STATISTICA 9.0

NEW FEATURES

Release Date: May 2009

A PDF of this document is available on your installation CD.

Even Faster!

Performance Improvements

STATISTICA 9 offers many new, unique features and added functionality to the entire *STATISTICA* product line. Many low-level optimizations have been implemented that greatly improve the overall performance of *STATISTICA*. These improvements can be found in both the 32-bit version and the new native 64-bit version of *STATISTICA*.

Because of the new technology and a variety of optimizations introduced to the main computational kernel of *STATISTICA*, not only the 64-bit version but also the 32-bit version of *STATISTICA 9* is significantly faster in most operations when compared to *STATISTICA 8* (which already was one of the fastest analytic applications on the market).

For example:

- Routine breakdown analyses (computing descriptive statistics broken down by hundreds or thousands of categories) can be twice as fast as before. In some cases, the performance is hundreds of times faster when the data include very long text variables.
- Decision Tree analyses (C&RT) run more than twice as fast.
- Neural Networks is almost three times faster depending on the specific neural network architecture.

Along with these enhancements to the analytic procedures, the spreadsheet performance has been further optimized to minimize needed computing resources when working with large data files.

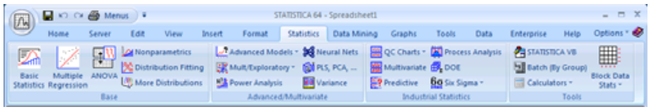
Native 64-Bit

Native 64-bit applications take full advantage of the 64-bit Operating Systems, allowing for better memory management and performance. Consequently, *STATISTICA 9* 64-bit can process designs of an increased size and further improves performance, more than doubling the speed of some computationally intensive analyses when compared to *STATISTICA 9* 32-bit. *STATISTICA* 64-bit is especially useful in data mining and other operations that use extremely large data sets and iterative, computationally demanding applications.

User Interface Enhancements

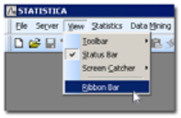
Enhanced User Experience

Application navigation is simpler and more intuitive with the addition of an Office 2007-style ribbon bar. Frequently used functionality is quickly visible, and related functionality is easily found.



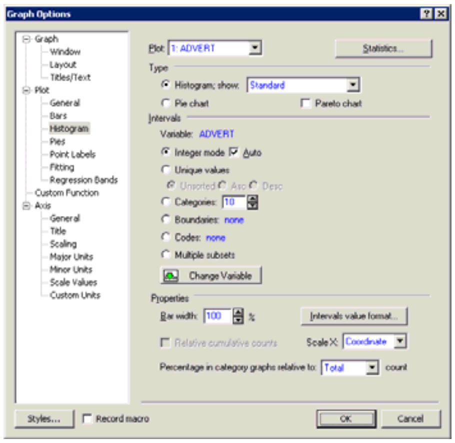
Note that the classic menus/toolbars will continue to be available, and you can switch between the two interfaces at any time.

To display the classic menus/toolbars, click *Menus* on the Quick Access toolbar in the upper-left corner of the ribbon bar.



To display the *STATISTICA* ribbon bar, select *Ribbon Bar* from the *View* menu.

Other improvements in the user interface can be seen with options that control the visual appearance of graphs, the behavior of spreadsheets, and the overall responsiveness of *STATISTICA*. In the more complex (multilayered) dialogs, it is now easier to find and explore your choices with new tree controls for application options [in the *Options* dialog, accessible by clicking *Options* on the *Tools* tab (ribbon bar) or by selecting *Options* from the *Tools* menu (classic toolbar)] and graph options [in the *Graph Options* dialog, accessible by clicking *Graph Options* on the *Format* tab (ribbon bar) or by selecting *Graph Options* from the *Format* menu (classic toolbar)].



