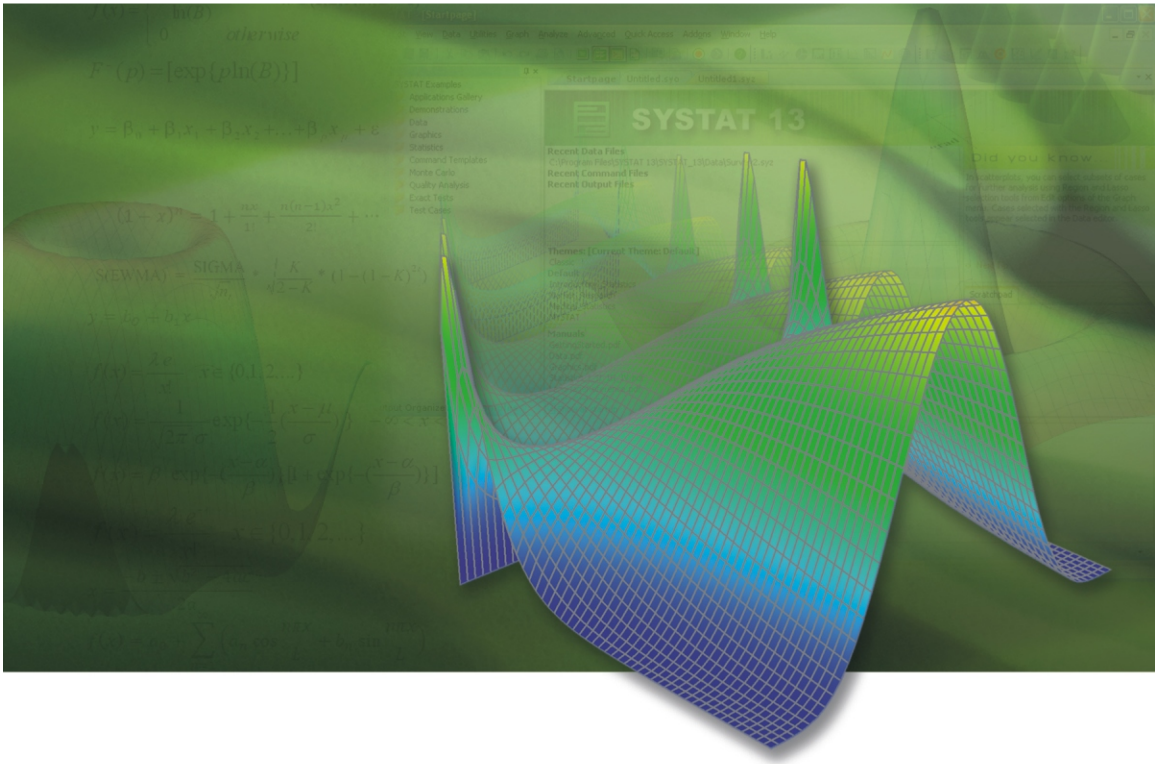
SYSTAT® 13



Getting Started

SYSTAT®

WWW.SYSTAT.COM

For more information about SYSTAT® software products, please visit our WWW site at *http://www.systat.com* or contact

Marketing Department
Systat Software, Inc.
225 W. Washington Street, Ste. 425
Chicago, IL 60606
Phone: (877) 797-8280
Fax: (312) 220-0070
Email: info-usa@systat.com

Windows is a registered trademark of Microsoft Corporation.

General notice: Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

The SOFTWARE and documentation are provided with RESTRICTED RIGHTS. Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subdivision (c)(1)(ii) of The Rights in Technical Data and Computer Software clause at 52.227-7013. Contractor/manufacturer is Systat Software, Inc., 225, W. Washington Street, Suite 425, Chicago, IL 60606. USA.

SYSTAT® 13 Getting Started
Copyright © 2009 by Systat Software, Inc.
Systat Software, Inc.
225 W. Washington Street, Ste. 425
Chicago, IL 60606
All rights reserved.
Printed in the United States of America.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

1 2 3 4 5 6 7 8 9 0   05 04 03 02 01 00

Contents

3 SYSTAT Basics 45

4 Data Analysis Quick Tour 99

6 Working with Output *185*

7 *Customization of the SYSTAT Environment* *203*

Chapter

1

What's New and Different in SYSTAT 13

This chapter gives a summary of new features and major changes in this version, relative to SYSTAT 12, in respect of GUI, data, commands, output, help, graphics, and statistics. Under each of these items, a list is given of new, modified and deleted features. This is followed by a brief description of each item in the same order, with the same serial number. More details are given in the appropriate chapters in the manual.

GENERAL FEATURES

Graphical User Interface

New Features

1. Autohide Spaces
2. Choice Tokens
3. Data Edit Bar
4. Data File Information
5. Default Format for Saving Command Files
6. Drag-and-Drop Data
7. Embedded Toolbars
8. Open Legacy Command Files
9. View Toolbars

10. Windows XP Style Grids
11. Trim Leading and Trailing Spaces in String Data

Modified Features

12. Autocomplete Commands
13. Command Coloring
14. Dialog Boxes
15. Rescue Report
16. Shortcut Keys
17. Status Bar
18. Themes

Deleted Features

19. Open Multiple Graphs (View and Active Modes)
20. Print Content of Data/Variable Editor

Data

New Features

21. Close Data Files
22. Default Variable Format
23. Save View Mode Data Files
24. Import Business Objects

Modified Features

25. Copy/Paste to Data/Variable Editor
26. Open Multiple Data Files
27. Recode Variables
28. Store and Retrieve Current Settings

Deleted Features

29. View Data Files

Commands

New Features

30. ACTIVE Command
31. Built-in Functions
32. FOCUS Command
33. Macros

Modified Features

34. FUNCTION Command
35. Multiple Option Values
36. PAGE NONE
37. Precedence Rules
38. String Subscripted Variables
39. Temporary Variables

Deleted Features

40. Built-In Variables

Output

New Features

41. Locales and Digit Grouping
42. Node and Link Captions

Graphics

New Features

43. Color using RGB Values
44. Gradient Colors for Surfaces
45. Label Dots in Dot (Summary) Charts
46. Built-In Colors
47. Colors for overlaid graphs, pie and stacked charts
48. Stacked Bar Charts with Grouping Variable
49. Individual Border Displays on Plots
50. Multiple Slices in Pie Charts
51. Numeric Case Labels

STATISTICAL FEATURES

New Features

1. ARCH and GARCH Models in Time Series
2. Best Subsets Regression
3. Confirmatory Factor Analysis
4. Environment Variables in Basic Statistics
5. Hypothesis Testing for Multivariate Mean
6. New Basic Statistics
7. Bootstrap Analysis in Hypothesis Testing
8. New Nonparametric Tests
9. Polynomial Regression

Modified Features

10. Analysis of Variance
11. Crosstabulations

12. Cluster Analysis
13. Fitting Distributions
14. Hypothesis Testing for Two-Sample Data in Columns
15. Least Squares Regression
16. Logistic Regression
17. Mixed Models

Descriptions for each of the above items are given in the following pages.

GUI

New Features

1. **Autohide Spaces**

   You can autohide the Workspace and Commandspace by clicking the button. For details about customizing the SYSTAT window, refer Chapter 7, *Customization of the SYSTAT Environment* in the Getting Started volume of the user manual.

2. **Choice Tokens**

   SYSTAT now allows you to define choice tokens using a new type of token dialog box where you may specify between 2 to 10 choices. Each choice may be linked to a SYSTAT command script so that, depending on the user's choice, the corresponding script will be executed. This gives you the ability to incorporate several (up to 10) sets of scripts, covering various possible scenarios for a given analysis, into a single SYSTAT command script. Depending on the user's choice, any given set may then be executed.

3. **Data Edit Bar**

   The Data/Variable Editor has a new toolbar called the Data Edit Bar. This allows you to navigate to any cell in the Data Editor, and view/edit data values. For more details about the Data Edit Bar, refer Chapter 3, Entering and Editing Data in the Data volume of the user manual.

4. **Data File Information**

   You can click the  button in the bar beside the Data and Variables tabs of the Data Editor to enter or edit comments related to the corresponding data file. Simply pause the mouse on the button to view the file comments currently entered for the data file.

5. **Default Format for Saving Command Files**

   Earlier versions of SYSTAT saved command files in the ANSI format and the previous version saved them in the Unicode format. SYSTAT now allows you to specify the format to save command files. There is also a setting in the Edit: Options dialog box where you may specify the default command file format.

6. **Drag-and-Drop Data**

   You may now drag and drop text into SYSTAT's Data Editor from editors that support dragging of content. This includes dragging and dropping text entered in the Commandspace of SYSTAT itself.

7. **Embedded Toolbars**

   The Format Bar, the Data Edit Bar and the Graph Editing toolbar, are now embedded in the Output Editor, Data Editor and Graph Editor tabs respectively.

8. **Open Legacy Command Files**

   You may now directly open and execute legacy command files if a VERSION command is inserted as the first line. The syntax is VERSION n where n may be either 11 or 12. Apart from this, the Translate Legacy Commands dialog box and the SYSTAT Command Translator also allow you to specify the version, whether it is 11 or 12, of the command file you want to translate.

9. **View Toolbars**

   You may now load one or more of SYSTAT's toolbars through the View menu. The entries corresponding to the toolbars that are loaded are prefixed by a check mark.

10. **Windows XP Style Grids**

    SYSTAT's Data/Variable Editor grid now adopts the current Windows XP theme that is applied to the Windows Desktop. Certain grid controls in dialog boxes like Data: Transform: If Then Let and Data: Select Cases also have the same look and feel.

11. Trim Leading and Trailing Spaces in String Data

You may now control the trimming of leading and trailing spaces in string data as you type/modify strings in the Data Editor. Check/uncheck this option in the Data tab of the Edit: Options dialog box.

Modified Features

12. Autocomplete Commands

Command arguments, options and option values will be "autocompleted" as they are typed in the Interactive or batch (Untitled) tab of the Commandspace. Arguments may be filenames, variable names, built-in function names or specific key words. If filenames or their paths involving spaces are selected, then they are automatically enclosed in quotes. Function names are automatically suffixed by parentheses.

13. Command Coloring

Coloring of command keywords is now an optional feature though set by default. You may set/suppress this option in the General tab of the Edit: Options dialog box. Also, variable names are now colored black, and option values are colored green.

14. Dialog Boxes

The tabbed dialog boxes of SYSTAT now have the tabs arranged vertically. This allows more tabs to be easily accessible with just a single click of the mouse.

15. Rescue Report

SYSTAT now attempts to restore a session that has just crashed. Also, if you click Send Report, the rescued files are automatically attached to the email message.

16. Shortcut Keys

SYSTAT now has the following new shortcut keys provided by default:

Ctrl + Q	Quit SYSTAT
Alt + backspace	Undo
Ctrl + Alt + Enter	Variable Properties
Ctrl + K	View Workspace

See the section on Keyboard Shortcuts in Chapter 7, *Customization of the SYSTAT Environment*, for a complete list.

17. Status Bar

The following enhancements have been made to the Status Bar:

- The page width can be set to Narrow, Wide or None by clicking PAGE on the Status Bar.
- The states of the Insert, Caps Lock, Num Lock and Scroll Lock keys on the keyboard can be toggled through the Status Bar. See Chapter 7, *Customization of the SYSTAT Environment* for a complete list of items on the Status Bar.

18. Themes

The following enhancements have been made to SYSTAT's Themes feature:

- Download Themes now has a dialog box interface wherein you may choose which themes to install.
- Theme files now have versions so that you will have the option to upgrade your theme file whenever a newer version is available on the SYSTAT server.
- When you apply a theme, you will be prompted to save the current theme.

Deleted Features

19. Open Multiple Graphs (View and Active Modes)

It is no longer possible to view multiple graphs in the Graph Editor. The latest graph, or a graph that you double-click on, will be displayed in the Graph Editor for editing.

20. Print Content of Data/Variable Editor

SYSTAT no longer supports printing the content of the Data/Variable Editor. To print data, list the variables in the output and print the output. To print variable information, click Utilities -> File Information -> Dictionary and print the resultant output.

Data

New Features

21. Close Data Files

You may now close data files using the context menu of the Data Editor, or the CLOSE command. Run CLOSE filename to close a particular file, or CLOSE / ALL to close all but the active data file.

22. Default Variable Format

You may now set a distinct default numeric variable format for new numeric variables in the Data Editor. This format is now independent of the numeric output format.

23. Save View Mode Data Files

You may now save data files that are in the view mode. Simply bring the desired 'view mode' tab into focus and click the Save button on the Standard toolbar, or click File -> Save.

24. Import Business Objects

SYSTAT now offers the option of using a "Business Objects Universe" as a data source (similar to the other choices such as ODBC, Excel, etc.). Business Objects is business intelligence platform organization, which supports pre-defined reports, ad-hoc reporting, dashboards, and extraction, transform, and load operations necessary when building data warehouse. The "Business Objects Universe" is a "semantic layer" which sits between the business end user and the complexities of the underlying database model. End users force the universe to access all the databases to which they have been given permission.

This feature allows you to login to the Business Objects platform, choose a universe to query, build a query, and process the resultant data in SYSTAT.

Modified Features

25. Copy/Paste to Data/Variable Editor

SYSTAT now allows you to copy a cell and paste it into a column. However it no longer supports the following:

- Pasting one or more cells in a row/column to a block of cells encompassing more than one row/column.
- Pasting an individual variable property to a new row in the Variable Editor.
- Pasting more than one property simultaneously to a block of variables.

26. Open Multiple Data Files

In the previous version of SYSTAT, the ability to work with multiple unmodified data files was tied to the global option to order output based on the input data file. The two options have been delinked in this version and, by default, you may have multiple unmodified data files open with output ordered chronologically. At a time, you may set any one of the files active for further processing. If you still want to work with a single active data file, SYSTAT provides a distinct global option to close the active data file when another is opened. Independent of this setting, you may order the output either chronologically or based on the input data file.

27. Recode Variables

SYSTAT now offers an option, ELSE, which will allow you to recode all values other than a given set of values to a certain specified value. Also, when you recode into a new variable, it inherits all non-recoded values from the old variable. Use the ELSE option if you do not want to inherit the non-recoded values.

28. Store and Retrieve Current Settings

SYSTAT now supports storing the current setting of the following:

- active data file
- value label display format
- variable label display format

The stored settings may then be retrieved at any subsequent instant during the current session.

Deleted Features

29. View Data Files

You will no longer be able to open data files directly in the view mode. However, by default, data file tabs will switch to view mode as before when another file is opened or set active.

Command Line Interface

New Features

30. ACTIVE Command

The ACTIVE command now activates a file that is in the view mode. It no longer opens the file from disk.

31. Built-in Functions

SYSTAT offers the following new built-in functions:

Mathematical:

ACSH	COSH	ODD
ASNH	EVEN	ROUND
CASE	FLOOR	SINH
CEIL	NCASE	
COLUMN	NVAR	

Multivariable:

COMPLETE

Groups and Intervals:

BOF	EOF	NCAT
BOG	EOG	

Character:

CHR$	SNUM
CODE	LEN

Date-Time:

FDAYM	LDAYM	MON
FDAYW	LDAYW	

Statistical:

BGCF	GDCF	P6CF
BGDF	GDDF	P6DF
BGIF	GDIF	P6IF
BGRN	GDRN	P6RN
EMCF	P5CF	PECF
EMDF	P5DF	PEDF
EMIF	P5IF	PEIF
EMRN	P5RN	PERN

32. FOCUS Command

SYSTAT now provides a FOCUS command for switching focus to the Data Editor, Graph Editor or Output Editor. Use it in command scripts to retain or force focus to be in a particular page of the Viewspace.

33. Macros

SYSTAT now allows you to define and call macros in your command scripts. A macro is a series of statements enclosed by the DEFMACRO and ENDMACRO commands. Macros may be used to execute a set of commands in many different places in a program.

34. FUNCTION Command

For user-defined functions, you now need to specify the type of the argument and the return type of the function as TMP. The syntax of the FUNCTION command is now as follows:

```
FUNCTION TMP funcname(TMP arg1, TMP arg2, …)
    statement1
    statement2
     -
    RETURN expression
ENDFUNC
```

35. Multiple Option Values

SYSTAT now expects multiple option values to be enclosed in braces ({}). For example, if you want to specify three colors for an overlaid graph, type the option as COLOR = {MAGENTA, BLUE, YELLOW}.

36. PAGE NONE

You can now set the page width to be unlimited using the PAGE NONE command.

37. Precedence Rules

The SYSTAT namespace, which consists of all its possible module names, commands, arguments, options and option values, now has the following precedence structure (highest to lowest):

- Class 0. SYSTAT module names, commands, options, and option values where such values are fixed keywords
- Class 1. Built-in function names
- Class 2. User-defined function, matrix, and array variable names
- Class 3. File variable names (in the currently active data file)

With the introduction of this precedence, there will not be restrictions on variable names that you use in data files. Depending on the context, a name will be treated as coming from the lowest-numbered class possible.

38. String Subscripted Variables

For string variable names that are subscripted, you now have to prefix the dollar sign before the subscript. For example, what was myvar(1)$ in the prior version should now be myvar$(1).

39. Temporary Variables

Temporary variable names should now be suffixed by the tilde (~) symbol, for example, mytmpvar~. Also, you need to use the TMP command to define temporary variables, for example, TMP mytmpvar~ = 10.

Deleted Features

40. Built-In Variables

The erstwhile CASE, COMPLETE, BOF, BOG, EOF, and EOG are no longer available as built-in variables. They are now functions that you may use as before just by suffixing parentheses '()' to the name. For example, SELECT COMPLETE() and IF CASE() < 10 THEN LET x = 5.

Output

New Features

41. Locales and Digit Grouping

You may now select the locale that SYSTAT should use while displaying numbers in the Output Editor. SYSTAT also determines the format of the number(s) you type in the Data Editor from this setting. That means, you can now type numbers using the decimal and digit grouping symbols of the selected locale. The default locale, corresponding to the entry "System default", is determined from the Regional and Language Settings in the Windows Control Panel.

42. Node and Link Captions

You may now set Output Organizer node and collapsible link captions using the NODE command. Run HELP NODE to know the command syntax for accessing this new feature.

Graphics

New Features

43. Color using RGB Values

SYSTAT now offers you the option of specifying colors in terms of their Red-Green-Blue component values. This is available for specifying the color of elements, axes and frame colors.

44. Gradient colors for surfaces through the dialog box

SYSTAT now allows you to specify the gradient style for surfaces through the dialog box. This is available in the Surface and Line Style tab of the dialog boxes for the relevant graph types.

45. Label Dots in Dot (Summary) Charts

SYSTAT now offers the option of labeling dots in dot (summary) charts.

Modified Features

46. Built-In Colors

SYSTAT now provides 45 built-in colors as against the 12 available in previous versions.

47. Colors for overlaid graphs, pie and stacked charts

Overlaid graphs, pie charts, and stacked bar charts will now be colored in such a way as to provide more contrast between adjacent elements.

48. Stacked Bar Charts with Grouping Variable

You may now stack bars in the case of grouped bar charts as well. A stacked chart is drawn for each group, and all the charts are laid out in the same frame.

49. Individual Border Displays on Plots

SYSTAT now provides options to separately specify the border displays for individual borders. This allows you to suppress the display along any given border, or specify different kinds of border displays along the two borders in all two-dimensional plots.

50. Multiple Slices in Pie Charts

You may now request separating multiple slices from a pie chart. Request specified slice numbers or all slices.

51. Numeric Case Labels

SYSTAT now allows you to specify a numeric variable for setting labels in plots, multivariate displays and maps. In prior versions, you could only use string variables for labeling elements.

Statistical Features

New Features

1. **ARCH and GARCH Models in Time Series**

   As part of its Time Series feature update, SYSTAT now offers:

   - Fitting of ARCH and GARCH models through BHHH, BFGS, and Newton-Raphson implementations of the maximum likelihood method. Various options for setting convergence criteria are provided.
   - Forecasts for error variances using the parameter estimates.
   - Jarque-Bera test for normality of errors.
   - McLeod and Lagrange Multiplier tests for ARCH effect.

2. **Best Subsets Regression**

   A new addition to SYSTAT's Regression suite, this feature includes:

   - Finding the best models (choice of predictors) given the number of predictors, the number varying from one to the total number available in the data set,
   - Identifying the best model by various criteria such as R-Square, Adjusted R-Square, Mallow's Cp, MSE, AIC, AICC and BIC, and
   - Performing a complete regression analysis on the data set chosen by the user (same as the training set or different) using the best model selected by any of the above criteria.

3. **Confirmatory Factor Analysis**

   As part of the Factor Analysis feature, SYSTAT now offers Confirmatory Factor Analysis (CFA) with:

   - Maximum likelihood, Generalized Least-Squares, and Weighted Least-Squares methods of estimation of parameters of the CFA model.
   - A wide of variety of goodness-of-fit indices to measure the degree of conformity of the postulated factor model to the data, which include Goodness-of-Fit Index (GIF), Root Mean Square Residual (RMR), Parsimonious Goodness-of- fit Index (PGFI), AIC, BIC, McDonald's Measure of Certainty, and Non-Normal Fit Index (NNFI)..

4. **Environment Variables in Basic Statistics**

   SYSTAT now provides environment variables in its Basic Statistics module. These are variables that contain the computed values of various statistics for a given session, a given data file, and given variables. These may be directly used in subsequent transformation statements for further processing of the computed statistics. For details, refer to Chapter 5, *Command Language*.

5. **Hypothesis Testing for Multivariate Mean**

   The Hypothesis Testing feature has been strengthened with tests for mean vectors of multivariate data:

   - One-sample Hotelling's T2 test for mean vector of multivariate data equal to a known vector.
   - Two-sample Hotelling's T2 test for equality of two mean vectors of multivariate data.

6. **New Basic Statistics**

   SYSTAT now offers the following new basic statistics:

   - Standard error and confidence interval for the trimmed mean.
   - Winsorized mean, its standard error and confidence interval.
   - Sample mode
   - Interquartile range

7. **Bootstrap Analysis in Hypothesis Testing**

   The Hypothesis Testing feature now provides:

   - Bootstrap-based p-values for all tests for mean (one-sample z, one-sample t, two- sample z, two-sample t, paired t, Poisson) and variance (single variance, two variances and several variances).

