resulting graphs and spreadsheets on the automatically generated (output) Web pages, or customize the appearance of the generated output pages by adding HTML directives as appropriate.

Security and authentication is a key design feature in the *STATISTICA Enterprise Server* application system. At the beginning of the session, users “sign on” to the system with their user name and password. System administrators are able to control access to data sources and scripts based either on user or group permissions. The highest level of the access privilege allows advanced users (or administrators) to execute virtually arbitrary scripts (e.g., in order to perform system administration or maintenance operations). This level requires a designated (highest) access privilege because, due to the general nature and power of the *STATISTICA* Visual Basic language, it gives access (to the authorized users) to all resources on the network.

Note that this system can be integrated with the “traditional” (i.e., non-Web-based) *STATISTICA* concurrent network or a *STATISTICA* enterprise system authentication scheme.

Competitive Advantages

The competitive advantages of *STATISTICA Enterprise Server* applications start with the complete list of unique features of *STATISTICA* itself. Further, unlike the competing products, we offer a complete application (a “solution”) with a Web-based user interface and not merely a “development kit” (although the development kit facilities are also available to extend or customize the system). Also, we do not require that a specific Web server software be installed first (which may or may not comply with the client’s security standards and other policies). Finally, our system is controlled by industry standard VB scripts, C++, HTML, and XML that can be easily modified by users or system administrators. In addition, our distributed processing and multithreading technology delivers performance and system responsiveness that is not matched by any competing products.

Knowledge Portal

A designated *Knowledge Portal* application is optionally available that enables users to effectively and securely distribute organized sets of output documents over the Web. It offers support for workgroups of users (each with different access privileges, and thus access to different parts of the database of output documents), intuitive tree-view organization of available materials, and options to broadcast documents updated on the Web server in real time.