Plot: 1: ADVERT

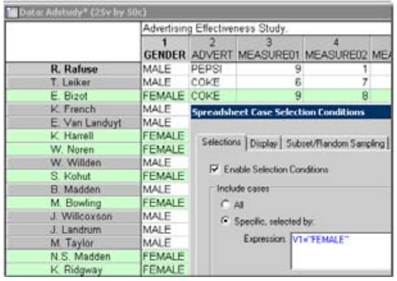
Type
● Histogram; show: Standard
○ Pie chart □ Pareto chart

Intervals
Variable: ADVERT
● Integer mode ☑ Auto
○ Unique values
● Unsorted ○ Asc ○ Desc
○ Categories: 10
○ Boundaries: none
○ Codes: none
○ Multiple subsets

Properties
Bar width: 100 %
□ Relative cumulative counts Scale X: Coordinate
Percentage in category graphs relative to: Total count

□ Record macro

A quick way to specify subsets of cases for analyses is to select the *Enable Selection Conditions* check box in the *Spreadsheet Case Selection Conditions* dialog. This new option enables the visual display of selected cases, identifying the cases using a light-green background by default. This enables you to see easily which cases will be used for the analysis.



Data: Adstudy* (25v by 50c)

Advertising Effectiveness Study.

	1 GENDER	2 ADVERT	3 MEASURE01	4 MEASURE02
R. Rafuse	MALE	PEPSI	9	1
T. Leiker	MALE	COKE	6	7
E. Bizot	FEMALE	COKE	9	8
K. French	MALE			
E. Van Landuyt	MALE			
K. Harrell	FEMALE			
W. Noren	FEMALE			
W. Willden	MALE			
S. Kohut	FEMALE			
B. Madden	MALE			
M. Bowling	FEMALE			
J. Willcoxson	MALE			
J. Landrum	MALE			
M. Taylor	MALE			
N.S. Madden	FEMALE			
K. Ridgway	FEMALE			

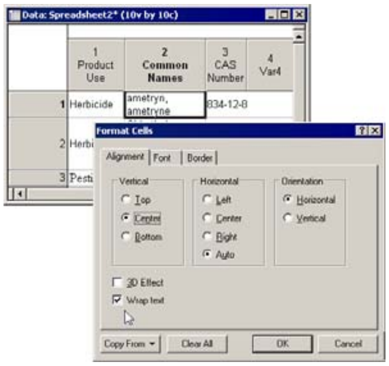
Spreadsheet Case Selection Conditions

Selections | Display | Subset/Random Sampling

☑ Enable Selection Conditions

Include cases
○ All
● Specific, selected by:
Expression: V1="FEMALE"

Spreadsheet cell-display enhancements apply to text that is too long to be displayed in a cell at the current column width. The improvements in version 9 are two-fold: with the default wrapping settings, if the adjacent cells are empty, the text will now extend into those adjacent cells. Secondly, you can now *Wrap text* within the spreadsheet data, allowing lengthy text to be displayed on multiple lines in a spreadsheet cell.



Data: Spreadsheet2* (10v by 10c)

	1 Product Use	2 Common Names	3 CAS Number	4 Var4
1	Herbicide	ametryn, ametryne	834-12-8	
2	Herbi			
3	Pesti			

Format Cells

Alignment | Font | Border

Vertical: ○ Top ● Center ○ Bottom
Horizontal: ○ Left ○ Center ○ Right ● Auto
Orientation: ● Horizontal ○ Vertical

□ 3D Effect
☑ Wrap text

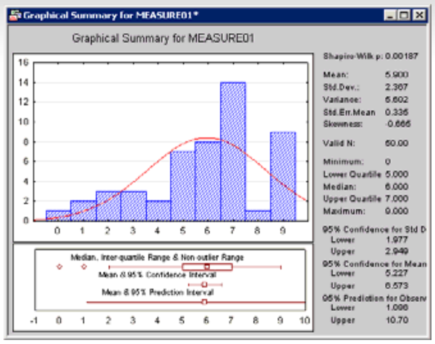
Copy From | Clear All | OK | Cancel

Graphics Enhancements

Visual Exploration of Data

One of *STATISTICA*'s strengths is visual exploration of data. StatSoft continues to build upon this strength with new visualizations for exploring data distribution, comparing variables, and creating color maps for correlation matrices within *Basic Statistics*.