8. **New Nonparametric Tests**

   The Nonparametric Tests feature has been updated to include:

   - Jonckheere-Terpstra test for ordered differences
   - Fligner-Wolfe test for control vs treatments

   The following pairwise comparison tests:

   - Dwass-Steel-Critchlow-Fligner
   - Conover-Inman

9. Polynomial Regression

SYSTAT offers polynomial regression on a single independent variable up to order 8:

- In natural form or in orthogonal form.
- Goodness-of fit-statistics (R2 and adj R2) and ANOVA with p-values for all models, starting from the order specified by the user, down to linear (order=1).
- Confidence and prediction interval plots along with estimates, and a plot of residuals vs. predicted values, as quick graphs.

Modified Features

10. Analysis of Variance

The Analysis of Variance feature now provides:

- Levene's test based on median for testing homogeneity of variances.
- A SUBCAT command that categorizes the desired factors just for the purpose of the analysis.

11. Crosstabulation

As part of its Crosstabulation feature, SYSTAT now offers:

- Relative Risk: In a 2 x 2 table, the relative risk is the ratio of the proportions of cases having a `positive' outcome in the two groups defined by row or column. Relative Risk is a common measure of association for dichotomous variables.
- Mode: SYSTAT gives an option to list only the first N categories in a one-way table (frequency distribution). This is done by adding a MODE = N option to the PLENGTH command within XTAB.
- Saved results with:
  - all requested columns in Multiway: Standardize
  - value labels of the input variables for the corresponding columns of the saved results file.
- Output categorized appropriately based on the type of table, and reorganized table of measures..

12. Cluster Analysis

In Cluster Analysis, the data file containing the saved results will preserve the value labels, if any, from the input data file.

13. Fitting Distributions

SYSTAT now performs the estimation of parameters for the beta, chi-square, Erlang, gamma, Gompertz, Gumbel, logistic, log-logistic, negative binomial, Weibull and Zipf distributions using the maximum likelihood method.

14. Hypothesis Testing for Two-Sample Data in Columns

For two-sample z, two-sample t, and test for two variances, option for input data in a layout where the data across the samples appear in different columns. This is in addition to the current indexed layout.

15. Least Squares Regression

The following enhancements are available in the Least Squares Regression feature:

- Save Standard Errors and Confidence Intervals in Least Squares Regression.
- A choice of bootstrapping residuals. Bootstrap Estimates of the Regression Coefficients, Bias, Standard Error and confidence intervals are then computed based on these.

16. Logistic Regression

SYSTAT provides the following enhancements to its Logistic regression feature:

- Simplified user interface and command line structure to analyze binary, multinomial, conditional, and discrete choice models separately.
- Option to specify the reference level for the binary and multinomial response models.
- Simpler form of input data to analyze matched sample case-control studies with one case and any number of controls per set.
- Discrete choice model provides two data layout inputs: Choice set and BY choice to model an individual's choices in response to the characteristics of the choices.
  - In the raw data layout choice set names for groups of variables can be defined, and variables can be created, edited, or deleted.
  - In the by choice framework, the choices sets already defined can be used in the data for the analyses.

17. Mixed Models

The Mixed Models feature performs significantly faster than in prior versions.

Chapter

2

Introducing SYSTAT

Keith Kroeger
(revised by Rajashree Kamath)

SYSTAT provides a powerful statistical and graphical analysis system in a graphical environment using descriptive menus and simple dialog boxes. Most tasks can be accomplished simply by pointing and clicking the mouse.

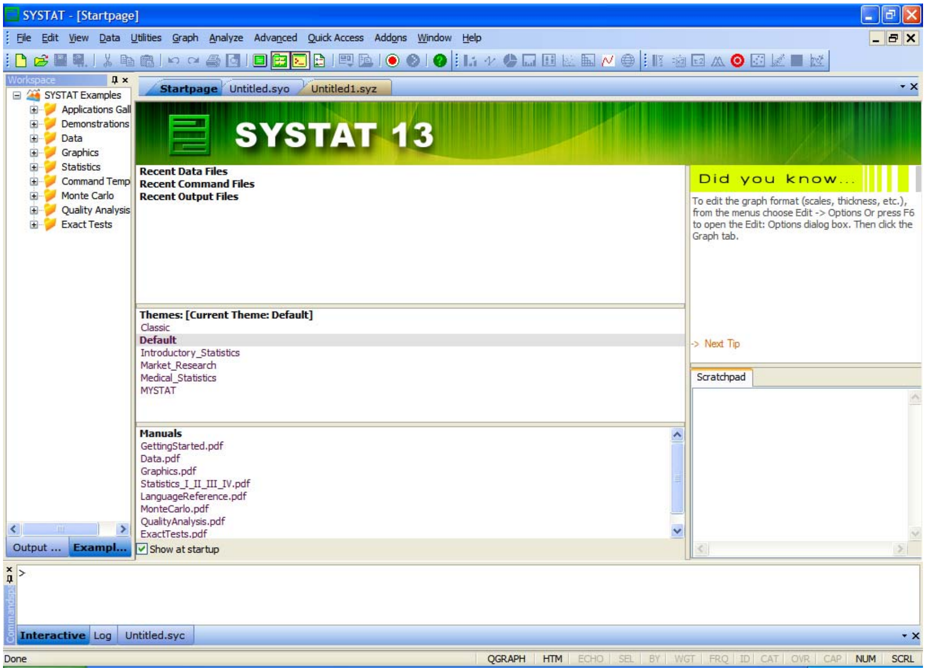
This chapter provides an overview of the windows, menus, dialog boxes, and Online Help available in SYSTAT. For information on using SYSTAT's command language, see Chapter 5.

User Interface

The SYSTAT window is made up of three panes, which we term as:

- Workspace
- Veiwspace
- Commandspace

Each pane consists of various tabs, or sets of tabs, and allows you to accomplish specific tasks. One pane, and one tab within it, will always be in focus. At any given moment, certain menu selections and their corresponding keyboard shortcuts (like Ctrl+C for copy) apply to the tab and/or pane that has the focus. To bring a pane into focus, click any of its constituent tabs. To bring a tab into focus, click it with the mouse, or select its name from the View menu. The user interface provides menus for running statistical analyses and producing graphs. It also contains toolbars to provide quick access to many standard statistical techniques and graphs.



Viewspace

The Viewspace consists of four components:

- Startpage
- Output editor (untitled .syo upon opening)
- Data/Variable editor (untitled .syz upon opening)
- Graph editor (graph1, when graph is in the Output editor).

Startpage. The Startpage is typically the first tab in the Viewspace, and it is divided into five panes:

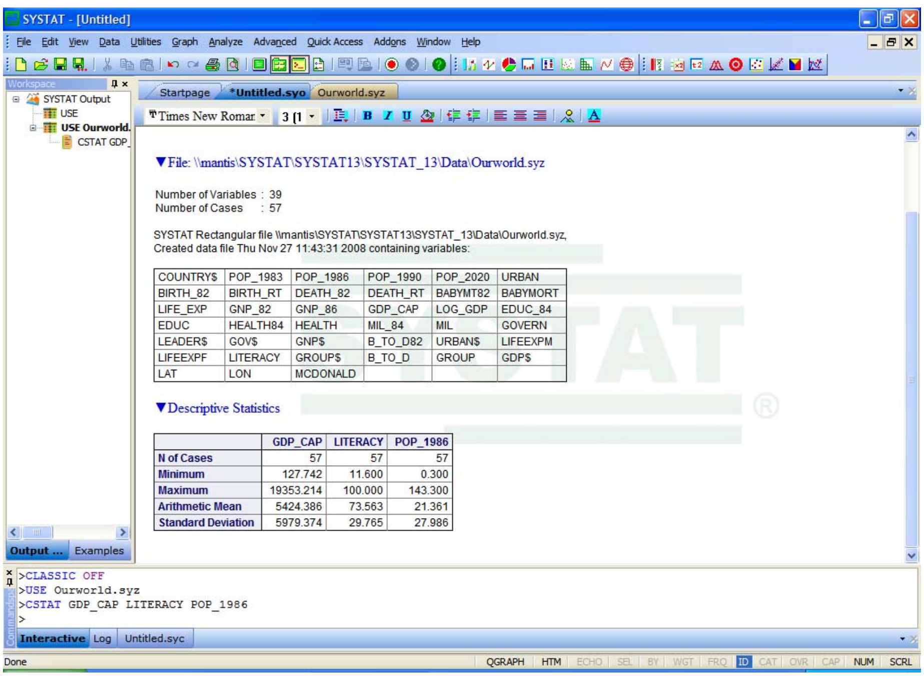
- Recent Files containing a list of all the recently opened data, command and output files; you can reopen these files just by double-clicking on their names.
- Themes contain a list of menu themes; double-click any one to apply it to the SYSTAT window.

- Manuals containing a list of the user manual documents; you can open the desired volume by double-clicking on its name.
- Tips providing useful tips about SYSTAT's features and how to achieve any given task; clicking Next Tip will allow you to scroll through any number of tips.
- Scratchpad for writing notes while you are working with SYSTAT. Anything that you enter here remains across sessions.

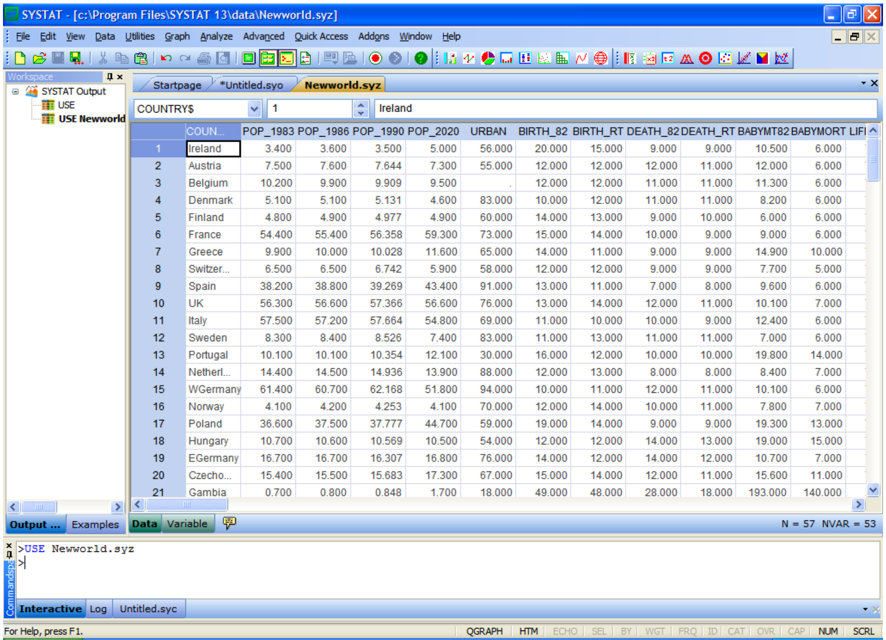
You can click on the bar at the top of the Startpage to know about the new features in the current version of SYSTAT. You can close the Startpage if you do not need it for the remainder of a session, or even prevent it from appearing when SYSTAT restarts.

Output editor. Graphs and statistical results appear in the Output editor. Collapsible links are created for each analysis or graph that you request. You can thus hide output that you do not need to see all the time. Simply click on the link once to collapse the corresponding output; click again to expand it.

You can perform some of the Output editor-related operations using the Format Bar that is embedded in the Output editor. For more information about the Output editor, see Chapter 6.



Data editor. The Data editor displays your data in a row-by-column format.



Each row is a case and each column is a variable. You can type new data into an empty Data editor, or you can edit and transform data.

- To define a variable, right-click on a column and choose Variable Properties. This opens the Variable Properties dialog box and allows you to name the variable, supply a label for it, select the variable type, indicate whether it is categorical, set display options, and specify comments.
- Use the Edit menu to cut, copy, delete, and paste rows, columns, and blocks of data.
- Use the Data menu to transform data and select subsets of cases.

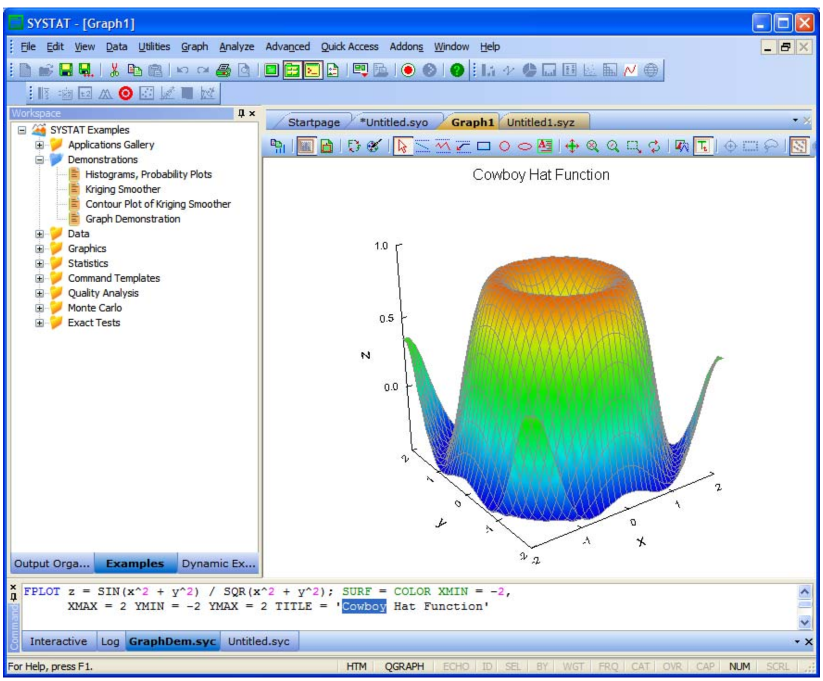
The data file that you create or open for use is called the active data file.You can open any number of data files using the File menu; a new tab is created in the Data editor for each file that you open. The currently active file automatically goes into the view mode when you create or open another file. You need to make it active only if you want to perform any data transformation or analyses on it. You can make a data file active using its context menu or the Output Organizer. You can thus have any number of data files available in the Data editor ready for use at just a click of the mouse.

Variable editor. Each data file, active or inactive, has a Data tab and a Variable tab. The Data tab allows you to edit data values directly in the grid that you see by default. The Variable tab allows you to edit the properties of variables directly. We will henceforth refer to the Variable tab as the Variable editor. The Variable editor has one row corresponding to each variable, and the row includes all the items that are in the Variable Properties dialog. With it, you can:

- Set any of the properties for any variable with a single click of the mouse.
- View and set the processing conditions in effect for the current data set, *viz*. information regarding frequency, weight, category and grouping variables defined if any, and any case selection conditions.

You can navigate to any specified column or row of the Data editor and veiw/edit the value stored in any cell using the Data Edit bar that is embedded in the Data editor. See Chapter 3: Entering and Editing Data of the *SYSTAT Data* volume for more information about the Data editor.

Graph editor. Double-clicking a graph in the Output editor or just clicking the Graph tab after drawing a graph opens the Graph editor.



You can perform many of the Graph editor-related operations using the Graph Editing toolbar that is embedded in the Graph editor. Use that and the menus to:

- Insert annotations and other text.
- Change font, color, fill, surface and line attributes.
- Rescale axes.
- Modify plot symbols.
- Customize labels.
- Edit legends.
- Identify individual points in scatterplots.
- Select a subset of cases using the Rectangular or Lasso tool.
- Zoom and rotate graphs.
- Change many other properties of a graph like changing its type, drawing various smoothers, specifying gradients for surfaces, connecting and partitioning plot points, slicing pie charts, and setting attributes for each individual axis line.

You can view any number of graphs using the context menu of the Output Organizer. See *SYSTAT Graphics* for more information about the Graph editor.

By default, the tabs of the Viewspace are arranged in the following order:

- Startpage
- Output editor
- Graph editor
- Active Data File
- Inactive Data Files

When a new tab is opened, it is inserted at the beginning of its group. You can click the arrow in the top right corner of the Viewspace and check [Active Tab at the Beginning] if you want a new tab to appear as the first tab of the Viewspace. You can bring a tab into focus by clicking the arrow and checking the name of the desired tab. If there are more tabs than are directly visible in the Viewspace, the tab becomes the first tab in the Viewspace or in its group depending on whether [Active Tab at the Beginning] is checked or not. This is especially useful when you have a lot of tabs open in the Viewspace.

You can close an active or inactive data file by right-clicking and selecting Close or by bringing the tab into focus and pressing the Close button in the top right corner of the Viewspace.

Workspace

The Workspace consists of three tabs:

- Output Organizer
- Examples
- Dynamic Explorer

Output Organizer. Use the Output Organizer primarily to navigate through the results of your statistical analysis. Selecting a completed procedure from the outline displays the corresponding results in the Output editor. You can also use the Output Organizer to select an item, and then copy, paste, delete, or move it, allowing you to tailor SYSTAT's output to your preferences. In addition, you can quickly move to specific portions of the output without having to use the Output editor scrollbars.

For more information about the Output Organizer, see Chapter 6.

Examples. Use the Examples tab to conveniently execute command scripts given in the user manual with just a click of the mouse. The SYSTAT Examples tree is organized by folders and nodes, the folders corresponding to each volume of the user manual. Double-click the nodes to run the underlying commands. You can also open these command scripts in the Commandspace for editing, and create links to your own command files for easy execution. You can even add example nodes to this tab using the Utilities menu.

See Chapter 5 to know more about the Examples tab.

Dynamic Explorer. The Dynamic Explorer becomes active when there is a graph in the Graph editor, and the Graph editor is active. Use the Dynamic Explorer to:

- Rotate and animate 3-D graphs.
- Zoom the graph in the direction of any of the axes.

See *SYSTAT Graphics* for more information about the Dynamic Explorer.

Commandspace

The Commandspace has three tabs:

- Interactive
- Batch (Untitled)
- Log

Interactive. Selecting the Interactive tab enables you to enter commands in the interactive mode, which issues the command after you press the Enter key. You can save the contents of the interactive tab (excluding the > prompts) and then use the file to submit a sequence of commands.

Batch (Untitled). Selecting the Batch (Untitled) tab enables you to work with command files in the batch mode. You can open any number of existing command fiels, and edit or submit any of these files. You could also type in an entire set of commands and then save or submit it, The name that you specify while saving any content that you may have typed here replaces the caption 'Untitled' on the tab.

Log. Selecting the Log tab enables you to examine the read-only log of the commands that you have run during your session. You can save the command log or even submit one or more of the generated commands.

By default, the tabs of the Commandspace are arranged in the following order:

- Interactive
- Log
- Command Files

When a new tab is opened, it is inserted at the beginning of its group (Batch). You can click the arrow in the bottom right corner of the Commandspace and check [Active Tab at the Beginning] if you want a new tab to appear as the first tab of the Commandspace. You can bring a tab into focus by clicking the arrow and checking the name of the desired tab. If you have opened more than 9 command files, the tab becomes the first tab in the Commandspace or in its group depending on whether [Active Tab at the Beginning] is checked or not. This is especially useful when you have a lot of tabs open in the Commandspace.

You can close the tab in focus by right-clicking and selecting Close or pressing the Close button in the bottom right corner of the Commandspace. You can close all open command files by right-clicking in any tab of the Commandspace and selecting Close All.

Reorganizing the User Interface

The Workspace,Viewspace and Commandspace can be resized if desired. To do so:

- Drag the boundaries of the panes (between Viewspace and Workspace, Workspace and Commandspace, and Viewspace and Commandspace) in the desired direction.

You can also reposition the panes. For this:

- Click the upper boundaries of the panes and drag the resulting outline to the new position. As you drag the outline, the border thins to indicate that the item will be docked to the main window at that location. To prevent docking, drag the item off the main window or hold down the Ctrl key as you drag. Double-clicking the upper boundary can undock docked items. Undocking items enlarges the remaining panes but can result in a cluttered desktop.

You can collapse the Workspace and Commandspace so that they are only visible when you pause the mouse on the corresponding vertical bar at the edge. To do this, click the at the top right corner of the pane.

The tabs of the Viewspace can be tiled so that you can view any two of the tabs simultaneously. To do this:

- Click the Window menu or right-click on the toolbar area and select Show Stacked or Show Side-by-Side All the panes in the Viewspace get laid out in a tiled fashion. Double-click one of the title bars to dock the panes to their default or previously docked positions.

Every toolbar can be repositioned by clicking and dragging the move handle . Toolbars can also be dragged and docked to the boundary between the Viewspace and Workspace. The Format Bar, Data and Graph Editing toolbars can be toggled by right-clicking on the Output editor, Data editor and Graph editor tabs respectively and selecting Show Toolbar.

You can also close the Workspace, Commandspace and toolbars so that more space is available for viewing the output, data and graphs. To do so:

- undock them and click in the upper right corner, or deselect their entry on the View menu. Closed items can be reopened via the View menu or using the keyboard. Keyboard short cuts are explained in Chapter 7.

Menus

SYSTAT has a common menu bar for all the panes and tabs. There are menus for opening, saving, and printing files, editing output, transforming data, matrix manipulation, generating experimental designs and random samples, performing statistical analyses, and creating graphs. At any given point of time, those menu items that are relevant to the active pane or tab are enabled. The menu can be customized using the Customize dialog from the View menu.

File. Use the File menu to create or open data, command and output files, import from databases, and save the contents of the active pane, all panes or newly created data files. The data file formats supported include SYSTAT, Excel, SPSS, SAS, MINITAB, S-PLUS, Statistica, Stata, JMP and ASCII files. You can save command files or the command log, and submit commands that are in the Commandspace, a command file, the Windows clipboard, or from a command file list. You can save output in the SYSTAT (.syo), or HTML (.mht) formats. You can also define page and printer settings, preview and print the content of the Output editor or Data editor, and Graph editor. Graphs can be reviewed using the Page Mode under the View menu. When the Graph Editor is active, you can also export and print graphs. You can export graphs in a variety of formats including WMF, PS, EPS, BMP, JPEG, GIF, TIFF, PNG, and PCT. The File menu can also be used to open recent data, commands, and output files.

Edit. Use the Edit menu to undo/redo a few steps, paste clipboard content to the active pane, define output related settings like ID variables, order of display of data values, and display of variable as well as value labels, change SYSTAT options including variable display order in dialog boxes, the algorithm to be used for random number generation, the behavior of the Enter key in the Data editor, font characteristics for output, data and graphs, display of statistical Quick Graphs, inclusion of command syntax in the output, and measurement units for graphs, reduction or enlargement of graphs, and file locations.