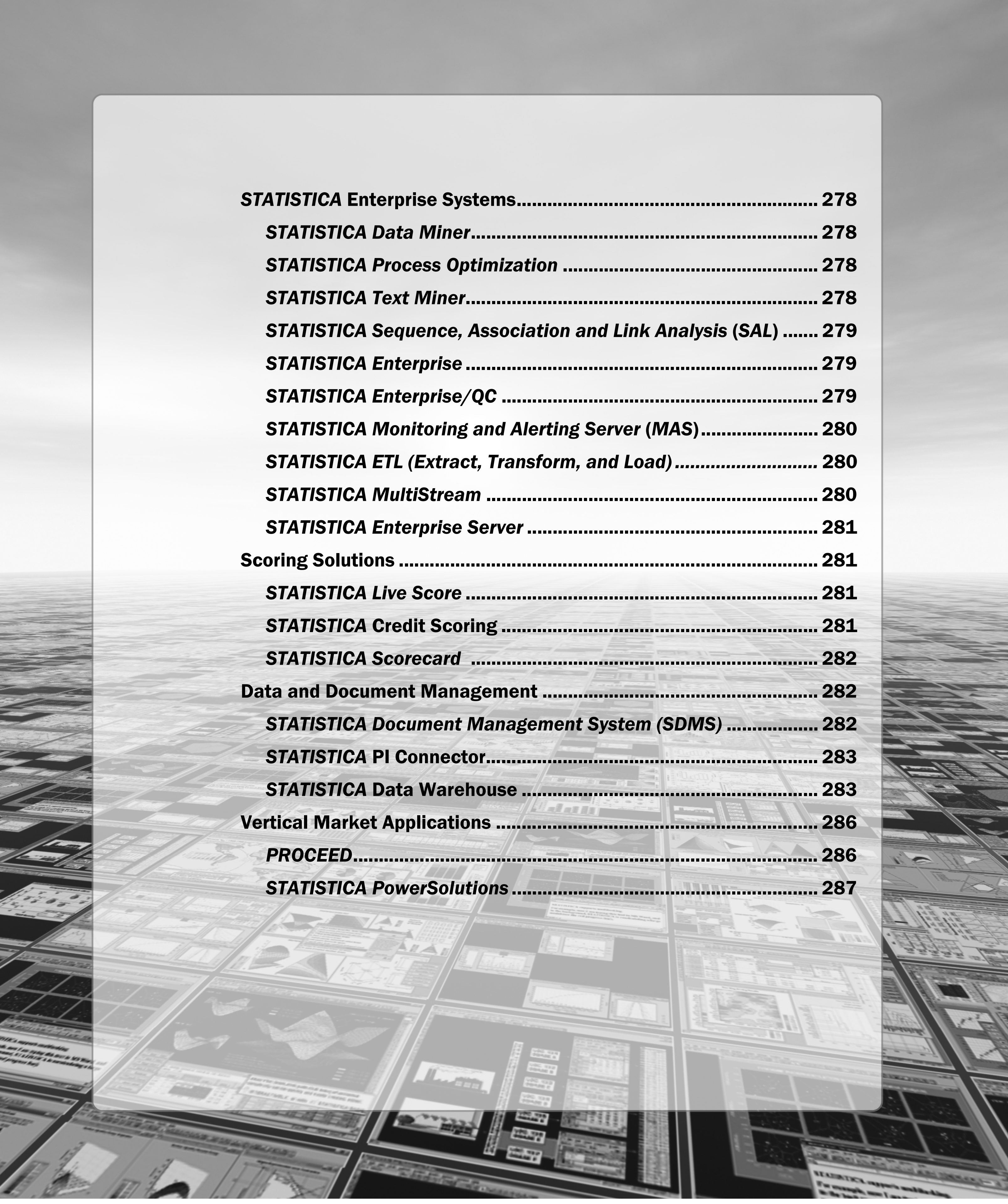
STATISTICA Enterprise Server Demo Movie

How does *STATISTICA Enterprise Server* work? Visit StatSoft’s Web site, ***www.StatSoft.com***, to view an informative presentation of the unique features of *STATISTICA* described here. The movie also includes a step-by-step example application.

C
APPENDIX

STATISTICA FAMILY OF PRODUCTS

continued ➲



C
APPENDIX

STATISTICA FAMILY OF PRODUCTS

Common system features. In addition to comprehensive, leading edge analytics, *STATISTICA* products offer a selection of fully customizable user interfaces (with simplified shortcut templates for novices), flexible, presentation-quality output management (including a variety of report formats, such as *.pdf*, Word, *.rtf*, *.html*, and output to Web portals), full OLE/ActiveX support, and Web enablement.

Also, all products include data management optimized to handle large data sets, interactive database query tools, and a wide selection of data import/export facilities. *STATISTICA* products can manage data sets of practically unlimited size and offer “quadruple” precision calculations; they support multiple input files, multiple instances, and multitasking. A broad selection of interactive visualization and graphics/drawing tools of the highest quality is fully integrated into each product, and each includes a complete set of automation options and a professional Visual Basic and .NET-compatible development environment with more than 14,000 externally accessible functions.

GENERAL-PURPOSE DESKTOP PRODUCTS

***STATISTICA* Base.** Offers a comprehensive set of essential statistics in a user-friendly package and all the performance, power, and ease of use of the *STATISTICA* technology.

- All *STATISTICA* graphics tools
- Basic Statistics, Breakdowns, and Tables
- Distribution Fitting
- Multiple Linear Regression
- Analysis of Variance
- Nonparametrics, and more

STATISTICA Advanced Linear/Nonlinear Models. Offers a wide array of the most advanced modeling and forecasting tools on the market, including automatic model selection facilities and extensive interactive visualization tools.

- General Linear Models
- Generalized Linear/Nonlinear Models
- General Regression Models
- General Partial Least Squares Models
- NIPALS Algorithm (PCA/PLS)
- Variance Components
- Survival Analysis
- Cox Proportional Hazards Models

- Nonlinear Estimation
- Fixed Nonlinear Regression
- Log-Linear Analysis of Frequency Tables
- Time Series/Forecasting
- Structural Equation Modeling, and more

STATISTICA Multivariate Exploratory Techniques. Offers a broad selection of exploratory techniques for various types of data, with extensive, interactive visualization tools.

- Cluster Analysis
- Factor Analysis
- Principal Components/Classification Analysis
- Canonical Analysis
- Discriminant Analysis
- General Discriminant Analysis Models
- Reliability/Item Analysis
- Classification Trees
- Correspondence Analysis
- Multidimensional Scaling, and more

STATISTICA Variance Estimation and Precision. A comprehensive set of techniques for analyzing data from experiments that include both fixed and random effects using REML (Restricted Maximum Likelihood Estimation). With *STATISTICA Variance Estimation and Precision*, you can obtain estimates of variance components and use them to make precision statements while at the same time comparing fixed effects in the presence of multiple sources of variation.

- Variability plots
- Multiple plot layouts to allow direct comparison of multiple dependent variables
- Expected mean squares and variance components with confidence intervals
- Flexible handling of multiple dependent variables: analyze several variables with the same or different designs at once
- Graph displays of variance components

STATISTICA Automated Neural Networks (SANN). Contains the most comprehensive neural network algorithms and training methods.