Select *Descriptive statistics* (in the *Basic Statistics and Tables* Startup Panel) to see two new compound graph options; *Graphs 2* and *Graphical comparative summary display*. Use *Graphs 2* to explore data distribution for one variable.



Graphical Summary for MEASURE01*

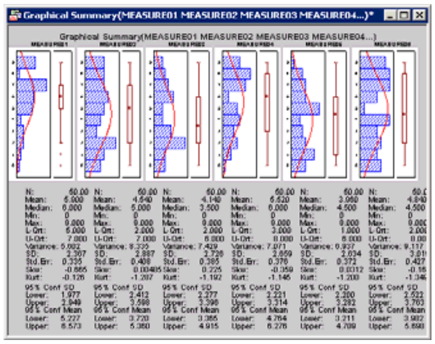
Graphical Summary for MEASURE01

Shapiro-Wilk p: 0.00187

Mean:	5.900
Std.Dev.:	2.367
Variance:	5.602
Std.Err.Mean	0.336
Skewness:	-0.666
Valid N:	60.00
Minimum:	0
Lower Quartile	5.000
Median:	6.000
Upper Quartile	7.000
Maximum:	9.000
95% Confidence for Std D	
Lower	1.977
Upper	2.949
95% Confidence for Mean	
Lower	5.227
Upper	6.573
95% Prediction for Observ	
Lower	1.090
Upper	10.70

16
14
12
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8 9
Median, Inter-quartile Range & Non-outlier Range
Mean & 95% Confidence Interval
Mean & 95% Prediction Interval
-1 0 1 2 3 4 5 6 7 8 9 10

Use *Graphical comparative summary display* to compare up to six variables in one graph. The histogram and box plot use the same scale per variable.

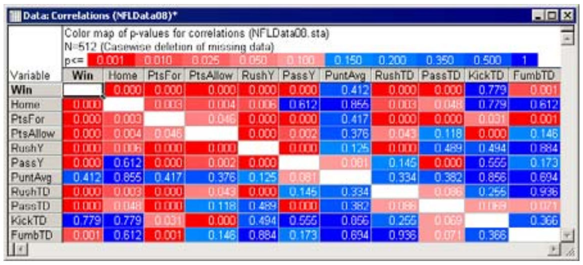


Graphical Summary(MEASURE01 MEASURE02 MEASURE03 MEASURE04...)*

Graphical Summary(MEASURE01 MEASURE02 MEASURE03 MEASURE04...)

	MEASURE01	MEASURE02	MEASURE03	MEASURE04	MEASURE05	MEASURE06
N:	60.00	60.00	60.00	60.00	60.00	60.0
Mean:	5.900	4.640	4.140	5.520	3.960	4.840
Median:	6.000	5.000	3.500	6.000	4.500	4.500
Min:	0	0	0	0	0	0
Max:	9.000	9.000	9.000	9.000	9.000	9.00
L-Qrt:	5.000	2.000	2.000	3.000	1.000	2.00
U-Qrt:	7.000	7.000	6.000	8.000	6.000	8.00
Variance:	5.602	8.335	7.429	7.071	6.937	9.117
SD:	2.367	2.887	2.726	2.659	2.634	3.01
Std.Err.:	0.335	0.408	0.385	0.376	0.372	0.427
Skw:	-0.666	0.00485	0.225	-0.359	0.0312	-0.16
Kurt:	-0.126	-1.207	-1.192	-1.146	-1.200	-1.34
95% Conf SD						
Lower:	1.977	2.412	2.277	2.221	2.200	2.522
Upper:	2.949	3.598	3.396	3.314	3.282	3.703
95% Conf Mean						
Lower:	5.227	3.720	3.365	4.764	3.211	3.982
Upper:	6.573	5.360	4.915	6.276	4.709	5.698

Select *Correlation matrices* to explore the new color maps option. For example, in the display shown below, it is easy to see that winning in football is highly correlated to playing on the home field, score, rushing, etc. (in this example color coding is based on the statistical significance level of the correlation coefficients, as shown in the legend displayed in the title area of the table).