- **Output editor.** In addition to the above options, when the Output editor is active, you can undo/redo a few steps of output, cut, copy, and paste statistical output and other text from and into the Output editor, find and replace text strings, clear text and output, change font characteristics (including color and size), create numbered and bulleted lists, outdent/indent text, align text, tables and graphs, insert images and page breaks into your output, and collapse/expand links created by graphical and statistical procedures.
- **Data editor.** When the Data editor is active, you can also undo/redo up to 32 data editing operations, cut, copy and paste data from and into the Data editor, add

empty rows in a new or existing data file, insert/delete cases and variables, find a specific variable, find/replace occurrences of a string or number in any given column, and go to a desired cell.

- **Graph editor.** When the Graph editor is active, you can also copy graphs.
- **Output Organizer.** When the Output Organizer is active, you can also cut, copy, paste and insert tree folders, set the selected data file node as active, rename nodes, expand/collapse trees and see detailed node captions.

View. Use the View menu to view or hide the Workspace, Commandspace, Startpage, processing conditions, toolbars and status bar, make tabs active, and launch a full screen view of the Viewspace. This menu also allows you to create and customize toolbars, keyboard shortcuts and context menus. When the Output editor is active, you can also view graphs as frames only. When the Graph editor is active, use the View menu to switch between the Graph View and Page View, and turn the display of rulers and graph tooltips on and off.

Data. Use the Data menu to define categorical variables, transform (including recode) data values, rank, center or standardize data, trim extreme values, sort cases in the data file based on the values of one or more variables, transpose cases (rows) and variables (columns), wrap/unwrap or stack variables, merge data files (cases or variables), define ID variables and order of display of data values, specify grouping variables that split the data file into two or more groups for analysis, select and extract subsets of cases, list data in the Output editor, define case frequencies, and weight data for analysis based on the value of a weight variable. When the Data editor is active, you can also define variable properties and value labels, as well as edit data.

Utilities. Use the Utilities menu to access SYSTAT's MATRIX module, perform probability calculations, generate random samples from a variety of univariate discrete and continuous probability distributions, generate a variety of experimental designs, perform power analysis and calculations involving functions available in SYSTAT (including probability calculations), retrieve data file information and current SYSTAT settings, record macros i.e. command scripts generated by actions of the user and play them, create command file lists and customized user menus, access recently invoked dialogs, save, apply and download SYSTAT menu themes, as well as add examples to the Examples tab.

Graph. Use the Graph menu to access the Graph Gallery and to create function plots, summary charts like pie, doughnut bar, line, profile, pyramid, cone, cylinder and high-low-close, density displays like histograms, dot densities and box plots, distribution plots like density functions, probability plots and quantile plots, scatterplots,

scatterplots matrices, parallel coordinate displays, Andrews's Fourier plots, icon plots and maps.You can also overlay various graphs in a single frame. When the Graph editor is active with a graph in it, you can realign any displaced graph frames with their original positions, edit various properties of the graph like font attributes of graph/frame titles, axes, tick mark, bar and case labels, zoom, rotation, layout (position, size and arrangement), title, background color, type (for summary and density charts), and coordinate system of graphs, axes/scale type, tick mark style and location, label, limit lines, grid lines, transformations, line style and scale ranges on the graph's axes, titles, labels, location and layout of graph legends, colors and fill patterns for the graph's elements, style and size of plot symbols, surface, gradient and wireframe styles, and various options for each graph type. The Graph menu also allows you to copy graphs, define text annotation font and graph annotation attributes, select the pointer tool or any of the annotation tools, select the panning or zooming tools, reset any panning or zooming done to a graph, highlight a point in a plot to view the corresponding case in the Data editor, choose the region or lasso selection tools, and show or hide any selection made using these tools in the plot.

Analyze. Use the Analyze menu to run fundamental statistical analyses including crosstabulation, column and row basic statistics and stem-and-leaf plots, fitting distributions, correspondence analysis, loglinear models, nonparametric and multinormal tests, hypothesis testing, (univariate tests and Hotelling's T-square tests), simple as well as set and canonical correlations, Cronbach's alpha, linear and robust regression methods,logistic regression, probit analysis, two-stage least squares, mixed as well as nonlinear regression methods, nonparametric smoothing, univariate and multivariate analysis of variance, general linear models, mixed models, discriminant (classical and robust), cluster as well as factor analyses, (exploratory and confirmatory), plotting, transforming, and smoothing time series, autocorrelation and cross correlation functions, seasonal adjustment, ARIMA, ARCH tests, GARCH, trend analysis, and Fourier transformation.

Advanced. Use the Advanced menu to perform advanced statistical analyses like missing value analysis, quality analysis (including Pareto, Box-and-Whisker, various control charts like Shewhart and X-MR, ARL and OCC computation, and process capability analysis), nonparametric, Cox and parametric survival analysis, response surface methods (estimation, optimization and plotting), path analysis, conjoint analysis, multidimensional scaling, perceptual mapping, partially ordered scalogram analysis, test item analysis, signal detection analysis, network analysis, spatial statistics, and C&RT.

Quick Access. Use the Quick Access menu to quickly access all the commonly used statistical procedures.You may want to customize this menu to contain those analyses that you frequently use so that you may access all of them in a single location.

Window. Use the Window menu to cascade, stack, show side-by-side, or arrange the tabs of the Viewspace.

Help. Use the Help menu to access SYSTAT's online Help system (Contents, Index or Search, Acronym Expansions), Frequently Asked Questions (FAQ), demos and tutorials on various SYSTAT features, a Quick Reference guide on SYSTAT commands, and a list of new and modified commands. Through this menu, you may also update the license for running SYSTAT beyond the specified period, check for updates to the current version of SYSTAT, access the SYSTAT website, and display the copyright, version number and license information of your copy of SYSTAT.

Context Menus

SYSTAT provides several context menus that appear on right-clicking in various components (tabs or nodes in the three panes) of its interface.The available menus are listed below with a brief description of each.

Startpage. You can specify whether you want the Startpage to show at startup, clear recent data, command and output files that are listed in the Recent Files quadrant, refresh the content of the Startpage, close it for the rest of the session, and invoke the Edit: Options dialog box.

Output editor. You can cut or copy the selected content in the Output editor to the Windows clipboard, paste content from the clipboard to the Output editor, copy all the content in the Output editor to the clipboard, view the HTML source, refresh, or preview the content for printing, collapse/expand links in the output, show the Format Bar, create a new output file, clear all or save the content in the Output editor, and invoke the Edit: Options dialog box.

Data/Variable editor. You can copy all the content in the Data editor, set one of the inactive data files in the Data editor as the active data file, switch between the Data and Variable editors, enter or view and edit comments for a data file, show the Data Edit bar and Data toolbars, create a new data file, save data files, invoke the Edit: Options dialog box, close a data file, and show the processing conditions in effect (if the Variable editor is active).

Graph Editor. You can invoke the Graph Properties dialog box, animate a 3-D graph, realign any graph frames you may have moved from their original positions, copy or preview (for printing) the graph in the Graph editor, show the Graph Editing toolbar, save the graph that is in the Graph editor, and invoke the Edit: Options dialog box.

Output Organizer. You can rename tree nodes and folders, expand or collapse the entire tree including any tree folders or multilevel nodes, insert tree folders, create a new output file, clear all or save the content in the Output editor, and request detailed node captions.When a data node is selected, you can also set it as the active data file.When a text node is selected, you can also cut or copy it (and the corresponding output in the Output editor) to the clipboard, paste one or more nodes after copying them to the clipboard, or even delete it (which will also delete the corresponding content in the Output editor).When a graph node is selected, you can also view the corresponding graph in the Graph editor.

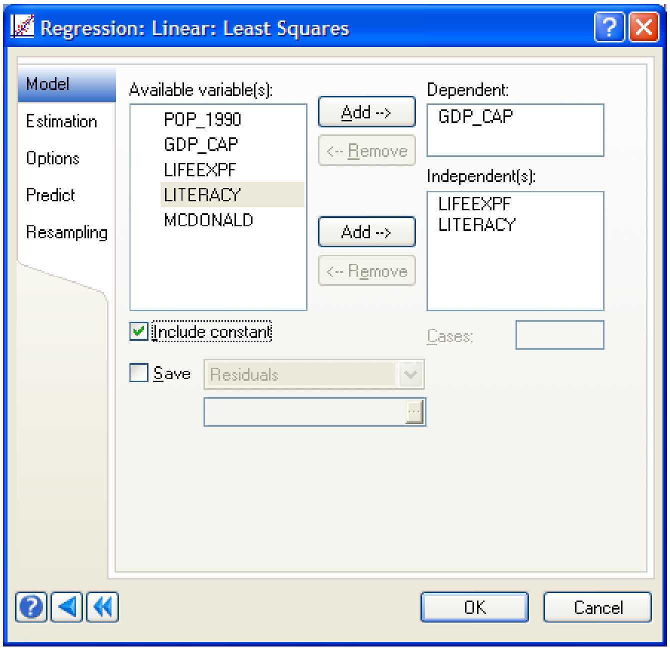
Examples. You can run the underlying example command file(s), expand or collapse the entire tree including any sub-folders or multilevel nodes. When an example node (not folder) is selected, you can also open the underlying command file in the Batch tab of the Commandspace.

Commandspace. Apart from the various options for editing and submitting commands, you can right-click on the Batch tab to create a new command file, open an existing command file, save the content of the tab, or close the tab.

In addition to these, context menus are available for cells , columns and rows in the Data editor, command files in the Batch, interactive and log tabs of the Commandspace, dialog box elements, status bar and the toolbar area. These menus provide shortcuts to various data editing, command submission, dialog actions, status bar content and menu actions respectively.

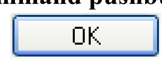
Dialog Boxes

Most menu selections in SYSTAT open dialog boxes, which you use to select variables and options for analysis. Each dialog box may have several basic components in separate tabs.



Tabs. Since many SYSTAT commands provide a great deal of flexibility, not all of the possible choices can be contained in a single dialog box. The main dialog box usually contains the minimum information required to run a command. Additional specifications are made in tabs. You can bring the content of a tab into view by clicking it with the mouse. Certain tabs require some input to be given in other tabs before they get enabled. A tab may get disabled if its contents are irrelevant for the existing selections.

Command pushbuttons. Buttons that instruct SYSTAT to perform an action.

- 

Runs the procedure for the selections you have made. This does not get enabled in some dialog boxes unless the minimum required input is given.
- 

Cancels the procedure. Any selections you may have made will be discarded.

-  Displays help related to the dialog box. If a dialog box has more than one tab, you will get help related to the active tab.
-  Resets the selections in the dialog box or active tab, to the defaults.
-  Resets the selections for all tabs in the dialog box.

Source variable list. A list of variables in the working data file. Only variable types (numeric and/or string) allowed by the selected command are displayed in the source list.

Target variable list(s). One or more lists, such as dependent and independent variable lists, indicating the variables you have chosen for the analysis. If an analysis compulsorily requires you to choose variables here, you will see '<Required>' in the list. If a list is empty, all variables in the source list will be used for the analysis.

Special lists. Some dialog boxes display lists with multiple columns, where you can input as many rows of input as you desire. Such lists can be customized using the two buttons:

- Insert a new row by pressing the  icon.
- Delete a row by pressing the  icon.

Pushbuttons. Dialog boxes contain pushbuttons for performing the following tasks:

- Add one or more variables to the desired target list by selecting them and then pressing the corresponding Add --> button. Alternatively, right-click on a variable or selection and select the "Add to *target list*" corresponding to the desired target list.
- Remove one or more variables from a target list by selecting them and then pressing the corresponding <-- Remove button. Alternatively, right-click on a variable or selection and select Remove.
- 'Cross' a variable in the source list with one in the target list by selecting them and then pressing the Cross --> button. You can also add crossed terms of multiple variables directly by selecting these variables in the source list and pressing the Cross button.
- Use the # --> button when you want to include the variables as well as all their crossed terms. You can also use this button with multiple variables.
- Use the Nest --> button to include nested terms in the target list.

Selecting variables. To add a single variable to the desired target list, you simply highlight it in the source variable list and click the Add --> button. Use the <-- Remove button to undo your selection. You can also double-click individual variables to move them from the source list to the target list, or vice versa. When there is more than one target list, this functionality will apply to one of them.

You can also select multiple variables:

- To highlight multiple variables that are grouped together on the variable list, click and drag the mouse cursor over the variables you want. Alternatively, you can click the first one and then Shift- click the last one in the group.
- To highlight multiple variables that are not grouped together on the variable list, use the Ctrl-click method. Click the first variable, and then Ctrl-click the other variables that you want. Avoid the name area while clicking and dragging.
- To select all the variables in a list, click inside the list and press Ctrl + A, or right-click and select Select All.

You can also right-click on a variable or a highlighted set of variables and use the menu that pops-up to add them to the desired target list, or remove them from the list.

Additional Features. Several additional features have been provided for the dialog boxes. They are:

- Keyboard shortcuts as an alternative to check boxes and radio buttons. Hold down the Alt key and press the underlined letter in the caption.
- The Tab key to navigate between items.
- For an edit text taking numeric values, tooltips indicating the valid range, displayed while pausing the mouse on the edit text.
- Edit texts taking integer values not accepting the decimal separator as input.
- Edit texts taking nonnegative values not accepting the negative (-) sign as input.
- Edit texts to contain filenames of files to be opened or saved, for features that require or support such options. Type the desired filename (with path), or press the ... button and select a file.

Getting Help

SYSTAT uses the standard HTML Help system to provide information you need to use SYSTAT and to understand the results. This section contains a brief description of the Help system and the kind of help provided with SYSTAT.

The best way to find out more about the Help system is to use it. You can ask for help in any of these ways:

- Click the  button in a SYSTAT dialog box. This takes you directly to a topic describing the use of the dialog box. This is the fastest way to learn how to use a dialog box.
- Right-click on any dialog box item, and select 'What's this?' to get help on that particular item.
- Hover the mouse on a menu item that would have opened a dialog box and press F1 to get help on that particular dialog box.
- Select Contents or Search from the Help menu.
- For help on any term or phrase that is listed in the Help Index, from the command prompt (on the Interactive tab of the Commandspace) type:

```
HELP "[phrase]"
```

The quotes are required only if the phrase contains spaces. This is very useful if you need help on SYSTAT commands. Refer the *Command Language* chapter for details.

Alternatively, type the term or phrase in any tab of the Commandspace, right-click on it and select HELP *phrase*. You will need to select the whole phrase before you right-click if it contains spaces.

Navigating the Help System

The SYSTAT Help system has the following tabs:

- **Contents.** The Contents button takes you to the table of contents of the Help system. Double-click book icons  in the Index listing to view the contents of that section. Selecting a topic with a page icon  opens the associated Help topic.
- **Index.** Provides a searchable index of Help topics. Enter the first few letters of the term you want to find and then double-click the topic in the list (or click and press the Display button) to view it.

- **Search.** Offers a full-text search of the Help system. Type the desired keyword and press the Enter key or the List Topics button. The Help system returns all topics containing the specified term. Double-click the desired topic in the list (or click and press the Display button) to view it. Check Search previous results to search for the keyword from within the previously listed topics. By default, all word forms of the keyword are located. Uncheck Match similar words if you want just the exact keyword to be located. Check Search titles only if you want to confine the search to the page titles alone.
- **Favorites.** Allows you to create and use a list of favorite help topics. The topic that you are currently viewing will automatically appear in the Current topic. You can either press Add to add this topic to the list, or you can type in a page title that you know exists in the Help system and then press Add. Select a topic in the list and press the Display button (or the Enter key) to view the topic. Use the Remove button to remove a selected topic from the list.

The following buttons are available in the toolbar of the Help system:

- **Hide/Show.** Hides or shows the Contents, Index and Search tabs.
- **Back.** Returns to the previous Help topic.
- **Forward.** Moves to the next Help topic, if you had pressed the Back button previously.
- **Stop.** Stops loading a page.
- **Refresh.** Refreshes the currently loaded page.
- **Home.** Loads the SYSTAT Help Copyright page.
- **Print.** Prints the current topic or all sub-topics under the current heading when you click this with the Contents tab active. When any other tab is active, use this to print the current page. Before printing, the Print dialog pops up so that you can specify the desired print settings.
- **Options.** Enables you to do any of the above, access the Windows Internet Options settings, or specify whether you want search keywords to be highlighted in the listed pages or not.

Depending on the topic displayed, the following buttons may appear in the current Help page:

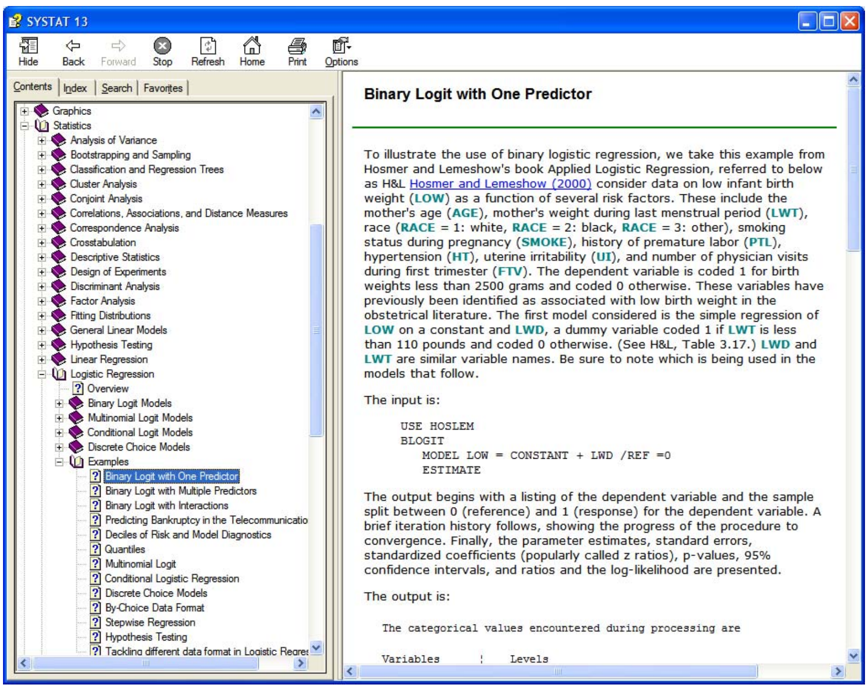
- **How To.** Provides minimum specifications for performing the analysis.
- **Syntax.** Describes the associated SYSTAT command. SYSTAT's command language offers some features not available in the dialog boxes.

- **Examples.** Offers examples of analyses, including SYSTAT command input and resulting output. Copy and paste the example input to the Batch tab of the Commandspace to submit the example as is, or modify the commands to your own analyses before submitting them. Make sure the file paths match the file locations you have opted for.
- **More.** Lists analysis options and related tabs. These topics are particularly useful for customizing your analyses.
- **See Also.** Lists related procedures or graphs.

You can select, cut, copy, paste and print the content of any Help page.

Examples

Often, the best way to learn about a procedure is through examples.The Help system provides several examples for each statistical procedure or graph. Select the example most relevant to your analysis or browse the examples to explore SYSTAT's capabilities.



The examples include all SYSTAT input. You can copy and paste the example input (also available as files in the 'Command' folder of the SYSTAT directory and having links in the Examples tab of the Workspace) to the Batch tab of the Commandspace to submit the example as is, or you can modify the commands to reflect your own analyses before submitting them.

The resulting output, including graphical results, follows the command input. Many of the examples include Discussion buttons throughout the output. Pressing any of these buttons yields a detailed explanation of the immediately preceding output. There may also be examples that are explained in more than one step, in which case More or Next buttons will be included in the page.

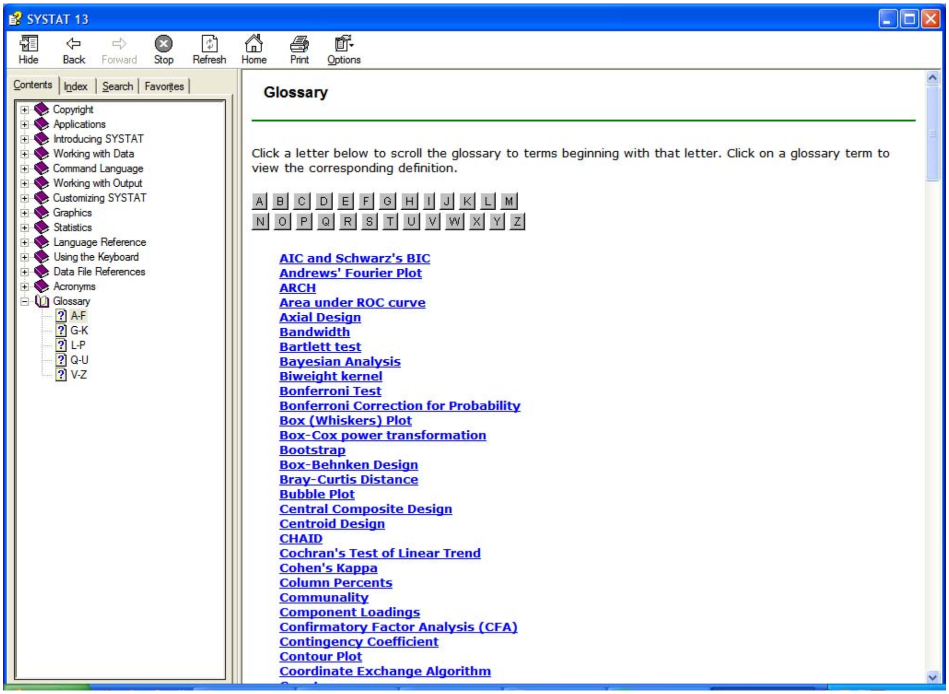
Example Command Files. The input commands for each example in the User Manual or in the Help system are available as command files in the "Command" folder of the SYSTAT directory. This provides an alternative way to run the examples. These files are organized in terms of the printed manual. Each file contains commands for one example and is named using six characters (xxyyzz.syc). The first two characters represent the corresponding volume of the printed manual as follows:

- 'da' for Data (called 'Data Volume' in the Command folder)
- 'gs' for Getting Started
- 'gr' for Graphics
- 's1' for Statistics I
- 's2' for Statistics II
- 's3' for Statistics III
- 's4' for Statistics IV
- 's5' for Quality Analysis (if installed)
- 's6' for Monte Carlo (if installed)
- 's7' for Exact Tests (if installed)

The next two digits represent the chapter number within the volume, and the last two digits represent the example number within the chapter. These files are organized in the 'Command' folder with nine subfolders, seven of them corresponding to the seven volumes mentioned above, a 'GraphDemo' subfolder and a 'Miscellaneous' one which contains commands of examples which are not numbered. The names of files in the 'Miscellaneous' folder are indicative of the examples they relate to. For example, to execute the commands given in Example 1 in Chapter 2 of *Statistics III*, submit the 's30201.syc' file. (Depending on your file location, you may have to define paths for files and rename them appropriately.)