- Automatic search for best architecture and network solutions
- Multilayer Perceptrons
- Radial Basis Function Networks
- Self-Organizing Feature Maps
- Time Series Neural Networks for both Regression and Classification problems
- A variety of algorithms for fast and efficient training of Neural Network Models including Gradient Descent, Conjugate Gradient, and BFGS
- Numerous analytical graphs to aid in generating results and drawing conclusions
- Sampling of data into subsets for optimizing network performance and enhancing the generalization ability
- Sensitivity Analysis, Lift Charts, and ROC Curves
- Creation of Ensembles out of already existing standalone networks
- C/C++/C#, PMML (Predictive Model Markup Language), Java, *STATISTICA Enterprise*, and SAS Neural Network Code Generators that are easy to deploy

STATISTICA Power Analysis. An extremely precise and user-friendly specialized tool for analyzing all aspects of statistical power and sample size calculation.

- Sample Size Calculation
- Confidence Interval Estimation
- Statistical Distribution Calculators, and more

INDUSTRIAL SOLUTIONS, SIX SIGMA TOOLS

STATISTICA Quality Control Charts. Offers fully customizable (e.g., callable from other environments), easy and quick to use, versatile charts with a selection of automation options and user-interface shortcuts to simplify routine work (a comprehensive tool for Six Sigma methods).

- Multiple Chart (Six Sigma Style) Reports and displays
- X-bar and R Charts; X-bar and S Charts; N_p, P, U, C Charts
- Pareto Charts
- Process Capability and Performance Indices
- Moving Average/Range Charts, EWMA Charts
- Short Run Charts (including Nominal and Target)
- CuSum (Cumulative Sum) Charts
- Runs Tests
- Interactive
- Causes and actions, customizable alarms, analytic brushing, and more

STATISTICA Process Analysis. A comprehensive package for Process Capability, Gage R&R, and other quality control/improvement applications (a comprehensive tool for Six Sigma methods).

- Process/Capability Analysis Charts
- Ishikawa (Cause and Effect) Diagrams
- Gage Repeatability & Reproducibility
- Variance Components for Random Effects
- Weibull Analysis
- Sampling plans, and more

STATISTICA Design of Experiments. Features the largest selection of DOE and related visualization techniques including interactive desirability profilers (a comprehensive tool for Six Sigma methods).

- Fractional Factorial Designs
- Mixture Designs
- Latin Squares
- Search for Optimal $2^{(k-p)}$ Designs
- Residual Analysis and Transformations
- Optimization of single/multiple response variables
- Central Composite Designs
- Taguchi Designs
- Minimum Aberration & Maximum Unconfounding
- $2^{(k-p)}$ Fractional Factorial Designs with Blocks
- Constrained Surfaces
- D- and A-Optimal Designs
- Desirability profilers, and more

STATISTICA Multivariate Statistical Process Control (MSPC). A complete solution for multivariate statistical process control, deployed within a scalable, secure analytics software platform.

- Univariate and multivariate statistical methods for quality control, predictive modeling, and data reduction
- Functions to determine the most critical process, raw materials, and environment factors and their optimal settings for delivering products of the highest quality
- Monitoring of process characteristics interactively or automatically during production stages
- Building, evaluating, and deploying predictive models based on the known outcomes from historical data
- Historical analysis, data exploration, data visualization, predictive model building and evaluation, model deployment to monitoring server
- Interactive monitoring with dashboard summary displays and automatic-updating results
- Automated monitoring with rules, alarm events, and configurable actions
- Multivariate techniques including Partial Least Squares, Principal Components, Neural Networks, Recursive Partitioning (Tree)

Methods, Support Vector Machines, Independent Components Analysis, Cluster Analysis, and more

STATISTICA ENTERPRISE SYSTEMS

In addition to the common features, *STATISTICA* Enterprise Systems optionally offer a wide selection of tools for collaborative work, Web browser-based user interfaces (using *STATISTICA Enterprise Server*), specialized databases, and a highly optimized interface to enterprise-wide data repositories, including options to rapidly process large data sets from remote servers in-place, without creating local copies. Deployment and on-site training services are available.