Data: Correlations (NFLData08)*

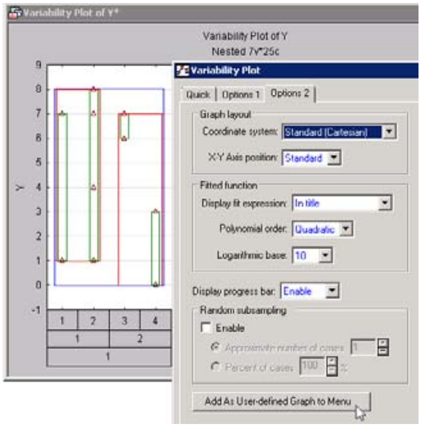
Color map of p-values for correlations (NFLData08.sta)
N=512 (Casewise deletion of missing data)

p<=	0.001	0.010	0.025	0.050	0.100	0.150	0.200	0.350	0.500	1

Variable	Win	Home	PtsFor	PtsAllow	RushY	PassY	PuntAvg	RushTD	PassTD	KickTD	FumbTD
Win		0.000	0.000	0.000	0.000	0.000	0.412	0.000	0.000	0.779	0.001
Home	0.000		0.003	0.004	0.006	0.612	0.855	0.003	0.048	0.779	0.612
PtsFor	0.000	0.003		0.046	0.000	0.000	0.417	0.000	0.000	0.031	0.001
PtsAllow	0.000	0.004	0.046		0.000	0.002	0.376	0.043	0.118	0.000	0.146
RushY	0.000	0.006	0.000	0.000		0.000	0.125	0.000	0.489	0.494	0.884
PassY	0.000	0.612	0.000	0.002	0.000		0.081	0.145	0.000	0.555	0.173
PuntAvg	0.412	0.855	0.417	0.376	0.125	0.081		0.334	0.382	0.856	0.694
RushTD	0.000	0.003	0.000	0.043	0.000	0.145	0.334		0.086	0.255	0.936
PassTD	0.000	0.048	0.000	0.118	0.489	0.000	0.382	0.086		0.069	0.071
KickTD	0.779	0.779	0.031	0.000	0.494	0.555	0.856	0.255	0.069		0.366
FumbTD	0.001	0.612	0.001	0.146	0.884	0.173	0.694	0.936	0.071	0.366	

A *Variability plot* is used to evaluate the variability of one factor within several other organizing factors and for visual breakdown and data mining analyses. Now the *Variability plot* can be used to create user-defined graphs.

As with other user-defined graphs, this can save time, i.e., you can save common settings and then apply them to another data set.



Variability Plot of Y*

Variability Plot of Y
Nested 7v*25c

Variability Plot

Quick | Options 1 | Options 2

Graph layout
Coordinate system: Standard (Cartesian)
X-Y Axis position: Standard

Fitted function
Display fit expression: In title
Polynomial order: Quadratic
Logarithmic base: 10

Display progress bar: Enable

Random subsampling
□ Enable
● Approximate number of cases 1
○ Percent of cases 100 %

Graph Categories

In previous versions, the number of graph categories was limited to 255; this limit has been increased to 1,000; e.g., it is now possible to create a multiple box plot with up to 1,000 boxes.

Graph Customization Macro Recording Options

Macro recording is a powerful option, and the macro object model architecture of *STATISTICA* and the scripting capabilities are easily available to anyone. No technical knowledge of *STATISTICA* Visual Basic is required. Common tasks can be automated or controlled for regulated work environments.

Macro recording enables you to access programmatically almost every aspect and virtually every detail of the functionality of the program. Even the most complex analyses and graphs can be recorded into *STATISTICA* Visual Basic (SVB) macro programs. They can later be run repeatedly or edited and used as building blocks of other applications.

Now this functionality has been expanded to include the recording of custom graph options. For example, suppose you create a line plot and want to change the line color and thickness. You open the *Graph Options* dialog and then modify the line options in the *Plot General* options pane. Select the *Record Macro* check box at the bottom of the *Graph Options* dialog, and click the *OK* button to generate a macro.