Glossary

The glossary offers an alphabetical listing of terms commonly encountered in statistical analyses. The buttons at the top of the glossary scroll the window to the corresponding letter. Clicking a glossary entry reveals the definition for that term.



Application Gallery

In addition to examples of each procedure, SYSTAT includes examples drawn from several fields of research. Chapter 8 provides a brief introduction to each application. You can access the complete applications from the Contents tab of the Help system. Double-click the Applications book icon and select Application Gallery. The available applications are listed with icons and a brief description. Clicking on any icon will open a page containing the detailed description, and buttons for the main Application Gallery page, Analyses page, and Sources page.

Chapter

3

SYSTAT Basics

This chapter provides simple step-by-step instructions for performing basic analysis tasks in SYSTAT, including:

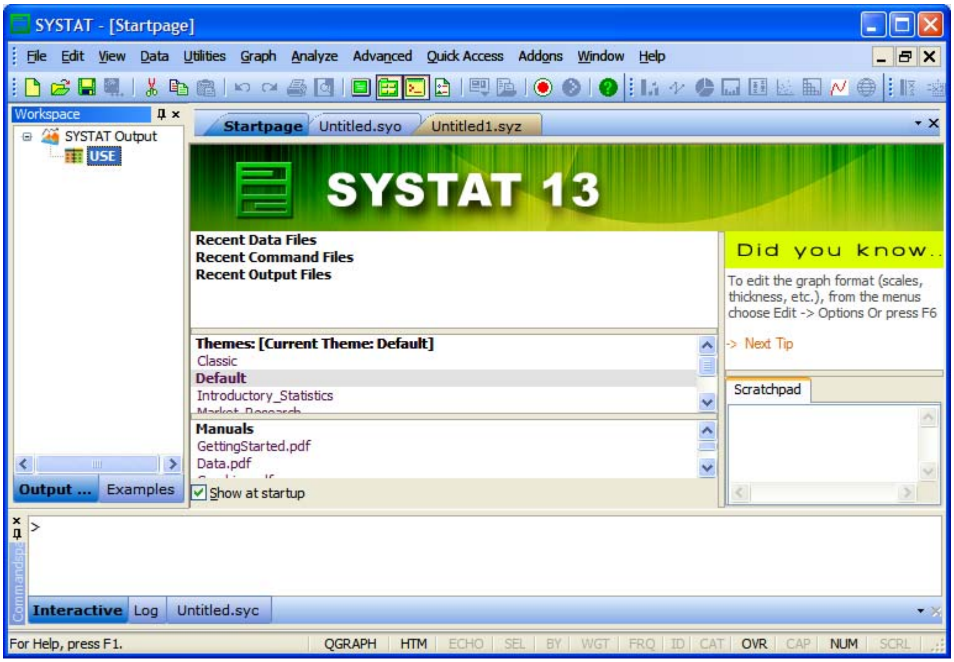
- Starting SYSTAT.
- Entering data in the Data Editor.
- Opening and saving data files.
- Using menus and dialog boxes to create charts and run statistical analyses.

Starting SYSTAT

To start SYSTAT for Windows XP, 2000, ME, and NT4:

- Choose:

Start
  Programs
    SYSTAT 13
      SYSTAT 13...



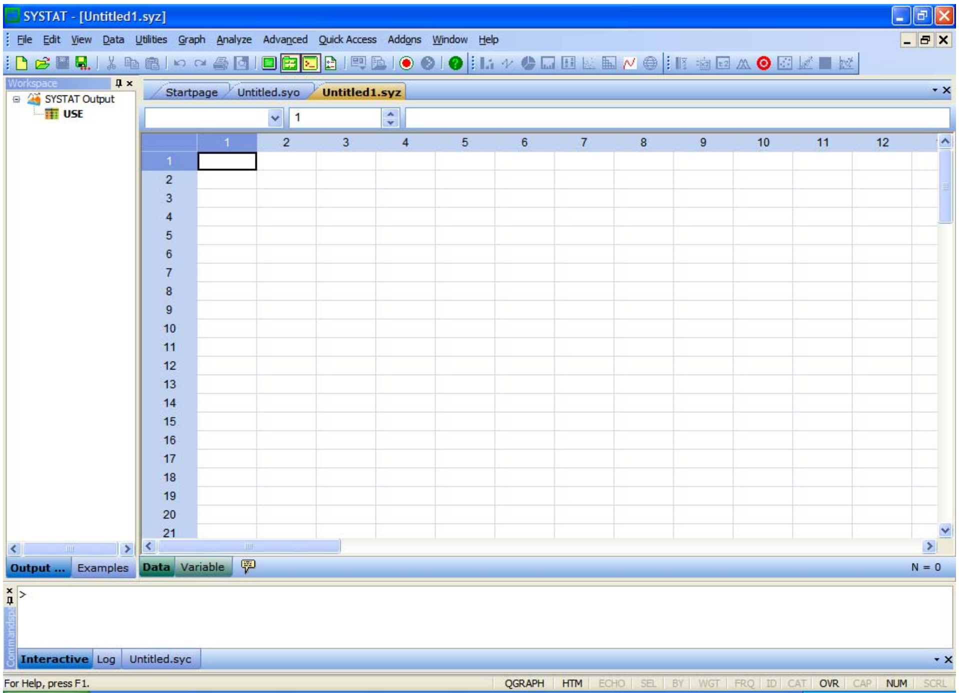
Entering Data

This section discusses how to enter data. If you prefer to start with data stored in a text file, see "Reading an ASCII Text File" on p. 51.

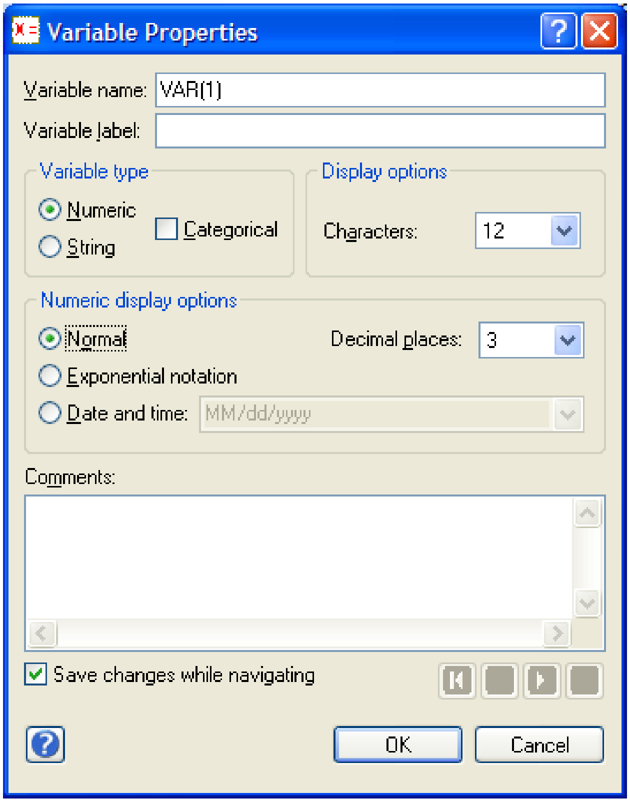
In the frozen-food section of the grocery store, we recorded this information about seven dinners:

Brand$	**Calories**	**Fat**
Lean Cuisine	240	5
Weight Watchers	220	6
Healthy Choice	250	3
Stouffer	370	19
Gourmet	440	26
Tyson	330	14
Swanson	300	12

Viewing, entering and editing data occurs in the Data editor. To open the Data editor, either choose Data editor from View menu or click on the Data editor tab (Untitled1.syz) in the Viewspace.



Open the Variable Properties dialog box, either from the menu Data->Variable Properties or by right-clicking on first column.



- Type BRAND$ for the variable name. The dollar sign ($) at the end of the variable name indicates that the variable contains character information.

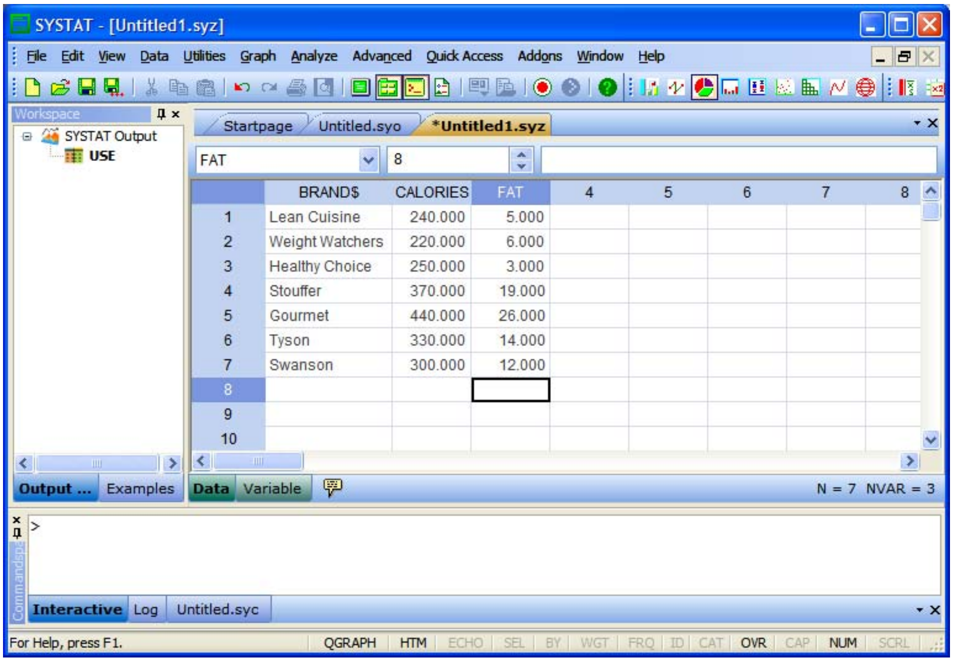
Note: Variable names cannot exceed 256 characters.

- In the Variable label edit box, you can type the alias for the variable name.
- Select String as the Variable type.
- Choose 15 from the drop-down list width edit box.
- Click OK to complete the variable definition.
- Repeat this process for the remaining variables, selecting Numeric as the variable type.

Note: In Numeric display options, the default decimal places are 3. This can be changed. Also, it is possible to change the display to Normal, Exponential notation or Date and time.

- Click the top left data cell (under the name of the first variable) and enter the data.
- To move across rows, press Tab after each entry. To move down columns, press the Enter key or down arrow key.

The Data editor will look like this:



- When you have finished entering the data, from the menus choose:

File
  Save As...

- Select the location for saving the file.
- Type SAMPLE as the name for the data file. SYSTAT adds the suffix *.SYZ* (*SAMPLE.SYZ*).

Reading an ASCII Text File

This section shows you how SYSTAT reads raw (ASCII) data files created in a text editor or word processor. SYSTAT can import ASCII files of the type .txt,.dat, and .csv.

SYSTAT can read alphanumeric characters, delimiters (spaces, commas, or tabs that separate consecutive values from each other), and carriage returns. SYSTAT cannnot read an ASCII file, which contains any unusual ASCII characters or page breaks, control characters, column markers, or similar formatting codes. See your word processor's documentation to find out how to save data as an ASCII text file.

Make sure that your text file satisfies the following criteria:

- Each case begins on a new line (to read ASCII files with two or more lines of data per case, use BASIC commands).
- Missing data are flagged with an appropriate code.

Imagine that someone used a text editor to enter 10 pieces of information (variables) about 28 frozen dinners:

BRAND$	Short names for brands
FOOD$	Words to identify each dinner as *chicken*, *pasta*, or *beef*
CALORIES	Calories per serving
FAT	Total fat in grams
PROTEIN	Protein in grams
VITAMIN A	Vitamin A, percentage daily value
CALCIUM	Calcium, percentage daily value
IRON	Iron, percentage daily value
COST	Price per dinner in U.S. dollars
DIET$	*Yes*, the dinner was shelved with dinners touted as “diet” or low in calories; *No*, it was shelved with regular dinners

BRAND$	FOOD$	CALORIES	FAT	PROTEIN	VITAMINA	CALCIUM	IRON	COST	DIET
lc	chicken	270	6	22	6	10	6	2.99	yes
lc	chicken	240	5	19	30	10	10	2.99	yes
lc	chicken	240	5	18	4	10	8	2.99	yes
lc	pasta	260	8	15	20	30	8	2.15	yes
lc	pasta	210	4	9	30	10	8	2.15	yes
ww	chicken	260	4	21	30	4	15	2.79	yes
ww	pasta	220	4	14	15	8	15	2.79	yes

BRAND$	FOOD$	CALORIES	FAT	PROTEIN	VITAMINA	CALCIUM	IRON	COST	DIET
ww	pasta	220	6	15	6	25	15	2.79	yes
hc	chicken	200	2	17	0	2	2	2.00	yes
hc	chicken	280	3	24	15	4	15	2.00	yes
ww	chicken	160	1	13	30	2	2	2.49	yes
hc	pasta	250	3	20	0	8	8	2.00	yes
ww	chicken	190	0	12	10	4	4	2.49	yes
st	beef	390	24	20	2	4	15	2.99	no
st	beef	370	19	24	2	20	15	2.99	no
st	chicken	320	10	27	10	15	8	2.69	no
st	chicken	330	16	18	2	2	4	2.99	no
gor	beef	290	8	18	15	4	10	1.75	no
gor	pasta	370	16	20	30	40	4	1.99	no
gor	pasta	440	26	20	100	35	10	1.75	no
gor	beef	300	34	22	15	10	20	1.75	no
ty	beef	330	14	24	8	10	10	3.00	no
ty	chicken	400	8	27	25	0	10	3.50	no
ty	chicken	340	7	31	70	0	15	3.50	no
ty	chicken	430	24	20	45	4	6	3.00	no
sw	chicken	550	25	22	0	6	15	2.25	no
sw	beef	330	9	25	10	2	25	2.85	no
sw	pasta	300	12	14	0	25	10	1.60	no

The first line contains names for the columns. SYSTAT will count these names (finding 10), and read 10 values for each case (dinner). We name this ASCII file *FOOD.DAT*. Let us read the *FOOD.DAT* file and convert it to a SYSTAT file called *FOOD.SYZ*.

■ From the menus choose:

File
  Open
    Data...

In the Open dialog box, select All Files from the drop-down list of file types, select *FOOD.DAT* and click Open.

The contents of the data file are displayed in the Data editor.

■ From the menus choose:

File
  Save As...

- Type FOOD for the filename in the Save dialog box and click OK.

The subsequent sections will show you how to create charts and run statistical analysis using SYSTAT menus and dialog boxes.

Graphics

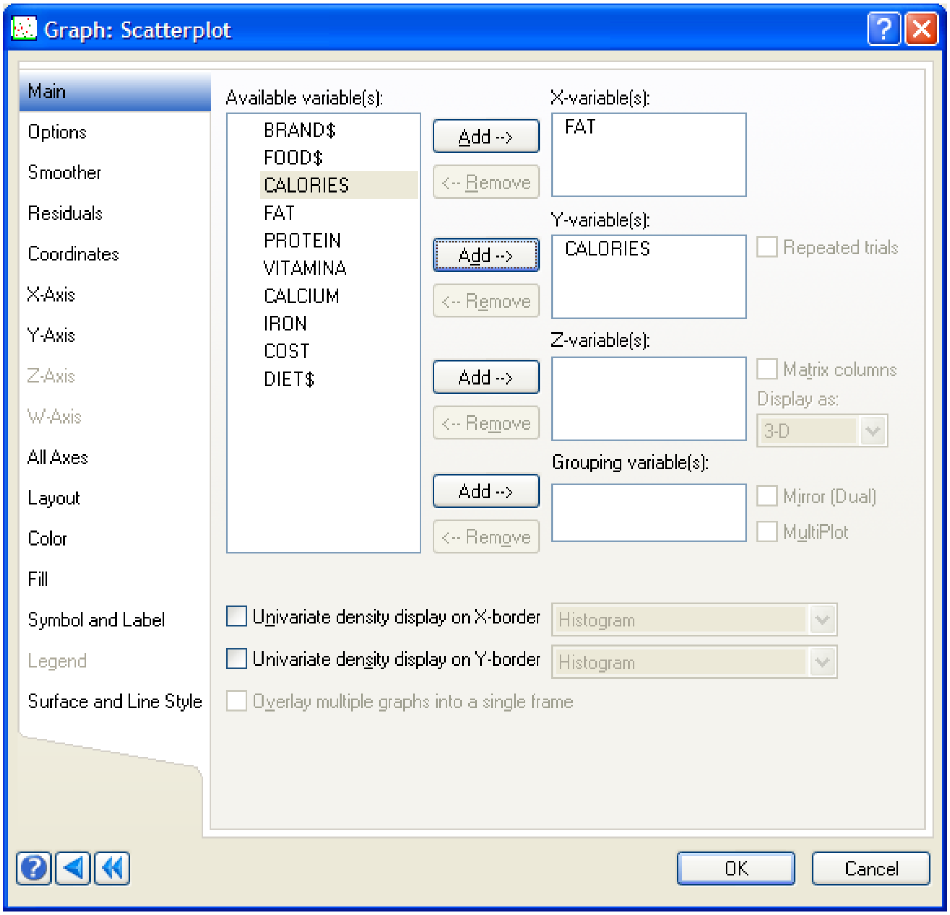
Scatterplots

Scatterplots provide a visual impression of the relation between two quantitative variables. Let us plot *CALORIES* versus *FAT* for this larger sample.

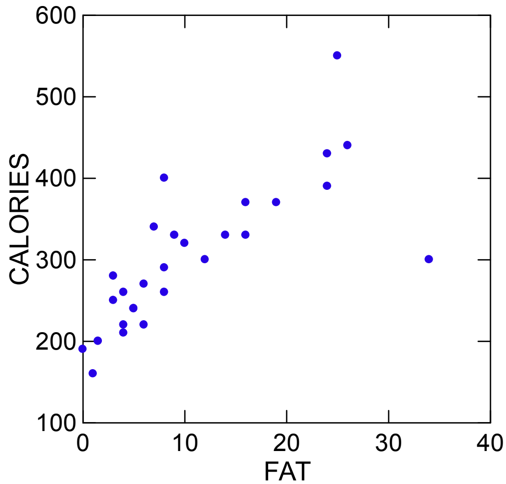
- From the menus choose:

Graph
  Scatterplot...

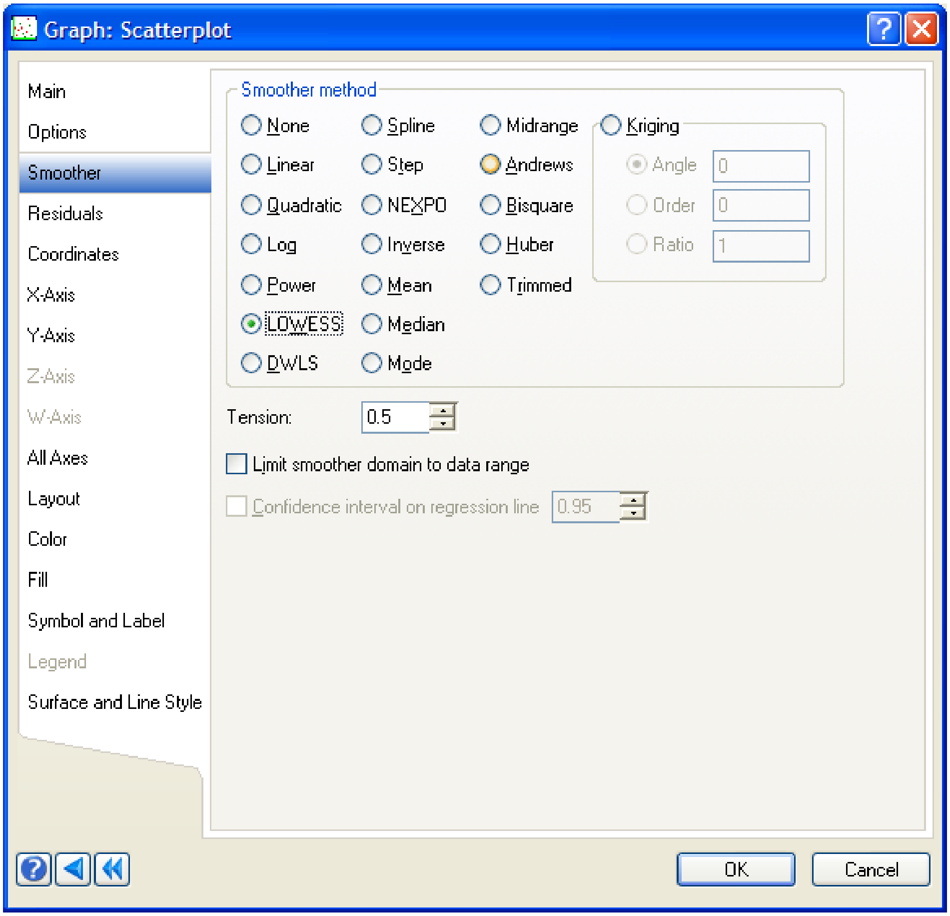
- In the Scatterplot dialog box, select *FAT* as the X-variable and *CALORIES* as the Y-variable.
- Click the Fill tab in the Scatterplot dialog box and select a solid fill for the first fill pattern.



- Click OK to execute the program.