STATISTICA Data Miner. The most comprehensive selection of data mining solutions on the market, with an icon-based, extremely easy-to-use user interface (optionally Web browser based via *STATISTICA Enterprise Server*, see page 281) and a deployment engine. It features a selection of completely integrated and automated, ready-to-deploy "as is" (but also easily customizable) systems of specific data mining solutions for a wide variety of business applications. A designated SPC version (*QC Data Miner*) to mine/analyze large streams of QC data is also available. The data mining solutions are driven by powerful procedures from five modules:

- General Slicer/Dicer Explorer (with optional OLAP)
- General Classifier
- General Modeler/Multivariate Explorer
- General Forecaster
- General Neural Networks Explorer, and more

STATISTICA Process Optimization. An add-on to *Data Miner*, *STATISTICA Process Optimization* is a powerful software solution designed to monitor processes and identify and anticipate problems related to quality control and improvement with unmatched sensitivity and effectiveness. *Process Optimization* integrates all *Quality Control Charts*, *Process Capability* analyses, *Experimental Design* procedures, and Six Sigma methods with a comprehensive library of cutting-edge techniques for exploratory and predictive data mining.

- Predict QC problems with cutting edge data mining methods
- Discover root causes of problem areas
- Monitor and improve ROI (Return On Investment)
- Generate suggestions for improvement
- Monitor processes in real time over the Web
- Create and deploy QC/SPC solutions over the Web
- Use multithreading and distributed processing to rapidly process extremely large streams of data

STATISTICA Text Miner. A powerful software solution for text mining, document retrieval, and mining of unstructured data. An optional add-on product for *STATISTICA Data Miner,* designed and optimized for accessing and analyzing documents (unstructured information) in a variety of formats: *.txt* (text), *.pdf* (Adobe), *.ps* (PostScript^{TM}), *.html*, *. xml* (Web-formats), and most Microsoft Office formats (e.g., *.doc, .rtf*); optimized access to Web pages (URL addresses) is also provided.

- Efficiently index very large collections of text documents; identify key terms and similarities between documents and terms, and extract the information relevant to your mission and goals

- Apply stub-lists (words to ignore) and language-specific stemming algorithms (various languages are supported)
- Includes numerous options for converting documents into numeric information for further processing (e.g., mapping, clustering, predictive data mining, classification of documents, etc.)
- Full support for multithreaded operation on multi-processor server installations for extremely fast indexing and searching of huge document repositories
- Can also be used to index, analyze, and mine other unstructured input, such as sound or image files (after domain-specific pre-processing is applied)
- Fully integrated into the *STATISTICA* and *STATISTICA Enterprise Server* systems, hence, the large number of available methods for supervised and unsupervised learning (clustering), mapping, data visualization, etc., are directly and immediately available; many of the algorithms available in *STATISTICA Data Miner*, such as the machine learning algorithms (*k*-Nearest Neighbor, Naive Bayes classifiers, advanced Support Vector Machines and Kernel classifiers), are particularly well suited for text mining or the analysis of other unstructured information

STATISTICA Sequence, Association and Link Analysis (SAL). Designed to address the needs of clients in retailing, banking, insurance, etc., industries by implementing the fastest known highly scalable algorithm with the ability to drive Association and Sequence rules in one single analysis. The program represents a stand-alone module that can be used for both model building and deployment. All tools in *STATISTICA Data Miner* can be quickly and effortlessly leveraged to analyze and "drill into" results generated via *STATISTICA SAL*.

- Uses a tree-building technique to extract Association and Sequence rules from data
- Uses efficient and thread-safe local relational database technology to store Association and Sequence models
- Handles multiple response, multiple dichotomy, and continuous variables in one analysis
- Performs Sequence Analysis while mining for Association rules in a single analysis
- Simultaneously extracts Association and Sequence rules for more than one dimension
- Given the ability to perform multidimensional Association and Sequence mining and the capacity to extract only rules for specific items, the program can be used for Predictive Data Mining
- Performs Hierarchical Single-Linkage Cluster Analysis, which can detect the more likely cluster of items that can occur. This has extremely useful, practical real-world applications, e.g., in retailing.

STATISTICA Enterprise. An integrated multi-user system designed for general-purpose data analysis and business intelligence applications in research. *STATISTICA Enterprise* can optionally offer the statistical functionality available in any or all *STATISTICA* products.

- Integration with data warehouses
- Intuitive query and filtering tools
- Easy-to-use administration tools
- Automatic report distribution
- Alarm notification, and more

STATISTICA Enterprise/QC. Designed for local and global enterprise quality control and improvement applications including Six Sigma. *STATISTICA Enterprise/ QC* offers a high-performance database (or an optimized interface to existing databases), real-time monitoring and alarm notification for the production floor, a comprehensive set of analytical tools for engineers, sophisticated reporting features for management, Six Sigma reporting options, and much more.