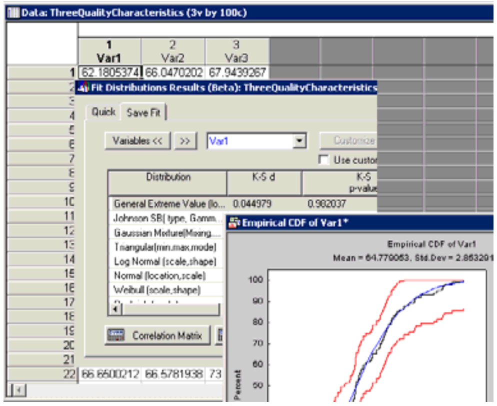
This powerful option also applies when recording master macros. You can now start a master macro recording session, run analyses, and customize the resulting graphs. If you re-run the master macro, the analyses will be replicated, along with the graphics customizations.

Statistics Enhancements

Distributions and Simulation

We are pleased to announce the beta release of *Distributions & Simulation*. This module enables users to automatically fit a large number of distributions for continuous and categorical variables to lists of variables. Standard distributions are available (normal, half-normal, log-normal, Weibull, etc.), but also included are specialized and general distributions (Johnson, Gaussian Mixture, Generalized Pareto, Generalized Extreme Value), and *STATISTICA* automatically ranks the quality of the fit for each selected distribution and variable.

In addition, the distributions fit to the list of selected variables and the covariance between the selected variables can be saved for deployment. The *Distributions & Simulation* module uses this deployment information to generate simulated data sets that not only faithfully reproduce the respective distributions, but also the covariances between variables. In short, in addition to facilitating efficient distribution fitting to large numbers of variables, this module enables users to fit general multivariate distributions, and simulate from those distributions, using cutting edge simulation techniques (e.g., Latin-Hypercube simulation).



Data: ThreeQualityCharacteristics (3v by 100c)

	1 Var1	2 Var2	3 Var3
1	62.1805374	66.0470202	67.9439267
22	66.6500212	66.5781938	73

Fit Distributions Results (Beta): ThreeQualityCharacteristics

Quick | Save Fit

Variables << | >> | Var1 | Customize

□ Use custo

Distribution	K-S d	K-S p-value
General Extreme Value (lo...	0.044979	0.982037
Johnson SB(type, Gamm...		
Gaussian Mixture(Mixing...		
Triangular(min,max,mode)		
Log Normal (scale,shape)		
Normal (location,scale)		
Weibull (scale,shape)		

Empirical CDF of Var1*

Empirical CDF of Var1
Mean = 64.770063, Std.Dev = 2.863291

These methods have proven useful in various domains such as modern DOE, reliability engineering, and risk modeling.

StatSoft welcomes your comments or observations regarding this new addition to the selection of analytic modules in *STATISTICA*. Please send your input to beta@statsoft.com.

Basic Statistics

Several new basic statistics have been added. The computation of Welch’s F statistic to test for equality of means when the variances are unequal is now available on the *ANOVA & tests* tab, located in the *Breakdown & one-way ANOVA* analysis results dialog.

Confidence interval estimates for the differences between means with confidence limits is now available on the *Options* tab of the *T-Test for Independent Samples by Variables* dialog and the *Advanced* tab of the *T-Test for Dependent Samples* dialog.

General Optimization

The *General Optimization* module, which is part of *STATISTICA Process Optimization*, is a unique, powerful, open-architecture product that enables users to optimize arbitrary functions of virtually any complexity, using Simplex, Genetic Algorithm, or Grid-Search methods. This module (released in beta version) has applications in virtually all domains in which there is a need to find best parameters that control specific processes to achieve optimal results according to user-specified criteria (e.g., process industries, business, finance, science). The function to be optimized can be specified in a simple *STATISTICA* Visual Basic (SVB) function or a set of formulas. This new module was specifically designed to make repeated invocation of other *STATISTICA* (or other, e.g., R) functions referenced in the optimization function very efficient. Therefore, optimization problems that involve multiple data mining prediction models (e.g., complex cost models) or simulation (for stochastic optimization, or for optimizing for multivariate process capability) can now be easily set up and solved efficiently.