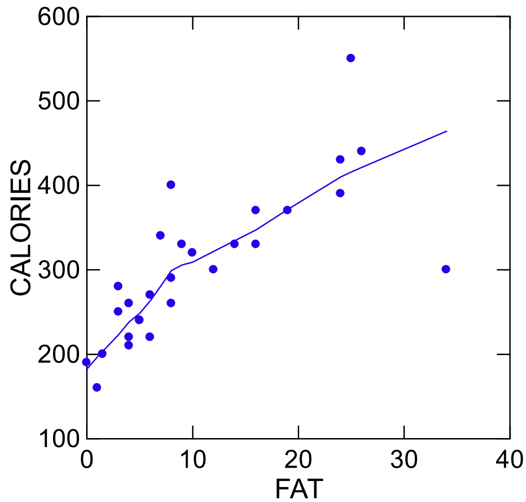


- Return to the Scatterplot dialog box by clicking the Scatterplot tool (). Notice that the previous settings are preserved.
- Click the Smoother tab in the Scatterplot dialog box, and select LOWESS smoother.



- Click OK to execute the program.

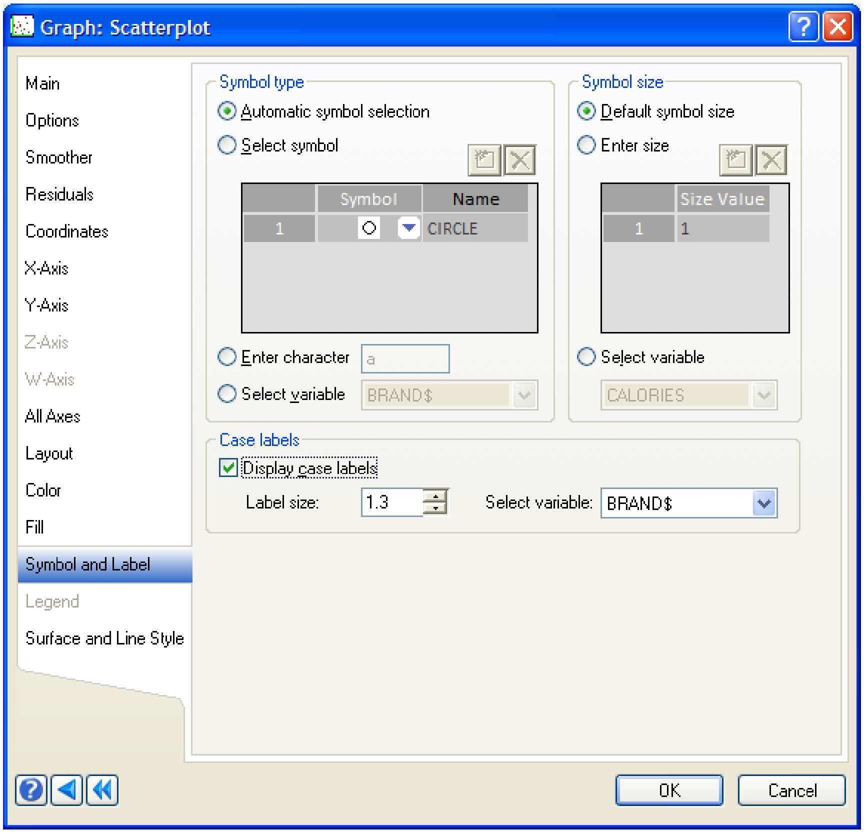
The resulting line displays a “typical” calorie value for each value of *FAT* without fitting a mathematical equation to the complete sample.



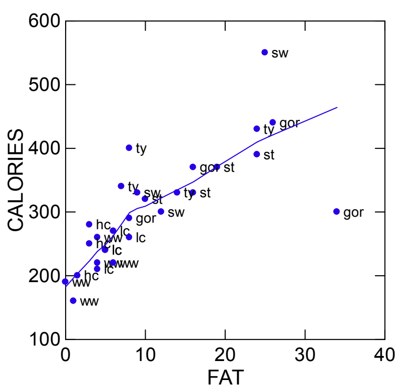
The smoother indicates, not surprisingly, that foods with a higher fat content tend to have more calories.

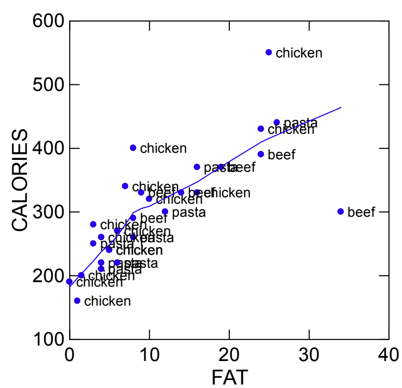
You may wonder what foods and what brands have the most calories? The fewest calories? The highest fat content? The lowest fat content?

- Return to the Scatterplot dialog box.
- Click the Symbol and Label tab in the Scatterplot dialog box, click Display case labels in the Case labels group, select *BRAND$* to label each plot point with the brand of the dinner, and set the case label size to 1.3. Repeat these steps for *FOOD$*.



■ Click OK to execute the program.

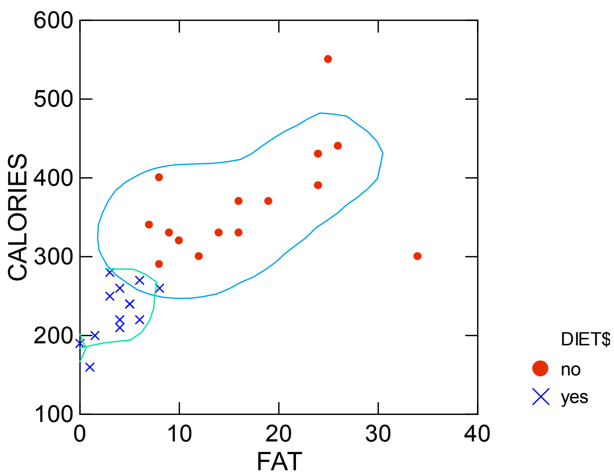




The top point in each plot is a chicken dinner made by *sw*—it must be fried chicken. Notice that the *beef* dinner by *gor* at the far right (close to the 300 calorie mark) contains considerably more fat than other dinners in the same calorie range.

Do diet dinners really have fewer calories and less fat than regular dinners? The dinners in the sample were selected from shelves where both regular and diet dinners were featured (*DIET$ no* and *yes*, respectively).

- Return to the Scatterplot dialog box.
- Select *DIET$* as the grouping variable.
- Select Overlay multiple graphs into a single frame.
- Deselect Display case labels in the Symbol and Label tab, and select None as the Smoother method in the Smoother tab.
- Click the Options tab in the Scatterplot dialog box.
- Select Confidence kernel and enter a p-value of 0.75 for a 75% confidence region.
- Click OK.

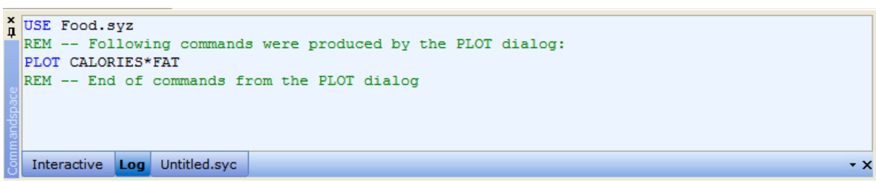


It is clear from the sample that the *DIET$ yes* dinners have fewer calories and less fat than the regular dinners.

Using Commandspace

Each time you use a dialog box to perform a step in an analysis, a command is generated. These "commands" are SYSTAT's instructions to perform the analysis. Instead of using dialog boxes to generate these commands, you can use the Commandspace and type them yourself. Whether generated by the dialog box or typed manually, the commands from each SYSTAT session can be saved in a file, modified, and resubmitted later. Although many users will use dialog boxes exclusively, we introduce commands here briefly to show how commands succinctly document the steps in your analysis. If you do not expect to use commands, you should skip the sections showing them.

You can type commands in the Commandspace of the SYSTAT window at the prompt (>) on the Interactive tab. When the Log tab is selected in the Commandspace, the commands corresponding to your dialog box choices are also displayed in the Commandspace. For example, the following command was generated by the Scatterplot dialog box selections.



If you enter commands from Interactive tab, you can recall previous commands by up and down arrow keys or by using F9 key.

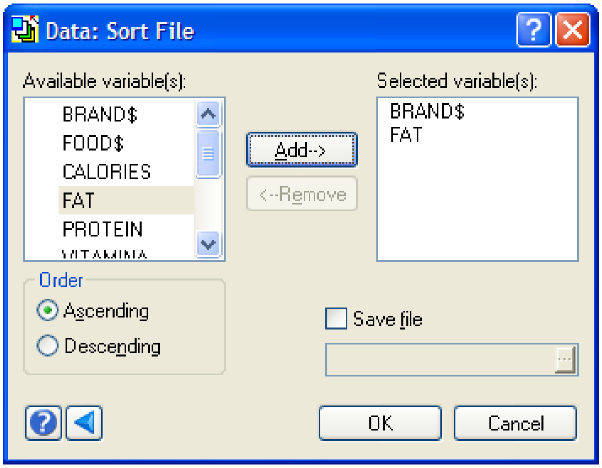
Sorting and Listing the Cases

Detailed graphics and statistics may not always be what you need—sometimes you can learn a lot simply by looking at numbers. This section shows you how to sort the dinners by type of food (*FOOD$*), and, within the foods, by fat content.

- From the menus choose:

Data
  Sort File...

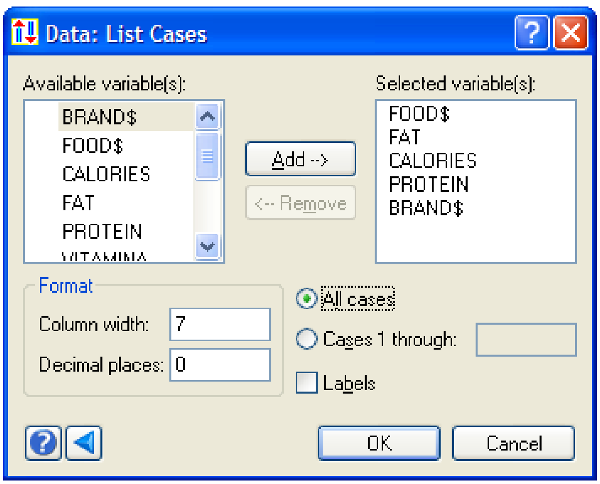
- In the Sort dialog box, select *FOOD$* and *FAT* as the variables, and then click OK.



- From the menus choose:

Data
  List Cases...

- Select *FOOD$*, *FAT*, *CALORIES*, *PROTEIN*, and *BRAND$* as the variables.
- In the Format group, enter 7 for Column width and 0 for Decimal places.



- Click OK.

Case	FOOD$	FAT	CALORIES	PROTEIN	BRAND$
1	beef	8	290	18	gor
2	beef	9	330	25	sw
3	beef	14	330	24	ty
4	beef	19	370	24	st
5	beef	24	390	20	st
6	beef	34	300	22	gor
7	chicken	0	190	12	ww
8	chicken	1	160	13	ww
9	chicken	2	200	17	hc
10	chicken	3	280	24	hc
11	chicken	4	260	21	ww
12	chicken	5	240	19	lc
13	chicken	5	240	18	lc
14	chicken	6	270	22	lc
15	chicken	7	340	31	ty
16	chicken	8	400	27	ty
17	chicken	10	320	27	st
18	chicken	16	330	18	st
19	chicken	24	430	20	ty
20	chicken	25	550	22	sw
21	pasta	3	250	20	hc
22	pasta	4	210	9	lc
23	pasta	4	220	14	ww
24	pasta	6	220	15	ww
25	pasta	8	260	15	lc
26	pasta	12	300	14	sw
27	pasta	16	370	20	gor
28	pasta	26	440	20	gor

Within each type of food, the fat content varies markedly. The diet brands *ww*, *lc*, and *hc* are the first entries under *chicken* and *pasta*. If the data file were larger, you would have to scan pages and pages of listings and it would be hard to see relationships (see the descriptors in the next section). Note that you can sort and list data in any procedure.

A Quick Description

As an early step in data screening, it is useful to summarize the values of grouping variables and to scan summary descriptors of quantitative variables.

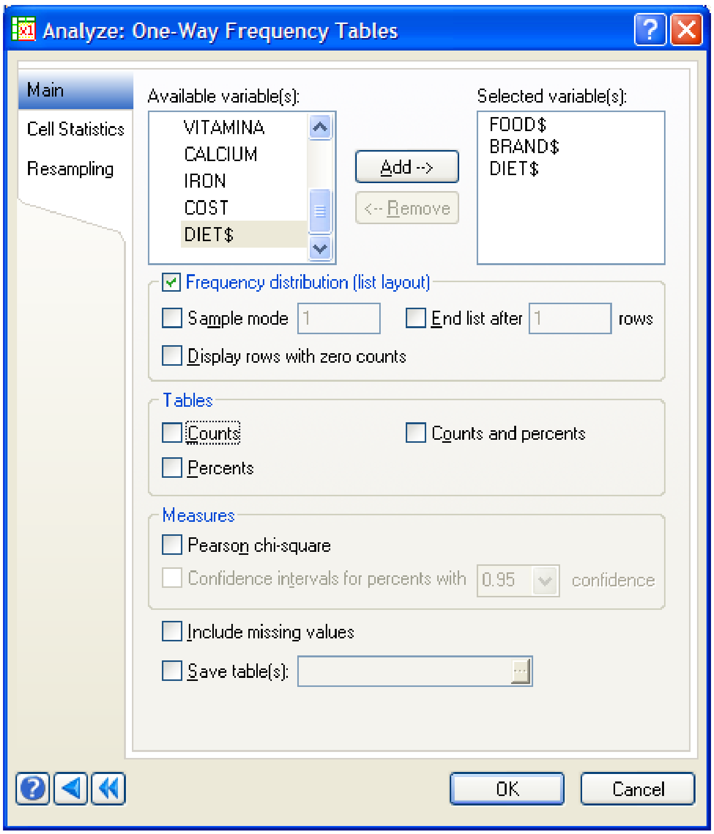
Frequency Counts and Percentages

The One-Way Frequency Tables on the Analyze menu, features many Print options that allow you to customize exactly what reports appear in your output. For example, the Frequency distribution option reports the number of times (*frequency*) each category of a grouping variable occurs and expresses it as a percentage of the total sample size. Cumulative frequencies and percentages are also available. In our "grabbing" sample strategy, we are interested in knowing what foods and how many of each brand and diet type we have.

- From the menus choose:

Analyze
  One-Way Frequency Tables...

- In the Tables group of the One-Way Tables dialog box, select Frequency distribution.
- Select *FOOD$, BRAND$,* and *DIET$* as the variables.



- Click OK

Frequency Distribution for FOOD$

FOOD$	Frequency	Cumulative Frequency	Percent	Cumulative Percent
beef	6	6	21.429	21.429
chicken	14	20	50.000	71.429
pasta	8	28	28.571	100.000

Frequency Distribution for BRAND$

BRAND$	Frequency	Cumulative Frequency	Percent	Cumulative Percent
gor	4	4	14.286	14.286
hc	3	7	10.714	25.000
lc	5	12	17.857	42.857
st	4	16	14.286	57.143
sw	3	19	10.714	67.857
ty	4	23	14.286	82.143
ww	5	28	17.857	100.000

Frequency Distribution for DIET$

DIET$	Frequency	Cumulative Frequency	Percent	Cumulative Percent
no	15	15	53.571	53.571
yes	13	28	46.429	100.000

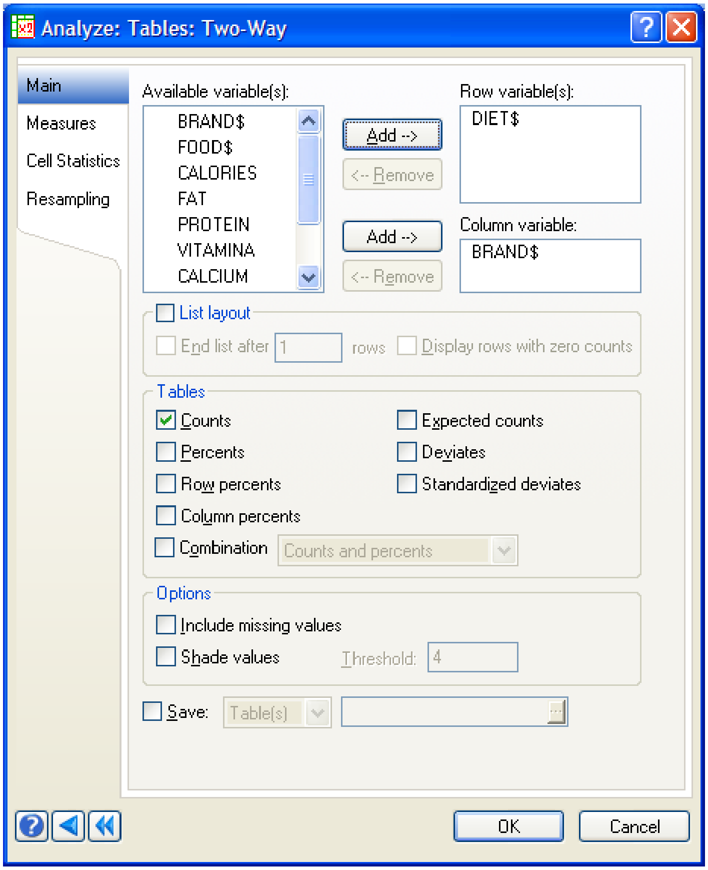
In above output, for *FOOD$* (the name appears at the top left in the first table), 14 of the 28 dinners in the sample (50% in the Pct column) are *chicken*, 28.6% are *pasta*, and 21.4% are *beef*. The number of dinners per *BRAND$* (second table) ranges from three to five. There are 15 regular (*DIET$ no*) dinners and 13 diet (*DIET$ yes*) dinners.

The List layout option in Two-Way Tables in the Analyze menu is useful for summarizing counts that result from cross-classifying two factors. Let us look at combinations of *DIET$* and *BRAND$*.

- From the menus choose:

Analyze
  Tables
    Two-Way...

- In the Options group of the Two-Way Tables dialog box, select List layout and deselect Counts.
- Select *DIET$* as the row variable and *BRAND$* as the column variable.



■ Click OK.

Frequency Distribution for DIET$ (rows) by BRAND$ (columns)

DIET$	BRAND$	Frequency	Cumulative Frequency	Percent	Cumulative Percent
no	gor	4	4	14.286	14.286
no	st	4	8	14.286	28.571
no	sw	3	11	10.714	39.286
no	ty	4	15	14.286	53.571
yes	hc	3	18	10.714	64.286
yes	lc	5	23	17.857	82.143
yes	ww	5	28	17.857	100.000

There are two *DIET$* and seven *BRAND$* categories—so there should be 14 combinations, but only 7 are shown here. The brands for the diet dinners differ from those for the regular dinners.

You may want to display frequencies for two factors as a two-way table. Let us deselect the List layout feature and look at *DIET$* by *FOOD$*.

- From the menus choose:

Analyze
 Tables
  Two-Way...

- Select *DIET$* as the row variable and *FOOD$* as the column variable.
- Deselect List layout (click the check box to deselect it if it is currently selected) and select Frequencies from the table box.

Counts

DIET$(rows) byFOOD$(columns)

	beef	chicken	pasta	Total
no	6	6	3	15
yes	0	8	5	13
Total	6	14	8	28

We failed to get any beef dinners in the *DIET$ yes* group.

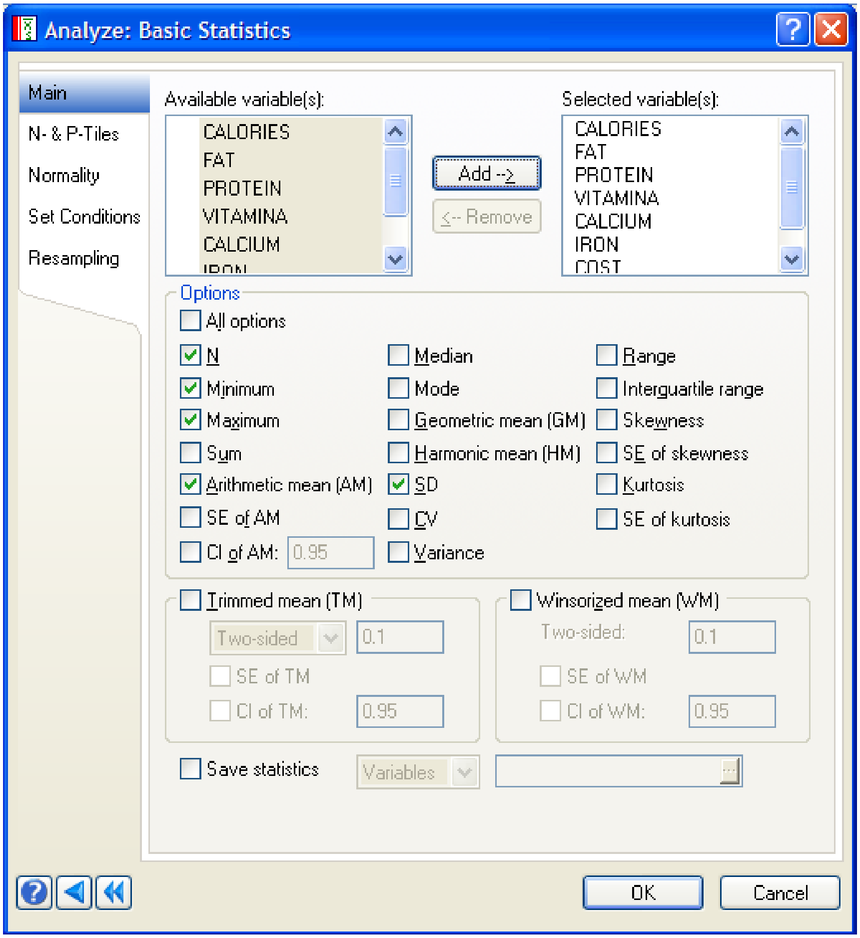
Descriptive Statistics

It is easy to request a panel of descriptive statistics. However, since we have not examined several of these distributions graphically, we should avoid reporting means and standard deviations (these statistics can be misleading when the shape of the distribution is highly skewed). It is helpful to scan the sample size for each variable to determine whether values are missing. The basic statistics are number of observations (N), minimum, maximum, arithmetic mean (AM), geometric mean, harmonic mean, sum, standard deviation, variance, coefficient of variation (CV), range, median, standard error of AM, etc.

- From the menus choose:

Analyze
 Basic Statistics...

- In the Analyze: Basic Statistics dialog box, select all of the variables in the source list (only numeric variables are available for this feature), and click OK to calculate the default statistics.