- Web-enabled user interface and reporting tools; interactive querying tools
- User-specific interfaces for operators, engineers, etc.
- Groupware functionality for sharing queries, special applications, etc.
- Open-ended alarm notification including cause/action prompts
- Scalable, customizable, and can be integrated into existing database/ERP systems, and more

STATISTICA Monitoring and Alerting Server (MAS). A system that enables users to automate the continual monitoring of hundreds or thousands of critical process and product parameters. The ongoing monitoring is an automated and efficient method for:

- Monitoring many critical parameters simultaneously
- Providing status “snapshots” from the results of these monitoring activities to personnel based on their responsibilities
- Dashboards associated with Users/Groups

STATISTICA ETL (Extract, Transform, and Load). Provides options to simplify and facilitate access to, aggregation, and alignment of data from multiple databases when some of the databases contain process data (using the optional PI Connector) while others contain “static” data (e.g., from Oracle or MS SQL Server). Provides for ad-hoc querying and aligning of data for subsequent analyses such as ad-hoc charting, etc., of data describing a specific time interval.

- Time-indexed *STATISTICA ETL* aggregates data from multiple data sources based on a date/time stamp variable. Data may be aligned by minute, hour, day, week, month, quarter, or year.
- ID-Based *STATISTICA ETL* aggregates data from multiple data sources based on an identifier variable (either number or text) and an optional time variable. If a time variable is defined, data may be optionally aligned by *N* equal intervals or *N* user-specified intervals.

STATISTICA MultiStream. A solution package for identifying and implementing effective strategies for advanced multivariate process monitoring and control. *STATISTICA MultiStream* was designed for process industries in general, but is particularly well suited to help power generation facilities leverage their data (collected into existing specialized process data bases for multivariate and predictive process control) for actionable advisory systems.

STATISTICA MultiStream is a complete enterprise system built on a robust, advanced client-server (and fully Web-enabled) architecture, offers central administration and management of deployment of models, as well as cutting edge root-cause analysis and predictive data mining technology, and its analytics are seamlessly integrated with a built-in document management system.

- Automated (nonlinear) root cause analysis and feature selection for thousands of parameters to clearly identify which ones are the most likely responsible for process problems
- Automated and interactive commonality analysis to identify parameters and processes that shifted or moved from

normal operations during particular time intervals

- Advanced linear and nonlinear (e.g., SVM, Recursive Partitioning, Neural Nets) models for creating sensitive multivariate control schemes and work flows to identify multivariate shifts and drifts early, before they cause problems
- Advanced data mining algorithms for predicting and optimizing key performance and quality indicators
- Tracks hundreds of data streams simultaneously
- Delivers simple summaries relevant to critical process parameters and outcomes via efficient and simple dashboards and drill-down workflows
- Delivers standard and customized analytic workflows for root cause analysis, leveraging cutting-edge data analysis and data mining technologies
- Warns of (predicted) problems and equipment failures before they occur (predictive alarming), thus avoiding costly shut-downs and unscheduled maintenance
- Watches "everything" that impacts your process performance in real time

STATISTICA Enterprise Server. The ultimate enterprise system that offers full Web enablement, including the ability to run *STATISTICA* interactively or in batch from a Web browser on any computer (including Linux, UNIX) and offload time-consuming tasks to the servers (using distributed processing). Uses multi-tier Client-Server architecture, supporting multithreading and distributed/ parallel processing that scales to multiple server computers.

SCORING SOLUTIONS

STATISTICA Live Score. *STATISTICA Enterprise Server* software within the *STATISTICA* Data Analysis and Data Mining Platform. Data are aggregated and cleaned and models are trained and validated using the *STATISTICA Data Miner* software. Once the models are validated, they are deployed to the *STATISTICA Live Score* server. *STATISTICA Live Score* provides multi-threaded, efficient, and platform-independent scoring of data from line-of-business applications. Some examples of the use of *STATISTICA Live Score*:

- Provides credit scorecards to customer service applications (e.g., call center systems and Web-based applications)
- Enables customer segmentation, up-sell/cross-sell, and customer churn identification to customer service and marketing representatives
- Provides proactive fraud detection alerts to analysts

STATISTICA Credit Scoring. The solution for any company to build in-house models for its various credit products and decision making. *STATISTICA Credit Scoring* covers all aspects of the credit scoring needs for your company.

- **In-house model building.** The *STATISTICA Credit Scoring* software solution enables the development and evaluation of predictive models to evaluate and assign a risk to applications for credit, either for a request for a new account or for requested changes (e.g., balance increase) to the terms of an existing credit account.
- **Scoring applications.** *STATISTICA Live Score* enables companies to score credit applications; it can be easily integrated with your existing customer service systems, self-service Websites for customers, etc.