Nonlinear Estimation

Changes have been made for *User-specified regression, least squares* and *User-specified regression, custom loss function*. While creating the estimated function, there is an option to review the variables. A common next step after reviewing variables is to use a variable in the function, but, in the past, you had to type the variable into the function. Now variables can be reviewed, selected, and inserted into the function via the *Review vars* button.

Occasionally there were name conflicts between function names and variable names when using *User-specified regression, least squares*. For example, one customer named a variable *PRED*, and there is a function named *PRED*, and the customer had to change the variable name before continuing with the analysis. Now *PRED* and *OBS* can be used as variable names for this analysis.

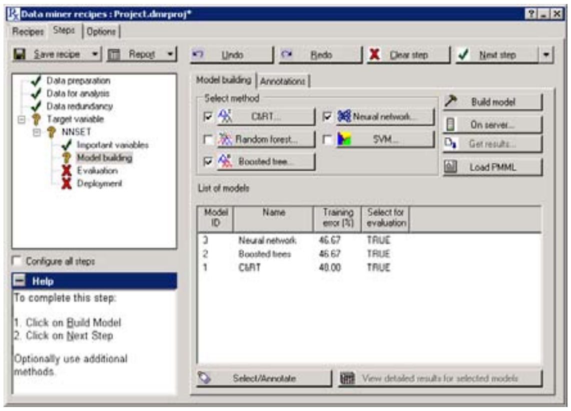
Principal Components & Classification Analysis

The *Results* dialog - *Descriptives* tab contains a *2D scatterplots* button. You can now select multiple variables and generate all the scatterplots with one click.

Data Mining Enhancements

Data Miner Recipes

STATISTICA 9 features the long awaited final release of *Data Miner Recipes (DMR)*, previously available only as a beta release. DMR is an easy and flexible step-by-step data mining guide, and it is now available to all of StatSoft's data miner customers. Novice data miners can quickly clean and analyze data, while advanced users can work more efficiently and have one more option to automate routine tasks. DMR explores the data and makes default decisions for you. You can easily modify these defaults as needed and save them for repeated use.



☑ C&RT... ☑ Neural network...
☐ Random forest... ☐ SVM...
☑ Boosted tree...

List of models

Model ID	Name	Training error (%)	Select for evaluation
3	Neural network	46.67	TRUE
2	Boosted trees	46.67	TRUE
1	C&RT	40.00	TRUE

☐ Configure all steps

To complete this step:

1. Click on Build Model
2. Click on Next Step

Optionally use additional methods.

Other Enhancements

New shortcut keys have been added for connecting data miner nodes in the workspace, and further drag/drop functionality was also added. Previous versions of *STATISTICA* required you to click on the second node to complete a connection. Now you can just drop the connection arrow on top of the second node to complete a connection.

Large classification and regression tree displays are now scrollable.

The *MARSplines* results spreadsheet for outcome, where the variable name used to be truncated to 8 characters, now displays the whole variable name for the independent variable.

The *General Optimization* module (see the description, above), which we are now releasing as a beta version, is included in *STATISTICA Process Optimization*. This module enables users to optimize arbitrary functions using Simplex, Genetic Algorithm, or Grid-Search methods. The function to be optimized can be specified in a simple *STATISTICA* Visual Basic (SVB) formula or program.

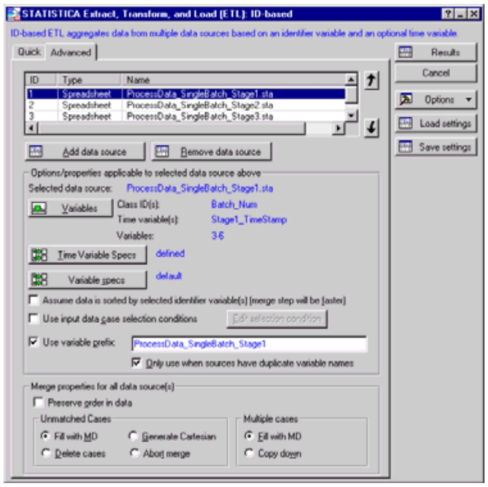
This new module was specifically designed to easily call (iteratively) any *STATISTICA* (or other) functions. Therefore, optimization problems that involve multiple data mining prediction models (e.g., complex cost models), and simulation (for stochastic optimization or for optimizing for multivariate process capability), can now be easily set up and solved efficiently.