	CALORIES	FAT	PROTEIN	VITAMINA	CALCIUM
N of Cases	28.000	28.000	28.000	28.000	28.000
Minimum	160.000	0.000	9.000	0.000	0.000
Maximum	550.000	34.000	31.000	100.000	40.000
Arithmetic Mean	303.214	10.804	19.679	18.929	10.857
Standard Deviation	87.815	8.959	5.019	22.593	10.845

	IRON	COST
N of Cases	28.000	28.000
Minimum	2.000	1.600
Maximum	25.000	3.500
Arithmetic Mean	10.464	2.544
Standard Deviation	5.467	0.548

For each variable, SYSTAT gives the number of cases with nonmissing values, the largest and smallest values, and the mean and standard deviation. *CALORIES* for a single dinner range from 160 to 550 with an average around 300 (303.214 to be exact). *VITAMINA* ranges from 0% to 100% with a mean of 18.9%. Since the mean is not close to the middle of the range, the distribution must be quite skewed or have a few extreme values.

Statistics By Group

You can use By Groups on the Data menu to stratify the analysis.

- From the menus choose:

Data
  By Groups...

- In the By Groups dialog box, select *DIET$* as the variable, and click OK.
- Return to the Basic Statistics dialog box.
- Select the following measures: N, Minimum, Maximum, Arithmetic mean (AM), CI of AM, and Median.
- Click OK.

```
Results for DIET$ = yes

                               ¦ CALORIES      FAT   PROTEIN   VITAMINA   CALCIUM
-------------------------------+--------------------------------------------------
 N of Cases                    ¦   13.000   13.000    13.000     13.000    13.000
 Minimum                       ¦  160.000    0.000     9.000      0.000     2.000
 Maximum                       ¦  280.000    8.000    24.000     30.000    30.000
 Median                        ¦  240.000    4.000    17.000     15.000     8.000
 Arithmetic Mean               ¦  230.769    3.885    16.846     15.077     9.769
 95.0% Lower Confidence Limit  ¦  209.769    2.544    14.225      7.921     4.629
 95.0% Upper Confidence Limit  ¦  251.770    5.225    19.467     22.233    14.910

                               ¦   IRON     COST
-------------------------------+----------------
 N of Cases                    ¦ 13.000   13.000
 Minimum                       ¦  2.000    2.000
 Maximum                       ¦ 15.000    2.990
 Median                        ¦  8.000    2.490
 Arithmetic Mean               ¦  8.923    2.509
 95.0% Lower Confidence Limit  ¦  5.999    2.265
 95.0% Upper Confidence Limit  ¦ 11.847    2.754

Results for DIET$ = no

                               ¦ CALORIES      FAT   PROTEIN   VITAMINA   CALCIUM
-------------------------------+--------------------------------------------------
 N of Cases                    ¦   15.000   15.000    15.000     15.000    15.000
 Minimum                       ¦  290.000    7.000    14.000      0.000     0.000
 Maximum                       ¦  550.000   34.000    31.000    100.000    40.000
 Median                        ¦  340.000   16.000    22.000     10.000     6.000
 Arithmetic Mean               ¦  366.000   16.800    22.133     22.267    11.800
 95.0% Lower Confidence Limit  ¦  327.873   12.247    19.748      6.231     4.735
 95.0% Upper Confidence Limit  ¦  404.127   21.353    24.519     38.302    18.865
```

	IRON	COST
N of Cases	15.000	15.000
Minimum	4.000	1.600
Maximum	25.000	3.500
Median	10.000	2.850
Arithmetic Mean	11.800	2.573
95.0% Lower Confidence Limit	8.597	2.207
95.0% Upper Confidence Limit	15.003	2.939

The median grams of protein for the 13 diet dinners is 17; the mean is 16.8. For the 15 regular dinners, these statistics are 22 and 22.1, respectively. Later we will request a two-sample *t* test to see if this is a significant difference. A 95% confidence interval for the average cost of a diet dinner ranges from $2.27 to $2.75. The confidence interval for the average cost of the regular dinners is larger—$2.21 to $2.94.

The *BY GROUPS* variable, *DIET$*, remains in effect for subsequent graphical displays and statistical analyses. To disengage it, return to the By Groups dialog box and select Turn off.

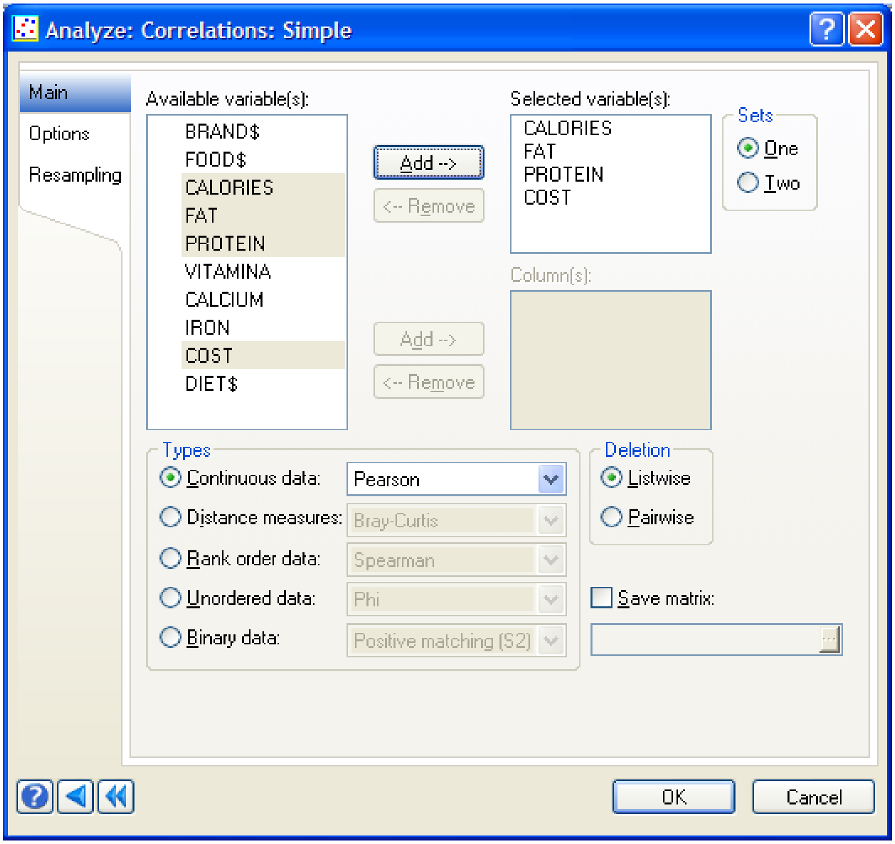
A First Look at Relations among Variables

What are the correlations among calories, fat content, protein, and cost? We can use correlations to quantify the linear relations among these variables.

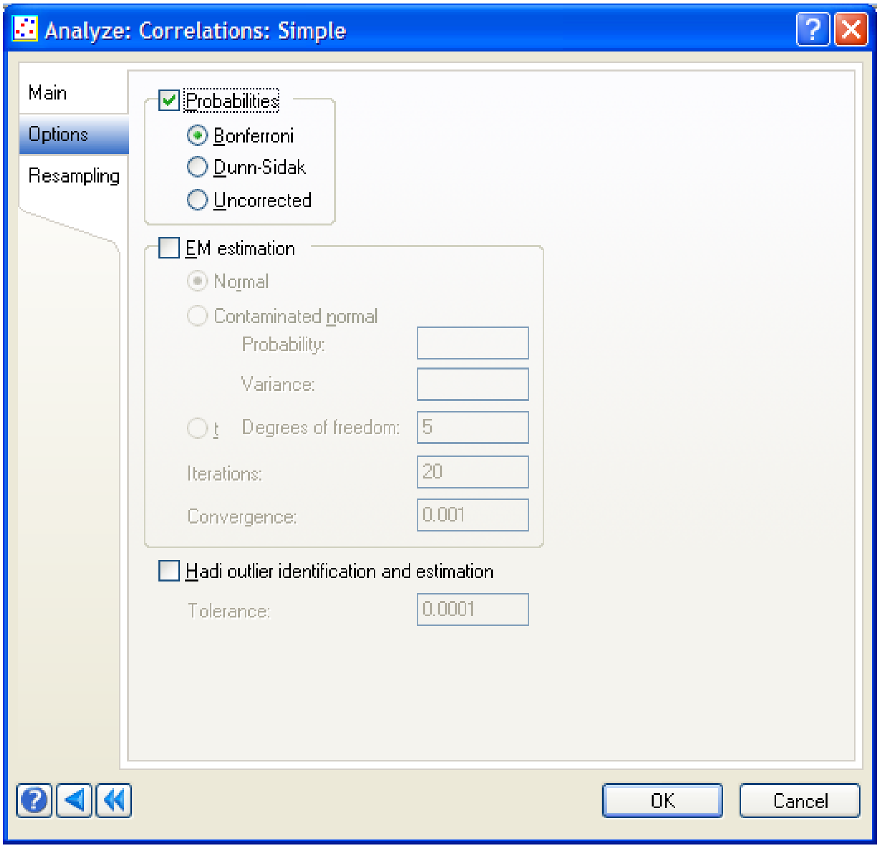
- From the menus choose:

Analyze
  Correlations
    Simple..

- In the Simple Correlations dialog box, select Continuous data type and select Pearson from the Continuous data drop-down list.
- Select *CALORIES*, *FAT*, *PROTEIN*, and *COST* as the variables.



- Click the Options tab and select Probabilities and Bonferroni. Because we study six correlations among four variables, we use Bonferroni adjusted probabilities to provide protection for multiple tests.



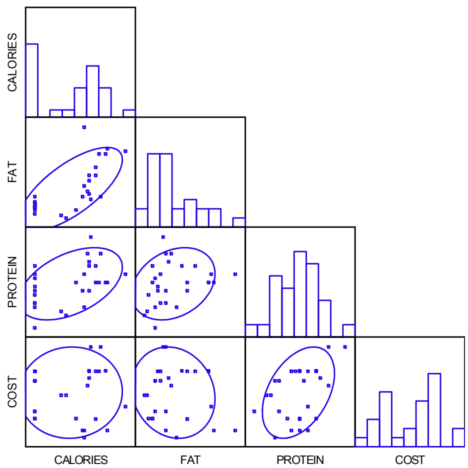
- Click OK.

```
Number of Observations: 28

Means

CALORIES       FAT    PROTEIN     COST
------------------------------------
 303.214    10.804     19.679    2.544

Pearson Correlation Matrix

         ¦ CALORIES       FAT    PROTEIN     COST
---------+-----------------------------------
CALORIES ¦    1.000
FAT      ¦    0.757 1.000
PROTEIN  ¦    0.550     0.278 1.000
COST     ¦    0.099    -0.134 0.420    1.000

Bartlett Chi-square Statistic : 38.865
df                            :      6
p-value                       :  0.000
```

Matrix of Bonferroni Probabilities

	CALORIES	FAT	PROTEIN	COST
CALORIES	0.000			
FAT	0.000	0.000		
PROTEIN	0.014	0.908	0.000	
COST	1.000	1.000	0.156	0.000

Scatter Plot Matrix



In above output one Quick Graph is generated. This is the Quick Graph that SYSTAT automatically generates when you request correlations. Quick Graphs are available for most statistical procedures. If you want to turn off a Quick Graph, use Options on the Edit menu.

The Quick Graph in this example is a scatterplot matrix (SPLOM). There is one bivariate scatterplot corresponding to each entry in the correlation matrix that follows. Univariate histograms for each variable are displayed along the diagonal, and 75% normal theory confidence ellipses are displayed within each plot.

The plot of *FAT* and *CALORIES* (top left) has the narrowest ellipse, and thus, the strongest correlation (that is, given that the configuration of the points is spread evenly, is not nonlinear, and has no anomalies).

In the Pearson correlation matrix displayed in above output, the correlation between *FAT* and *CALORIES* is 0.758. The *p*-value (or Bonferroni adjusted probability) associated with 0.758 is printed as 0.000 (or less than 0.0005). As the scatterplot seemed to indicate, the *FAT* and *CALORIES* Pearson correlation matrix is correlated.

PROTEIN also has a significant correlation with *CALORIES* (*r* = 0.55, *p-value* = 0.014). We are unable to detect significant correlations between *COST* and *CALORIES*, *FAT*, and *PROTEIN*.

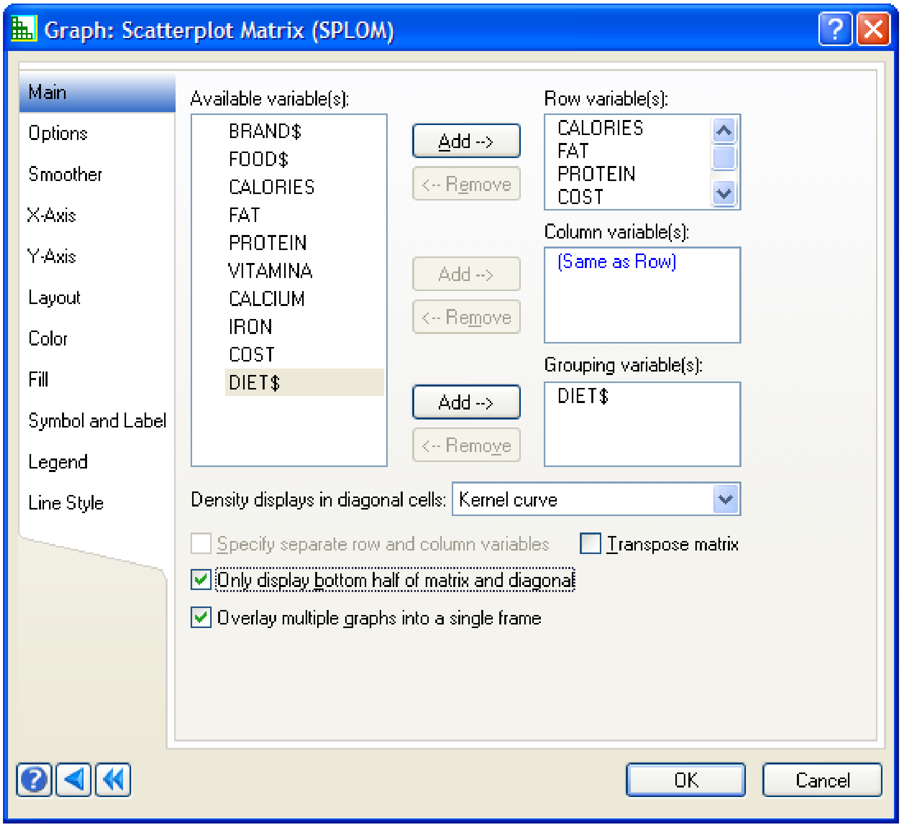
Subpopulations

The presence of subpopulations can mask or falsely enhance the size of a correlation. With Correlations, we could specify *DIET$* as a *BY GROUPS* variable as we did previously. Instead, let us examine the data graphically and use 75% nonparametric kernel density contours to identify the diet *yes* and *no* groups. We will also look at univariate kernel density curves for the groups.

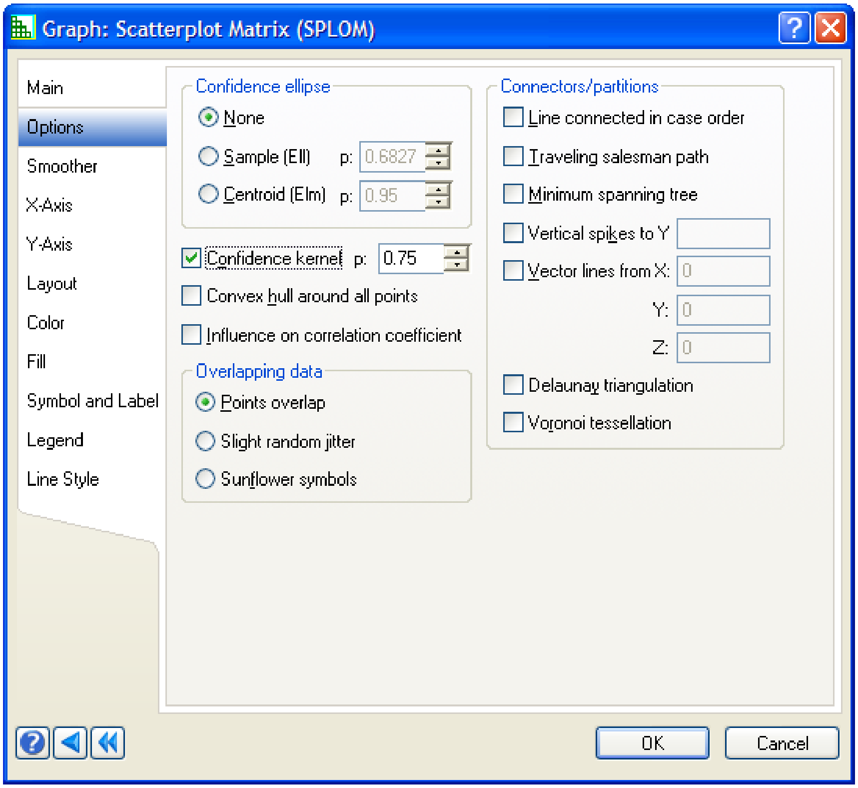
- From the menus choose:

Graph
  Scatterplot Matrix (SPLOM)...

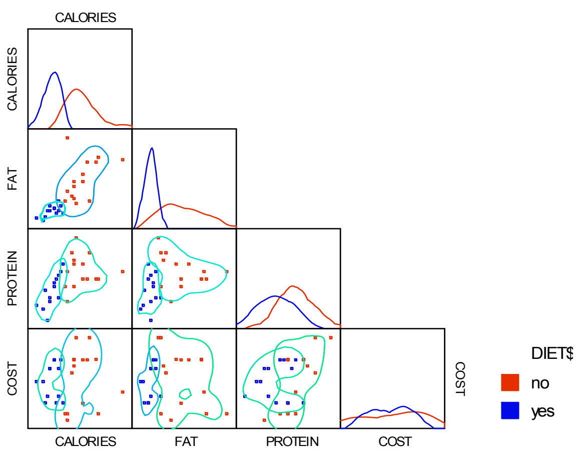
- Select *CALORIES*, *FAT*, *PROTEIN*, and *COST* as the Row variables.
- Select *DIET$* as the Grouping variable.
- Select Kernel Curve from the drop-down list for Density displays in diagonal cells.
- Select Only display bottom half of matrix and diagonal and Overlay multiple graphs into a single frame.



- Click the Options tab in the Scatterplot Matrix dialog box.
- Select Confidence kernel and enter the value of *p* as 0.75.

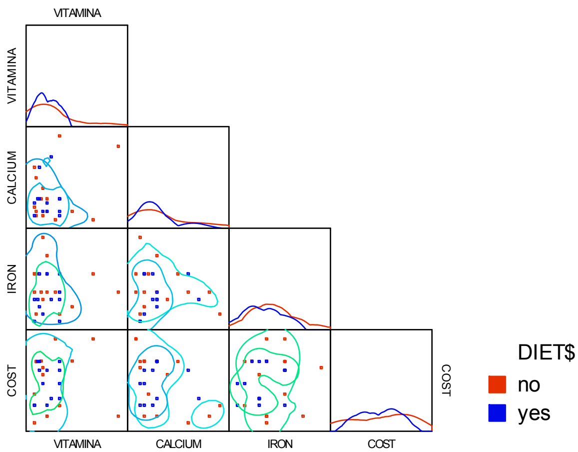

Graph: Scatterplot Matrix (SPLOM)
Main
Options
Smoother
X-Axis
Y-Axis
Layout
Color
Fill
Symbol and Label
Legend
Line Style
Confidence ellipse
None
Sample (Ell)
p:
0.6827
Centroid (Elm)
p:
0.95
Confidence kernel
p:
0.75
Convex hull around all points
Influence on correlation coefficient
Overlapping data
Points overlap
Slight random jitter
Sunflower symbols
Connectors/partitions
Line connected in case order
Traveling salesman path
Minimum spanning tree
Vertical spikes to Y
Vector lines from X:
0
Y:
0
Z:
0
Delaunay triangulation
Voronoi tessellation
OK
Cancel

- Click OK.


CALORIES
FAT
PROTEIN
COST
CALORIES
FAT
PROTEIN
COST
COST
DIET$
no
yes

For *CALORIES* and *FAT*, look at the separation of the univariate densities on the diagonal of the display. Notice that the price range (*COST*) at the bottom right for the diet dinners is within that for the regular dinners. *COST* is the Y-variable in the bottom row of plots. Within each group, *COST* appears to have little relation to *CALORIES* or *FAT*. It is possible that *COST* has a positive association with *PROTEIN* for the regular dinners (open circles in the *COST* versus *PROTEIN* plot).

Is there a relationship between cost and nutritive value as measured by the percentage daily value for vitamin A, calcium, and iron? Repeat the steps for the previous plot, but select *VITAMINA, CALCIUM, IRON,* and *COST* as the row variables.



COST is the Y-variable for each plot on the bottom row. There is no strong relationship between cost and nutritive value (as measured by *VITAMINA*, *CALCIUM*, and *IRON*), but there is a small cluster of low-cost dinners with high-calcium content. Later, we will find that these are *pasta* dinners.

3-D Displays

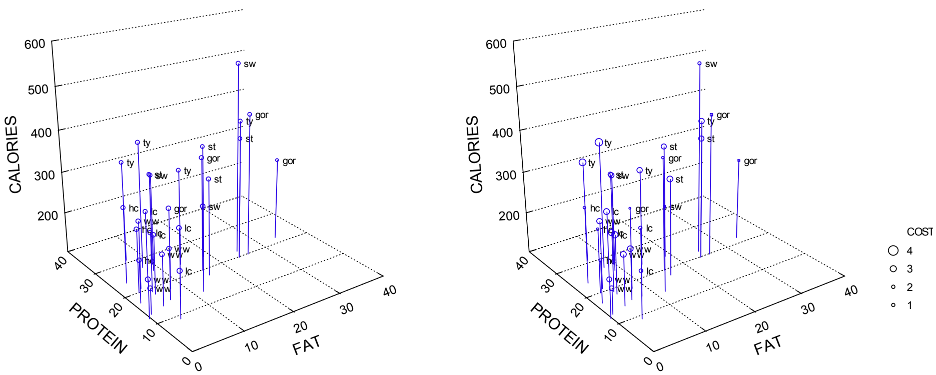
In this section, we use 3-D displays for another look at calories, protein, and fat. In the display on the left, we label each dinner with its brand code; in the display on the right, we use the cost of the dinner to determine the size of the plot symbol.