- **Evaluate performance.** *STATISTICA Credit Scoring* provides built-in monitoring and evaluation of the ongoing performance of the models to enable the evaluation of outcomes and key metrics and to make decisions about when models may need to be updated.

What makes the *STATISTICA Credit Scoring* solution unique?

- **The Approach.** *STATISTICA Credit Scoring* includes both traditional methods for developing credit scoring models (such as scorecards based on logistic regression) as well as more advanced methods for predictive modeling that often provide better accuracy, which translates into decreased risk, increased approval rates, and increased profits. *STATISTICA Credit Scoring* includes *STATISTICA Scorecard*, a dedicated solution for development, evaluating, and monitoring scorecards including steps for Feature Selection, Attribute Building, Scorecard Building, Cutoff Point Selection, Reject Inference, and Population Stability.
- **Real-time Scoring.** *STATISTICA Credit Scoring* includes *STATISTICA Live Score*, the solution for enabling scoring decisions directly from customer applications via Customer Service Agents, Websites, and other line of business systems.
- **Sources of Data.** Unlike generic scorecards, *STATISTICA Credit Scoring* can be tailored to meet your specific needs. For example, it provides the flexibility to include various data sources such as behavior scoring, utilizing the transactional record of the account to inform recommendations for credit line increases, incentives, cross-sell or up-sell, or other changes in terms.
- **Flexibility and Capabilities.** *STATISTICA Credit Scoring* is specific to building credit scoring models, but the same approaches and techniques can also be applied to modeling customer churn, increasing the ability to detect fraud, response modeling for marketing campaigns, and other applications within your company.

STATISTICA Scorecard. *STATISTICA Scorecard* is a dedicated solution for developing, evaluating, and monitoring scorecards, including steps for Feature Selection, Attribute Building, Scorecard Building, Cutoff Point Selection, Reject Inference, and Population Stability.

DATA AND DOCUMENT MANAGEMENT

STATISTICA Document Management System (SDMS). A complete, highly scalable, database solution package for managing electronic documents. With the *STATISTICA Document Management System*, you can quickly, efficiently, and securely manage documents of any type [e.g., find them, access them, search for content, review, organize, edit (with trail logging and versioning), approve, etc.].

- Extremely transparent and easy to use
- Flexible, customizable (optionally browser/Web-enabled) user interface
- Electronic signatures
- Comprehensive auditing trails, approvals
- Optimized searches
- Document comparison tools
- Security
- Satisfies the FDA 21 CFR Part 11 requirements
- Satisfies ISO 9000 (9001, 14001) documentation requirements
- Unlimited scalability (from desktop or network Client-Server versions, to the

ultimate size, Web-based worldwide systems)
- Open architecture and compatibility with industry standards

STATISTICA PI Connector. Allows for direct integration to data stored in the PI data historian. The *STATISTICA PI Connector* utilizes the PI user access control and security model, allows for interactive browsing of tags, and takes advantages of dedicated PI functionality for interpolation and snapshot data. *STATISTICA* integrated with the PI system is being used for streamlined and automated analyses for applications such as Process Analytical Technology (PAT) in FDA-regulated industries, Advanced Process Control (APC) systems in Chemical and Petrochemical industries, and advisory systems for process optimization and compliance in the Energy Utility industry.

STATISTICA Data Warehouse. A complete, powerful, scalable, and customizable intelligent data warehouse solution, which also optionally offers the most complete analytic functionality available on the market, fully integrated into the system. *STATISTICA Data Warehouse* consists of a suite of powerful, flexible component applications, including:

- *STATISTICA* Data Warehouse Server Database
- *STATISTICA* Data Warehouse Query (featuring *STATISTICA Enterprise Server Query)*
- *STATISTICA* Data Warehouse Analyzer (featuring *STATISTICA Enterprise Server Data Miner*, *STATISTICA Enterprise Server Text Miner*, *STATISTICA Enterprise Server Process Optimization*, or the complete set of *STATISTICA Enterprise Server* analytics)
- *STATISTICA* Data Warehouse Reporter (featuring *STATISTICA Knowledge Portal* and/or *STATISTICA Enterprise Server* Interactive Knowledge Portal)
- *STATISTICA* Data Warehouse Document Repository (featuring *STATISTICA Enterprise Server Document Management System*)
- *STATISTICA* Data Warehouse Scheduler
- STATISTICA Data Warehouse Real Time Monitor and Reporter (featuring *STATISTICA Enterprise Server* or *STATISTICA Enterprise/QC Server*)

If you are new to data warehousing, StatSoft consultants will guide you step by step through the entire process of designing the optimal data warehouse architecture – from a comprehensive review of your information storage and extraction/analysis needs, to the final training of your employees and support of your daily operations.