Miscellaneous Enhancements

STATISTICA ETL (Extract, Transform, and Load)

STATISTICA Extract, Transform, and Load (ETL) offers powerful tools to align, merge, and intelligently combine data from databases and submit them to the powerful *STATISTICA* data processing capabilities for data filtering, aggregation, alignment, and analyses.

ID-Based *STATISTICA ETL* can be used to align data from disparate sources by batch number and time interval, and/or by one or more ID fields. Merging many-to-one data sets is a common scenario in many process and manufacturing industries.



ID-based ETL aggregates data from multiple data sources based on an identifier variable and an optional time variable.

ID	Type	Name
1	Spreadsheet	ProcessData_SingleBatch_Stage1.sta
2	Spreadsheet	ProcessData_SingleBatch_Stage2.sta
3	Spreadsheet	ProcessData_SingleBatch_Stage3.sta

Options/properties applicable to selected data source above

Selected data source: ProcessData_SingleBatch_Stage1.sta

Class ID(s): Batch_Num
Time variable(s): Stage1_TimeStamp
Variables: 3-6

Time Variable Specs: defined
Variable specs: default

☐ Assume data is sorted by selected identifier variable(s) (merge step will be faster)
☐ Use input data case selection conditions
☑ Use variable prefix: ProcessData_SingleBatch_Stage1
☑ Only use when sources have duplicate variable names

Merge properties for all data source(s)

☐ Preserve order in data

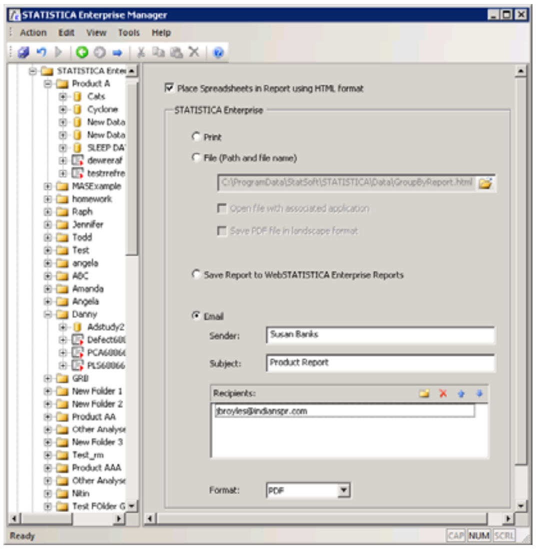
Unmatched Cases: ● Fill with MD ○ Generate Cartesian ○ Delete cases ○ Abort merge

Multiple cases: ● Fill with MD ○ Copy down

For example, using *STATISTICA ETL* it is easy to pre-process batch manufacturing data to constant batch length (solve the "unequal batch length" problem for *PLS/PCA* model-based quality control with *STATISTICA* or on-line real-time *STATISTICA Enterprise*). Other typical applications include process data collected at different time intervals that need to be aligned either by aggregating the values for the variables collected at the higher frequency, or by replicating the values for the variables collected at the lower frequency (e.g., when analyzing historical process data describing the performance of a furnace, to align combustion parameters collected at 1-minute intervals with fuel quality data collected daily). In version 9, the product has been further updated for improved performance and scalability and more detailed reporting (e.g., to report the actual intervals in the output results).

Enterprise Reports

STATISTICA Enterprise's reporting interface allows you to create reports in HTML, PDF, and RTF formats. Now the system can automatically email these reports when they are run.