To produce 3-D displays:

- From the menus choose:

Graph
  Scatterplot...

- In the Scatterplot dialog box, select *FAT* as the X-variable, *PROTEIN* as the Y-variable, and *CALORIES* as the Z -variable.
- Select Display grid lines in the X-Axis, Y-Axis, and Z-Axis tabs.
- Click the Options tab and select Vertical spikes to Y from the Connectors/partitions group.
- To produce the plot on the left, click the Symbol and Label tab, click Display case labels in the Case labels group, and select *BRAND$* to label each plot point with the brand of the dinner.
- To produce the plot on the right, click the Symbol and Label tab, click Select variable in the Symbol size group, and select *COST* as the symbol size variable.



Notice the back corner of the display on the left—the tallest spike extends to *sw*, indicating the dinner with the most calories. On the floor of the display, we read that its fat content is between 20 and 30 grams and that its protein is a little over 20 grams. We see this same point in the display on the right—the size of its circle is not extreme, indicating a mid-range price. Notice the small circle toward the far right—this dinner costs much less than the *sw* dinner and has a higher fat content and a similar protein value. The most expensive dinners (that is, the larger circles) do not concentrate in a particular region.

A Two-Sample t-Test

One of the most common situations in statistical practice involves comparing the means for two groups. For example, does the average response for the treatment group differ from that for the control group? Ideally, the subjects should be randomly assigned to the groups.

For the food data, we are interested in possible differences in *PROTEIN* and *CALCIUM* between the diet and regular dinners. Thus, the dinners are not randomly assigned to groups. In a real observational study, a researcher should carefully explore the data to ensure that other factors are not masking or enhancing a difference in means.

In the t-test, we test the hypothesis,

H_0: Means of diet and regular dinners are equal.

The alternative to this hypothesis could be

H_1: Mean of Diet is "greater" than mean of regular, or

H_1: Mean of Diet is "not equal" to mean of regular, or

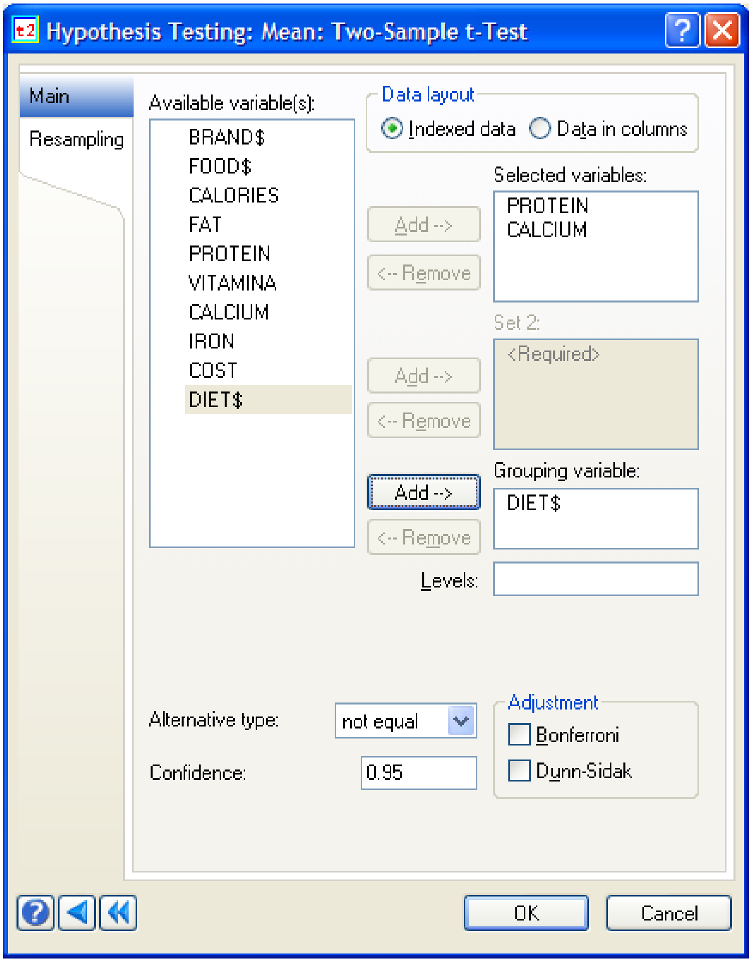
H_1: Mean of Diet is "less" than mean of regular.

Since we have no information, let us choose the second alternative H_1: Mean of diet is “not equal” to mean of regular. In other words, do diet and regular dinners differ in protein and calcium content? In this example, we use the t-test procedure.

- From the menus choose:

Analyze
  Hypothesis Testing
    Mean
      Two Sample t-Test...

- In the Two-Sample t-Test dialog box, select *PROTEIN* and *CALCIUM* as the variables, and select *DIET$* as the grouping variable.
- In the Alternative type, choose ‘not equal’.
- Click OK.



H0: Mean1 = Mean2 vs. H1: Mean1 <> Mean2

Grouping Variable = DIET$

Variable	DIET$	N	Mean	Standard Deviation
PROTEIN	no	15.000	22.133	4.307
	yes	13.000	16.846	4.337
CALCIUM	no	15.000	11.800	12.757
	yes	13.000	9.769	8.506

Separate Variance

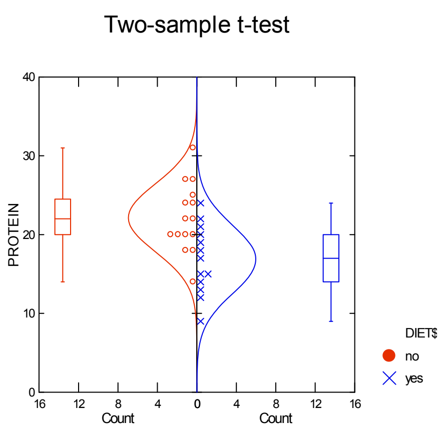
Variable	DIET$	Mean Difference	95.00% Confidence Interval Lower Limit	Upper Limit	t	df
PROTEIN	no	5.287	1.916	8.658	3.228	25.385
	yes					
CALCIUM	no	2.031	-6.322	10.384	0.501	24.520
	yes					

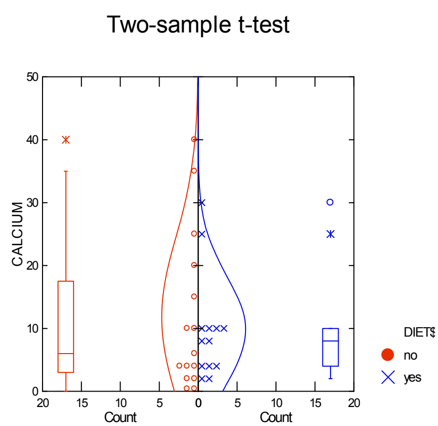
Variable	p-Value
PROTEIN	0.003
CALCIUM	0.621

Pooled Variance

Variable	DIET$	Mean Difference	95.00% Confidence Interval Lower Limit	Upper Limit	t	df
PROTEIN	no	5.287	1.922	8.653	3.229	26.000
	yes					
CALCIUM	no	2.031	-6.538	10.600	0.487	26.000
	yes					

Variable	p-Value
PROTEIN	0.003
CALCIUM	0.630





The t-test procedure produces two density plots as Quick Graphs. On the far left and right sides of the density plot for each test variable are box plots for each category of the grouping variable. The box plot on the left side of each graph is for the *DIET$ no* group, and the box plot on the right side of each graph is for the *DIET$ yes* group.

The middle portion of each graph shows the actual distribution of data points, with a normal curve for comparison.

The results in the box plots for *PROTEIN* are desirable. The median (horizontal line in each box) is in the center of the box, and the lengths of the boxes are similar. Also, the peaks of the normal curves, which represent the mean for a normal distribution, are very close to the median values. This indicates that the distributions are symmetric and have approximately the same spread (variance). This is not true for *CALCIUM*. These distributions are right skewed and possibly should be transformed before analysis.

The mean values for *PROTEIN* are the same as those in the By Groups statistics—22.133 and 16.846. The standard deviations differ little (4.307 and 4.337), confirming what we observed in the box plots. This means that we can use the results of the pooled-variance *t* test printed below the means. This test is usually the first one you see in introductory texts and assumes that the distributions have the same shape (that is, the variances do not differ). For *PROTEIN*, we conclude that the mean of 22.1 for the regular dinners does differ significantly from the mean of 16.8 for the diet dinners ($t = 3.229$, p-value = 0.0003).

The separate-variance *t* test does not require the assumption of equal variances. Considering the distributions for *CALCIUM* displayed in the box plots and that the standard deviations for the groups are 12.757 and 8.506, we use the separate-variance *t* test results. We are unable to report a difference in average *CALCIUM* values for the regular and diet dinners ($t = 0.501$, p-value = 0.621).

The discussion of SYSTAT's procedures is very exploratory at this stage, so you should not conclude that *CALCIUM* values are homogeneous. Always take the time to think about what possible subgroups might be influencing or obscuring results.

A One-Way Analysis of Variance (ANOVA)

Does the cost of a dinner vary by brand? Let us try an analysis of variance (ANOVA) to determine whether the average price of frozen dinners varies by brand. After looking at the graphics earlier in this chapter, we assume that differences do exist, so we also request the Tukey HSD test for post hoc comparison of means. This test provides protection for testing many pairs of means simultaneously; allowing us to make statements about which brand's average cost differs significantly from another brand's. Before we run the analysis of variance, we will specify how the brands should be ordered in the output (results will be easier to follow if we order the brands from least to most expensive).

- From the menus choose:

Data
  Order of Display...

- In the Order dialog box, select *BRAND$* as the variable.
- Select Enter sort and type 'gor', 'hc', 'sw', 'lc', 'ww', 'st', 'ty'.
- Click OK.
- From the menus choose:

Edit
  Options...

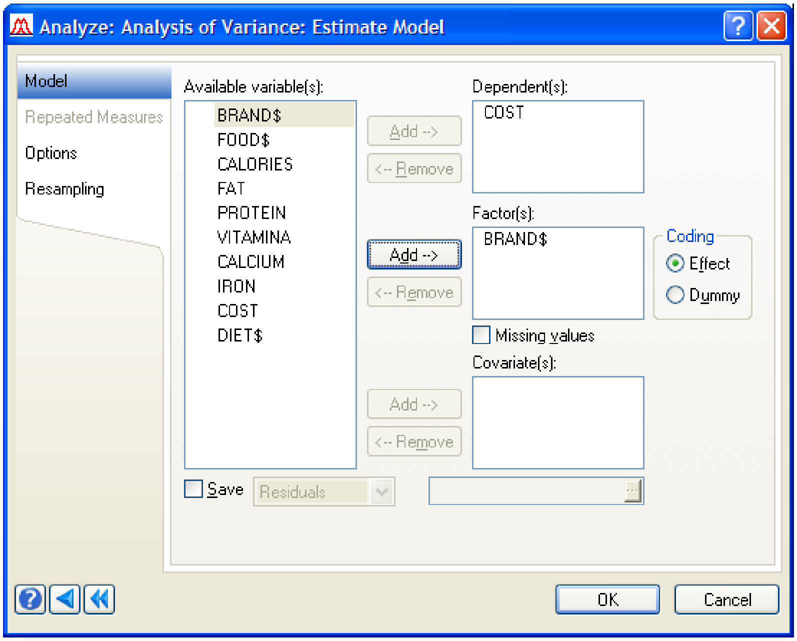
- In the Output Results group on the Output tab, select Long from the Length drop-down list. (This will provide extended results for the analysis of variance.)
- Click OK.

To request an analysis of variance:

- From the menus choose:

Analyze
  Analysis of Variance
    Estimate Model...

- In the Analysis of Variance: Estimate Model dialog box, select *COST* as the dependent variable and *BRAND$* as the factor variable.
- Click OK.




```
Effects coding used for categorical variables in model.
The categorical values encountered during processing are

Variables             ¦             Levels
----------------------+-------------------------
BRAND$  (7 levels)    ¦ gor    hc     sw     lc     ww
                      ¦ st     ty


Dependent Variable    ¦   COST
N                     ¦     28
Multiple R            ¦  0.861
Squared Multiple R    ¦  0.742


Estimates of Effects B = (X'X)-1X'Y

Factor     ¦ Level      COST
-----------+---------------
CONSTANT   ¦            2.505
BRAND$     ¦ gor       -0.695
BRAND$     ¦ hc        -0.505
BRAND$     ¦ sw        -0.271
BRAND$     ¦ lc         0.149
BRAND$     ¦ ww         0.165
BRAND$     ¦ st         0.410
```

Analysis of Variance

```
Source  ¦ Type III SS    df    Mean Squares    F-Ratio    p-Value
--------+-------------------------------------------------------
BRAND$  ¦       6.017     6           1.003     10.042      0.000
Error   ¦       2.097    21           0.100
```

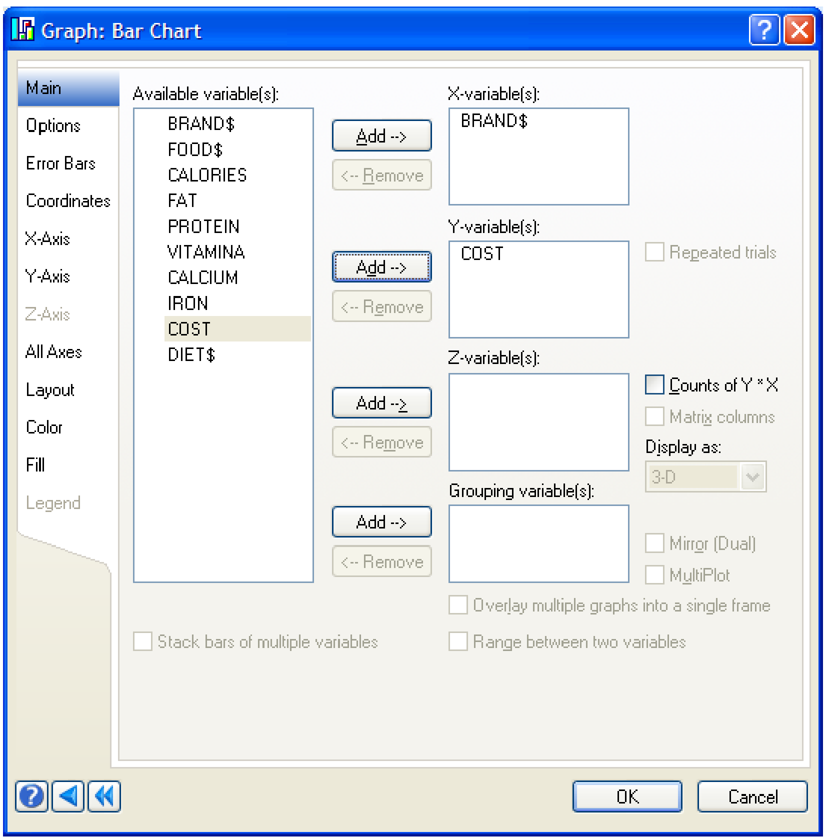
```
Least Squares Means

Factor ¦ Level   LS Mean   Standard Error       N
-------+---------------------------------------------
BRAND$ ¦ gor       1.810            0.158   4.000
BRAND$ ¦ hc        2.000            0.182   3.000
BRAND$ ¦ sw        2.233            0.182   3.000
BRAND$ ¦ lc        2.654            0.141   5.000
BRAND$ ¦ ww        2.670            0.141   5.000
BRAND$ ¦ st        2.915            0.158   4.000
BRAND$ ¦ ty        3.250            0.158   4.000
```

We can point out that the means are ordered by increasing cost because of the Order feature. This feature also pertains to graphical displays.

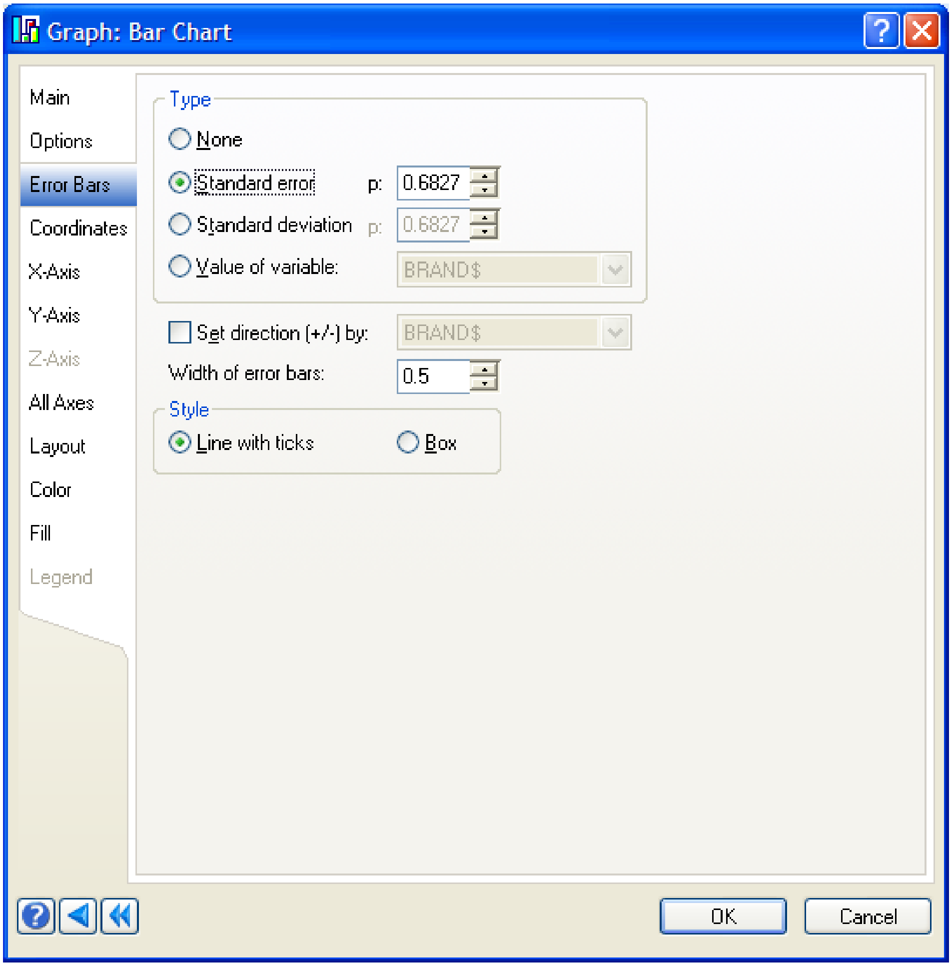
- From the menus choose:

Graph
  Bar Chart...

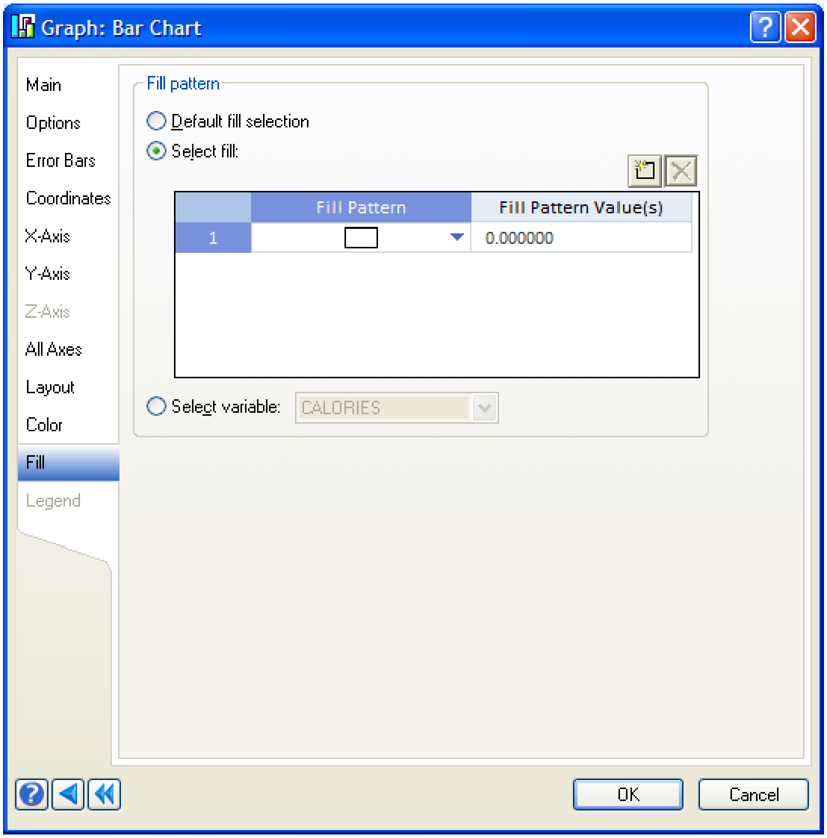
- Select *BRAND$* as the X-variable and *COST* as the Y-variable.



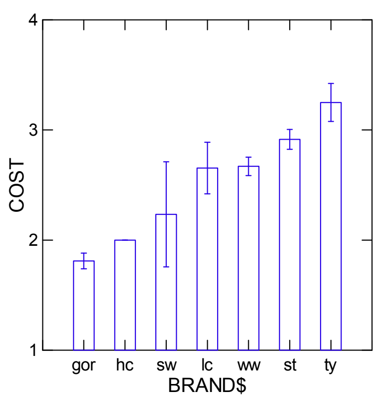
■ Click the Error Bars tab and select Standard error from the Type group.



■ Click the Fill tab, select Select fill from the Fill pattern group, and select as the Fill Pattern.


Graph: Bar Chart
Main
Options
Error Bars
Coordinates
X-Axis
Y-Axis
Z-Axis
All Axes
Layout
Color
Fill
Legend
Fill pattern
Default fill selection
Select fill:
Fill Pattern
Fill Pattern Value(s)
1
0.000000
Select variable:
CALORIES
OK
Cancel

- Click OK.


4
3
2
1
COST
gor
hc
sw
lc
ww
st
ty
BRAND$

The *F-ratio* in the Analysis of Variance table at the beginning of the output indicates that there are one or more differences in average price among the seven brands (*F-ratio* = 10.0415, *p-value* < 0.0005).