Crucial features and benefits. The crucial features and benefits of *STATISTICA Data Warehouse* solutions include, among many others:

- Complete data warehousing application tailored to your business
- Platform independent architecture for seamless integration with your existing infrastructure
- Facilities to integrate data from a wide variety of sources
- Virtually unlimited scalability
- Options to update/synchronize data from multiple sources via automatic schedulers or on demand
- Completely Web-enabled system architecture to provide ultimate enterprise functionality for all company locations around the world (e.g., access via Web browsers from any location)

- Advanced security model and authentication of users
- Complete document management options to optimize management of documents of any types and satisfy regulatory requirements (e.g., FDA 21 CFR Part 11, ISO 9000)
- Advanced analytic components to clean/verify data and to integrate automated data mining, artificial intelligence, and real-time process monitoring
- Options to automatically run and post on Knowledge Portals (or broadcast) highly customized reports, including interactive (i.e., drillable, sliceable, and user-customizable) reports and results of advanced analytics
- Backup and archiving options
- Programmable, customizable, and expandable to adapt to specific mission profiles (open architecture, exposed to extensions using the most industry standard languages, such as VB, C++, Java, HTML)
- Built on robust, well tested, highly scalable, cutting-edge technology to leverage your investment [including highly optimized in-place database processing (IDP) technology, true multithreading, distributed/parallel processing, and support for pooling CPU resources of multiple servers to deliver supercomputer-like performance]

STATISTICA Data Warehouse is a complete intelligent data storage and information delivery/distribution solution that enables you to customize the flow of information through your organization and provide all authorized members of your organization with flexible, secure, and rapid access to critical information and intelligent reporting.

The system is virtually platform independent and will fit into any existing database architecture and hardware environment. It will efficiently combine information from multiple database formats and sources (from manual data entry forms to large batteries of automatic data collection devices). The system can be further enhanced through integration with other fully compatible components of the *STATISTICA* line of applications and solutions; to name just a few:

- *STATISTICA Data Miner* for advanced data mining and artificial intelligence (e.g., neural networks) based solutions to provide decision support through cutting-edge methods for knowledge extraction and prediction
- *Quality Control Miner* and *Enterprise/QC* for tight integration with quality control, process control, and yield management activities
- *STATISTICA Text Miner* for automatic processing of unstructured information in documents, databases, or Web directories (Web-crawling of URLs)
- *STATISTICA Knowledge Portal* for presenting summary reports, charts, and action items to end users (management, sales force, engineers, etc.) through secure access portals via the Web; to deliver key intelligence and decision support to stakeholders worldwide

Architecture and connectivity. *STATISTICA Data Warehouse* connects to any platform, database, or data source, and will scale to businesses and applications of any size. The program is built on a database and database schema customized for your particular business. The solution can be installed either inclusive of a high performance database engine (SQL Server) or as a (virtual) database schema compatible with most industry standard databases; therefore, it will seamlessly integrate into existing database systems.

Because *STATISTICA Data Warehouse* does not depend on one particular database vendor or hardware platform, it is itself entirely platform-independent. The main *Data Warehouse* software will connect to any database format and, hence, can efficiently combine and pool information from multiple sources.

STATISTICA Data Warehouse application software will run on servers with multiple processors or banks of multiple-processor servers for super-computer like performance. The system will scale effortlessly and economically to even huge data sizes and analysis (intelligence) problems.

Web enablement. *STATISTICA Data Warehouse* extracts information from sources anywhere in the world and delivers intelligence anywhere in the world.

The Web component of the system is built on the proven *STATISTICA Enterprise Server* technology that is used by organizations worldwide to provide secure access via standard Web browsers. Unlike other Web-based solutions, *STATISTICA Data Warehouse* does not require any additional components to be installed on the (thin) client machines.

Advanced security and authentication. The *STATISTICA Data Warehouse* implements a detailed and sophisticated security system to ensure that your proprietary knowledge and intelligence is safe from unauthorized access. The system will likely become the most important repository of business intelligence and decision support resources in your organization. Therefore, the security of the system is a crucial priority so that those valuable resources are shielded from unauthorized access.

STATISTICA Data Warehouse implements the highest level of security by establishing groups of users with different levels of authority (regarding the information that is accessible and the operations that can be performed), requiring regularly updated passwords, etc. Special methods are also in place to detect and guard against systematic electronic intrusions (“hacking”).