☑ Place Spreadsheets in Report using HTML format

STATISTICA Enterprise

○ Print

○ File (Path and file name)

C:\ProgramData\StatSoft\STATISTICA\Data\GroupByReport.html

□ Open file with associated application

□ Save PDF file in landscape format

○ Save Report to WebSTATISTICA Enterprise Reports

● Email

Sender: Susan Banks

Subject: Product Report

Recipients:

jbroyles@indianspr.com

Format: PDF

Changes have also been made to the audit logging facilities to always record audit log times in universal coordinated time. The conversion is made to the local time zone when the audit log is displayed. This enables users in different time zones to have an accurate view of what changed when.

WebSTATISTICA Licensing

WebSTATISTICA has always been licensed per-processor. Now the licensing system has been modified to allow for deploying separate *WebSTATISTICA* instances on separate servers without requiring special licensing. For instance, a site licensed for 8 CPUs can either deploy this license on a single 8 CPU server or on two separate 4 CPU servers pointing at the same license file.

Microsoft Installer (MSI) Support

With the release of version 9, we have changed the *STATISTICA* application installation platform to Microsoft Installer (MSI) instead of InstallShield installer that we used in version 8 and before.

When installing interactively, the user experience is similar to that from version 8, but the dialogs are more attractive.

However, the real benefit from using MSI is in how the *STATISTICA* installer can now be integrated into other installation packages and enterprise installation tools. The MSI allows for a totally "silent" installation, where all the information the user enters (CDKEY, serial number, netID, install code, and user registration information) can be passed either directly on the command line as command-line parameters or in a parameter file that the command line references.

There are three main use cases that we see for the MSI installer:

1. Integrating our installer with other installers.
2. Working with enterprise-level deployment solutions for single-user installations.
3. Working with enterprise-level deployment solutions for concurrent workstation installations.

For more information on these new installation options, please contact your local StatSoft office.

Expanded Interfaces for Developers

In version 9 of *STATISTICA*, a variety of enhancements have been added for developers and system integrators.

1. A new lightweight *STATISTICA* Spreadsheet library is now distributed with *STATISTICA.* It is available at no cost to third party developers who intend to read or write *STATISTICA* data files. It is multi-threaded and has a separate multi-threaded library for .NET access.
2. Graphs have a new event interface, OnGroupingSelect. This event is used for graphs that are categorized or aggregated in some fashion, for instance, a histogram (data are categorized into bars), a box whisker plot (separate bars represent separate categories), or categorized graphs. When the user selects items in the plots, the OnGroupingSelect event is fired to provide information about what groups/categories that selection represents. This new interface now allows applications using *STATISTICA* graphs to implement drill-down capability
3. *STATISTICA* will no longer by default include all macro references in a newly created macro. Instead, each individual module will add its specific references when a macro is recorded. Suppressing all the references makes macros start up more quickly, but you could run into unresolved references if you copy/paste code from one macro into another. Therefore, the program will now check when you are copying/pasting between macros that have a different list of references, and offer to copy the additional references.
4. A new command-line parameter /MacroArgument has been added that can be used in conjunction with the /RunMacro argument. This will allow you to pass a parameter to the macro being run, which the macro can access with the GetScriptArgument call.
5. *STATISTICA* offers a comprehensive set of integration options to use with procedures written in R, which is a highly extensible programming language and environment for statistical computing (http://www.r-project.org). All versions of R (up to 2.8.1, the most current version as of this writing) can be executed within *STATISTICA*. R results can be displayed in native *STATISTICA* Spreadsheets and Graphs. A variety of integration options are offered including execution of R code on *STATISTICA* servers. See the *Integration Options and Features to Leverage Specialized R Functionality in STATISTICA and WebSTATISTICA Solutions* (http://www.statsoft.com/products/webserver.htm) white paper for more details.
6. ANSI-92 SQL JOIN syntax is now supported by *STATISTICA Query*. Newer versions of SQL Server will require these types of joins to be used. By default, this option isn't selected. You can set the option per query or for all queries.

Please contact StatSoft at **918-749-1119** or info@statsoft.com if you have any questions about your *STATISTICA 9* upgrade.