Document control. *STATISTICA Data Warehouse* enables full document management, compliant with government and industry standards.

STATISTICA Document Management System can be seamlessly integrated into your *STATISTICA Data Warehouse* application to optimize the flow of information within your organization and thus increase your productivity. This system can also be configured to comply with all (corporate) documentation management policies or regulatory requirements for document security, audit trails, and electronic signatures/authentication (as, for example, stipulated by FDA 21 CFR Part 11: *Electronic Records; Electronic Signatures*; or ISO 9001 4.5: *Document and data control*).

Advanced analytics. *STATISTICA Data Warehouse* can incorporate the most advanced data analysis and knowledge extraction methods available; you can go far beyond OLAP to simplify and extract knowledge about even the most complex – and inaccessible to other applications – patterns in the data.

Because *STATISTICA Data Warehouse* is built from the same high performance components as the entire *STATISTICA* line of analytic solutions software, those analytic solutions can easily and seamlessly be integrated into your Data Warehouse. *STATISTICA* offers the most comprehensive set of tools for data mining, text mining, data analysis, graphics and visualization, quality and process control (including Six Sigma), etc. on the market. These resources and technologies can be connected to the data sources in the *STATISTICA Data Warehouse* to leverage the most advanced technologies and algorithms available for analyzing and extracting key intelligence from all sources. For example, you can apply hundreds of neural networks architectures, highest performance tree classifiers (e.g., stochastic gradient boosting trees), flexible root cause analyses, control charting methods, powerful business forecasting methods, or sophisticated analytic graphics methods to convert raw data in the Data Warehouse into useful and actionable intelligence with clear implications for decisions affecting your business.

Programmability and customizability. *STATISTICA Data Warehouse* is an open-architecture system that will not lock you into a relation with a single vendor or solution; you can respond quickly to new business demands and requirements that need to be incorporated into the Data Warehouse.

As with all applications and solutions in the *STATISTICA* family of products, *STATISTICA Data Warehouse* is fully programmable and customizable, using industry standard programming tools such as Visual Basic, C++, Java, or HTML. This feature is of key importance when your business depends on your ability to quickly adapt to new information and business realities. Because you can customize the system without being forced to rely on the programmers of a single vendor or knowledge of idiosyncratic scripting conventions (required by many competing solutions), you have the freedom to develop your proprietary extensions to the Data Warehouse and to add not only your own reports but also custom analytic and data transformation/cleaning procedures, using widely available resources and industry standard tools (e.g., VB, C++, Java, or HTML tools and programmers). Of course, StatSoft can always offer you a full complement of consulting, system integration, and programming services delivered by an experienced staff.

VERTICAL MARKET APPLICATIONS

PROCEED. A turnkey manufacturing software solution that distills fundamental causal relationships between products and the processes that produce them, using data that is already collected and managed. *PROCEED* implements the patent-pending approach developed and proven at Caterpillar Inc. and powered by the *STATISTICA* Enterprise Analytics Software Platform.

High tech manufacturing enterprises today collect vast amounts of data.

- Data about the production processes.
- Data about tests of raw materials, subassemblies, and materials in process.
- Data about the critical to quality attributes of finished products.

All of these data collection and storage efforts continue to be fueled by increases in automation, technology advances in the storage capabilities of data repositories, and the advances in sensors and other techniques for measurement. Today's manufacturers are sitting on a gold mine of information . . . only if they are able to translate it into actionable information.

Collecting data is not sufficient to drive enterprise change. To create change, we need to translate these data into knowledge and then communicate that knowledge in a format that enables the people who are empowered to act on it. Now is the time for this Return on Investment from data using *PROCEED*.

PROCEED combines novel and traditional knowledge extraction methods to:

- Derive and validate simple to complex causal relationships between manufacturing processes and product quality outcomes
- Deploy actionable information to enable process owners and knowledge workers to compare what-if scenarios and simultaneously optimize multiple competing outcomes

STATISTICA PowerSolutions. A solution package aimed for use at power generation companies to optimize power plant performance, increase efficiency, and reduce emissions. This product offers a highly economical alternative to multimillion dollar investments in new or upgraded equipment (hardware). Based on more than 20 years of experience in applying advanced data driven, predictive data mining/optimization technologies for process optimization in various industries, *STATISTICA PowerSolutions* enables power plants to get the most out of their existing equipment and control systems by leveraging all data collected at their sites to identify opportunities for improvement, even for older designs such as coal-fired Cyclone furnaces (as well as wall-fired or T-fired designs).

INDEX