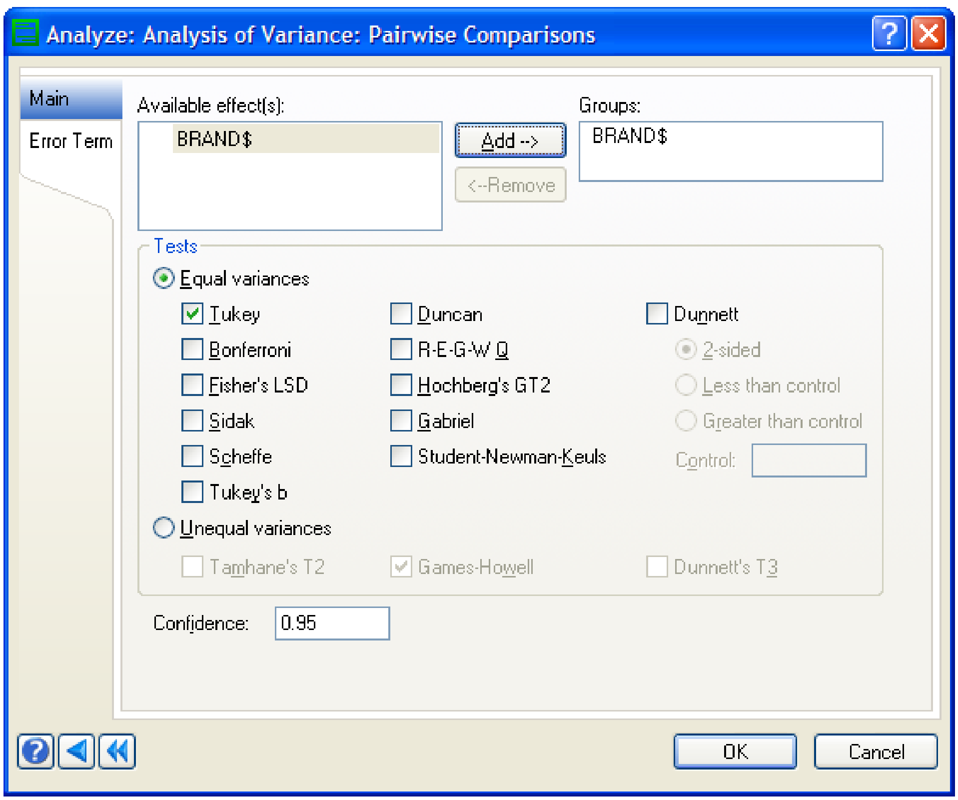
Tukey Pairwise Mean Comparisons

Let us use SYSTAT's advanced hypothesis testing capability to request Tukey's Pairwise Mean Comparison test.

From the menus choose:

Analyze
  Analysis of Variance
    Pairwise Comparisons...

- Specify *BRAND$* under Groups and select Tukey under Tests.



■ Click OK.

```
Post Hoc Test of COST
Using least squares means.
Using model MSE of 0.100 with 21 df.
```

Tukey's Honestly-Significant-Difference Test

BRAND$(i)	BRAND$(j)	Difference	p-Value	95% Confidence Interval Lower	Upper
gor	hc	-0.190	0.984	-0.975	0.595
gor	sw	-0.423	0.590	-1.208	0.361
gor	lc	-0.844	0.010	-1.533	-0.155
gor	ww	-0.860	0.009	-1.549	-0.171
gor	st	-1.105	0.001	-1.831	-0.379
gor	ty	-1.440	0.000	-2.166	-0.714
hc	sw	-0.233	0.968	-1.072	0.605
hc	lc	-0.654	0.115	-1.404	0.096
hc	ww	-0.670	0.100	-1.420	0.080
hc	st	-0.915	0.016	-1.700	-0.130
hc	ty	-1.250	0.001	-2.035	-0.465
sw	lc	-0.421	0.548	-1.171	0.330
sw	ww	-0.437	0.506	-1.187	0.314
sw	st	-0.682	0.117	-1.466	0.103
sw	ty	-1.017	0.006	-1.801	-0.232
lc	ww	-0.016	1.000	-0.666	0.634
lc	st	-0.261	0.874	-0.950	0.428
lc	ty	-0.596	0.120	-1.285	0.093
ww	st	-0.245	0.903	-0.934	0.444
ww	ty	-0.580	0.138	-1.269	0.109
st	ty	-0.335	0.742	-1.061	0.391

Let us read the Tukey results appearing above. The first and second columns represent the pair and the third column indicates the difference in cost for each pair of means. Differences between the *gor* brand and the others are reported in column 3 ($0.19 with *hc*, $0.42 with *sw*, and $1.44 with *ty*). The fourth column reports the probability associated with each difference. *Gor* is significantly less expensive than all brands except *hc* and *sw*.

In column 3, notice that, on the average, the *hc* brand costs $0.915 less than the *st* brand and $1.25 less than the *ty* brand. From the probability table, these differences are significant with probabilities of 0.015650 and 0.000672, respectively. The only other significant difference is that the average price for the *sw* brand costs $1.02 less than the *ty* brand.

A Two-Way ANOVA with Interaction

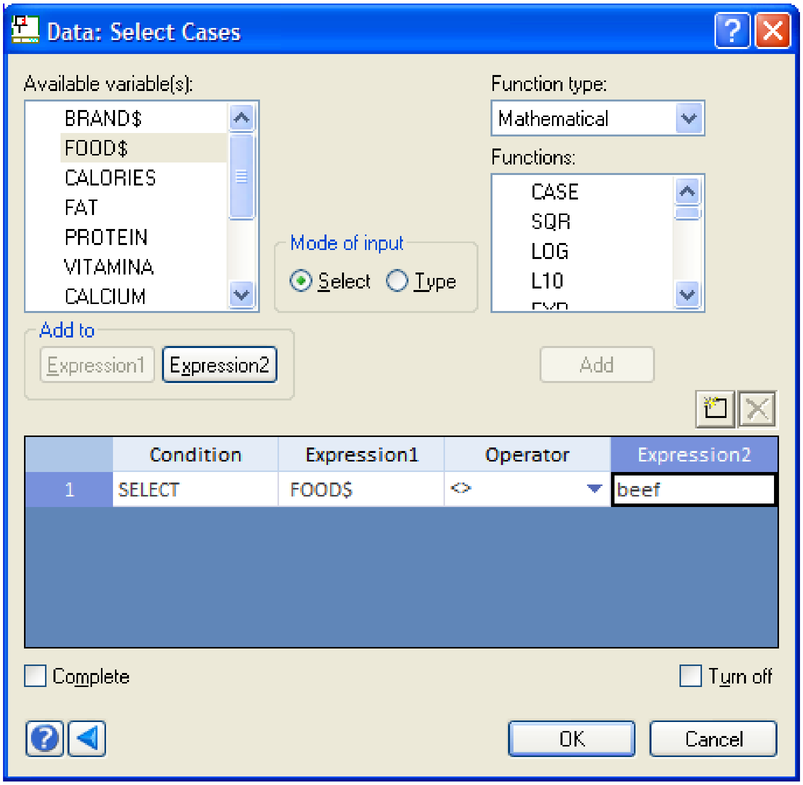
Do nutrients vary by type of food? Earlier, in a scatterplot matrix, we observed a small cluster of dinners that had higher calcium values than the others. In the two-sample *t*-test, we were unable to detect differences in average calcium values between the diet and regular dinners. Let us explore further by using both food type and dinner type to define cells—that is, we request a two-way analysis of variance. Using the Counts feature in Two-Way Tables, we found that although our sample has beef, chicken, and pasta dinners, there were no beef dinners in the *DIET$ yes* group. (SYSTAT can analyze ANOVA designs with missing cells. See SYSTAT, *Statistics* II, Chapter 3 for more information.)

Let us use Select Cases on the Data menu to omit the beef dinners, and then request an analysis of variance for a two-by-two design (*DIET$ yes* and *no* by *chicken* and *pasta*).

- From the menus choose:

Data
  Select Cases...

- In the Select dialog box, select *FOOD$* as Expression1.
- Select <> (not equal) from the drop-down list of operators.
- For Expression2, type 'beef' (include the quotation marks while working with commands, the dialog box takes care of this.).
- Click OK.

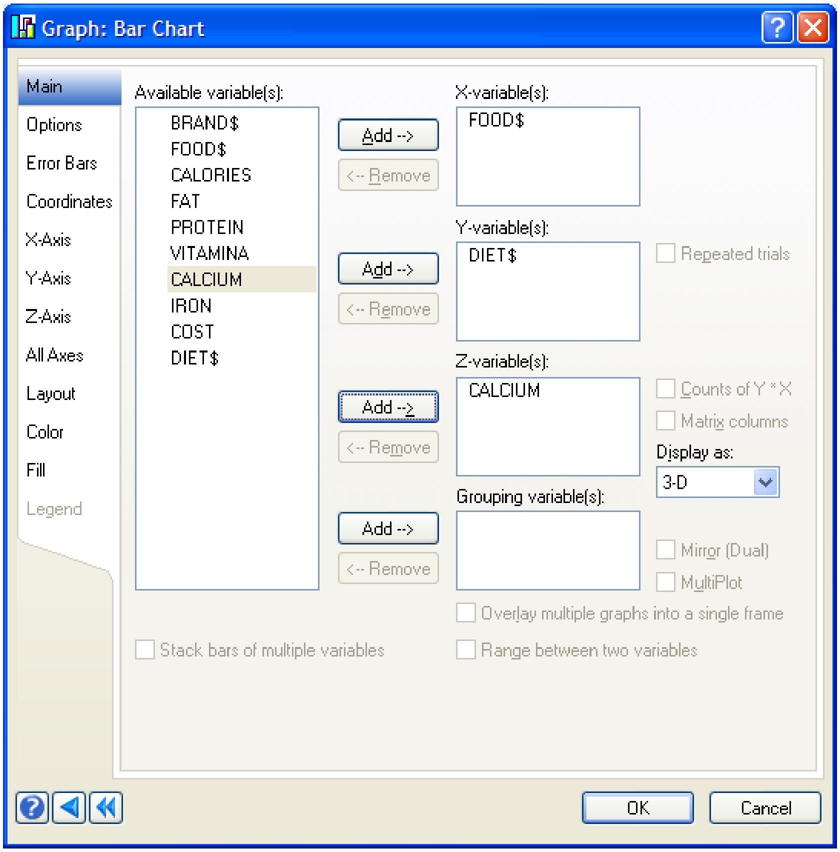


To get a bar chart of the cell means:

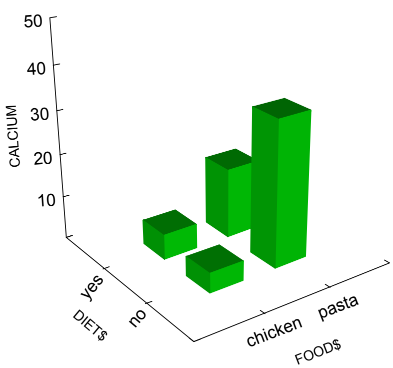
- From the menus choose:

Graph
  Bar Chart...

- Select *CALCIUM* as the Z-variable, *DIET$* as the Y-variable, and *FOOD$* as the X-variable.
- Click the Error Bar tab and select none from the type group.
- Click the Fill tab, select Select fill from the Fill pattern group, and select solid Fill Pattern.


Graph: Bar Chart
Main
Options
Error Bars
Coordinates
X-Axis
Y-Axis
Z-Axis
All Axes
Layout
Color
Fill
Legend
Available variable(s):
BRAND$
FOOD$
CALORIES
FAT
PROTEIN
VITAMINA
CALCIUM
IRON
COST
DIET$
Add -->
<-- Remove
X-variable(s):
FOOD$
Y-variable(s):
DIET$
Repeated trials
Z-variable(s):
CALCIUM
Counts of Y * X
Matrix columns
Display as:
3-D
Grouping variable(s):
Mirror (Dual)
MultiPlot
Overlay multiple graphs into a single frame
Stack bars of multiple variables
Range between two variables
OK
Cancel

■ Click OK.


50
40
30
20
10
CALCIUM
yes
no
DIET$
chicken
pasta
FOOD$

Suggestion. Try using the Dynamic Explorer to rotate this 3-D bar chart.

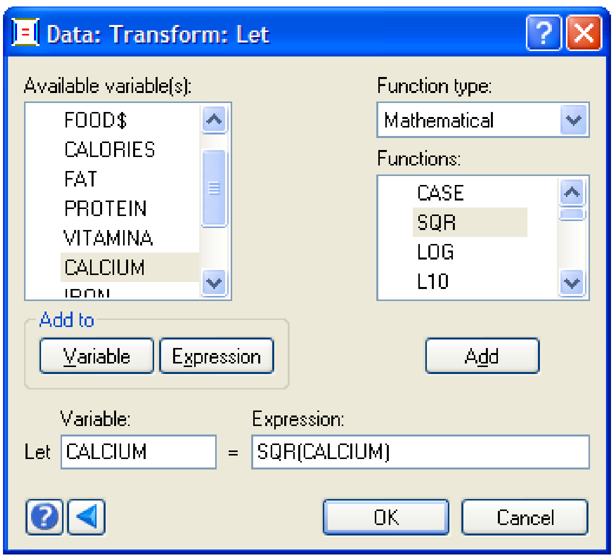
The box plot in the two-sample *t*-test example shows that the distributions of calcium for the *yes* and *no* groups are skewed and have unequal spreads. Let us use a root transformation of *CALCIUM* to make its distribution symmetric.

Before requesting the analysis of variance, we will transform *CALCIUM*, taking the square root of each value.

- From the menus choose:

Data
  Transform
    Let...

In the Let dialog box, select *CALCIUM* as the variable, select SQR from the list of mathematical functions, and select *CALCIUM* from the variable list and add it to the expression. The Expression box should now look like this: SQR(CALCIUM).



- Click OK.
- Now request the analysis of variance, repeating the steps in the last example, except that here we use *CALCIUM* as dependant variable and both *DIET$* and *FOOD$* as the factor variables.

```
Data for the following results were selected according to
SELECT ( FOOD$ <> 'beef')

Effects coding used for categorical variables in model.
The categorical values encountered during processing are
```

Variables	Levels	
DIET$ (2 levels)	no	yes
FOOD$ (2 levels)	chicken	pasta

Dependent Variable	CALCIUM
N	22
Multiple R	0.804
Squared Multiple R	0.647

Estimates of Effects B = (X'X)-1X'Y

Factor	Level	CALCIUM
CONSTANT		3.380
DIET$	no	0.305
FOOD$	chicken	-1.423
DIET$*FOOD$	no*chicken	-0.639

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-Ratio	p-Value
DIET$	1.807	1	1.807	1.432	0.247
FOOD$	39.298	1	39.298	31.136	0.000
DIET$*FOOD$	7.908	1	7.908	6.266	0.022
Error	22.719	18	1.262		

Least Squares Means

Factor	Level	LS Mean	Standard Error	N
DIET$	no	3.685	0.397	9.000
DIET$	yes	3.074	0.320	13.000

Least Squares Means

Factor	Level	LS Mean	Standard Error	N
FOOD$	chicken	1.956	0.303	14.000
FOOD$	pasta	4.803	0.410	8.000

Least Squares Means

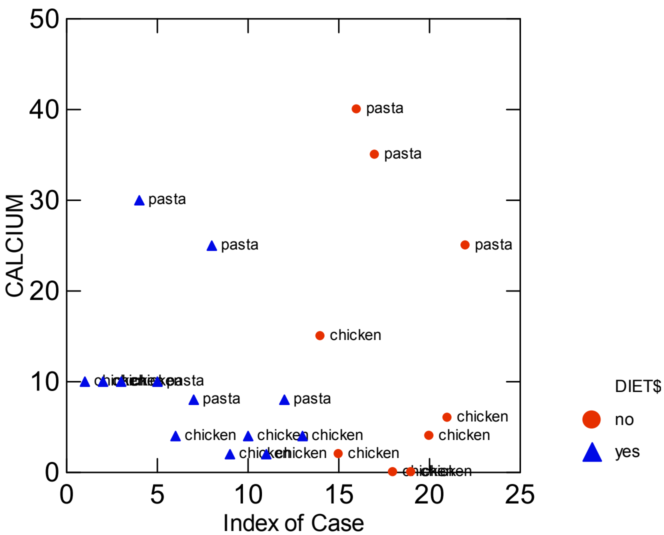
Factor	Level	LS Mean	Standard Error	N
DIET$*FOOD$	no*chicken	1.623	0.459	6.000
DIET$*FOOD$	no*pasta	5.747	0.649	3.000
DIET$*FOOD$	yes*chicken	2.289	0.397	8.000
DIET$*FOOD$	yes*pasta	3.859	0.502	5.000

The significant *DIET$* by *FOOD$* interaction suggests exercising caution when interpreting main effects. The main effect for *DIET$* does not appear to be significant (*p-value* = 0.247)—but let us look at a scatterplot and see if that tells us anything more.

- From the menus choose:

Graph
  Scatterplot...

- Select *CALCIUM* as the Y-variable and *DIET$* as the grouping variable. (SYSTAT will automatically use the case number as the X-variable.)
- Select Overlay multiple graphs into a single frame.
- Click the Symbol and Label tab, click Select symbol, select a circle for the first symbol and a triangle for the second.
- Check Display case labels in the Case labels group and select *FOOD$* as the case label variable.
- Click the Fill tab, click Select fill in the Fill pattern group, and select a solid fill for both the first and second fill patterns.
- Click OK.



The scatterplot shows that all of the dinners with a square root value for *CALCIUM* over 4 are *pasta* dinners (which is consistent with the significant main effect for *FOOD$*)—but it also shows that the highest values are also regular (*DIET$* = *no*) dinners. This suggests that further investigation might be warranted.

Bonferroni Pairwise Mean Comparisons

Since we have a significant *DIET$* by *FOOD$* interaction, we should be cautious about interpreting main effects. Let us use SYSTAT's advanced hypothesis testing capability to request Bonferroni adjusted probabilities for tests of pairwise mean differences.

- From the menus choose:

Analyze
  Analysis of Variance
    Pairwise Comparisons...

- Specify *DIET$* ∗ *FOOD$* under Groups and select Bonferroni under Test group.
- Click OK.

```
Post Hoc Test of CALCIUM
Using least squares means.
Using model MSE of 1.262 with 18 df.

Bonferroni Test

DIET$(i)*FOOD$-   DIET$(j)*FOOD$-
(i)               (j)                Difference   p-Value      95% Confidence Interval
                                                                    Lower          Upper
------------------------------------------------------------------------------------
no*chicken        no*pasta               -4.124     0.000          -6.478         -1.770
no*chicken        yes*chicken            -0.667     1.000          -2.464          1.131
no*chicken        yes*pasta              -2.236     0.025          -4.252         -0.221
no*pasta          yes*chicken             3.457     0.002           1.204          5.711
no*pasta          yes*pasta               1.888     0.201          -0.543          4.318
yes*chicken       yes*pasta              -1.570     0.148          -3.467          0.328
```

We are interested in four of the six differences (and probabilities) in these panels. First we look within diets and then within food types. For the:

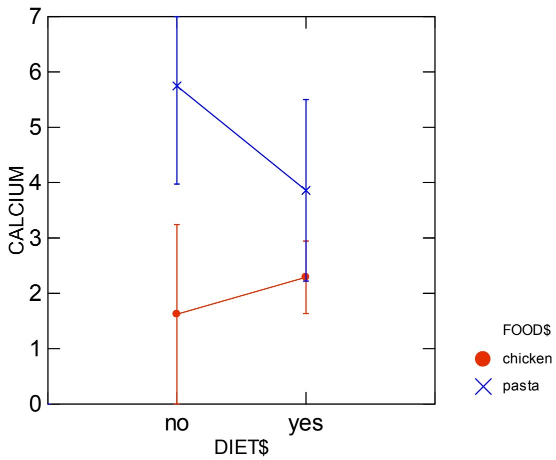
- *regular* meals (*DIET$ no*), the difference in average *CALCIUM* content between *chicken* and *pasta* meals is highly significant (the difference in square root units is 4.124, *p-value* = 0.001).
- *diet* meals (*DIET$ yes*), the difference in average *CALCIUM* content between *chicken* and *pasta* is not significant (1.570, *p-value* = 0.247).
- *pasta* meals, the difference in average *CALCIUM* content between the *DIET$ yes* and *no* groups is not significant (–1.888, *p-value* = 0.336).
- *chicken* meals, the difference in average *CALCIUM* content between *DIET$ yes* and *no* groups is not significant (0.667, *p-value* = 1.000).

It will be more clear if you see a dot display of these means.

- Select

Graph
Summary Charts
Dot...

- Choose *CALCIUM* as the Y-variable and *DIET$* as the X-variable.
- Specify *FOOD$* as the grouping variable.
- Select Overlay multiple graphs into a single frame.
- Click the Error Bars tab, choose Standard error from the Type group, and specify a value of 0.9545.
- Click Options tab and select Line connected in left-to-right order.
- Click OK.



For the *regular* meals (*DIET$ no)*, the error bars do not overlap, indicating a significant difference in calcium content between pasta and chicken. However, for the *diet* meals (*DIET$ yes*), the overlapping error bars suggest no significant difference between the meal types.

Focusing on the *pasta* meals, the average calcium content for the *diet* meals is within two standard errors of the average calcium content for the *regular* meals. Similar observations can be made for the *chicken* meals.

Summary

The first step in any data analysis is to look at your data. SYSTAT provides a wide variety of graphs that can help you identify possible relationships between variables, spot outliers that may unduly effect results, and reveal patterns that may suggest data transformations for more meaningful analysis.

SYSTAT also provides a wide variety of statistical procedures for analyzing your data. We have covered some of the most common and basic statistical techniques in this chapter, and we have still barely scratched the surface.

Chapter
4

Data Analysis Quick Tour

This chapter provides a quick tour of SYSTAT's capabilities, using data from a survey of uranium found in groundwater.

Groundwater Uranium Overview

The U.S. Department of Energy collected samples of groundwater in west Texas as part of a project to estimate the uranium reserves in the United States. Samples were taken from five different locations, called producing horizons, and then measured for various chemical components. In addition, the latitude and longitude for each sample location were recorded. Several questions are of interest:

- Does the uranium concentration vary by producing horizon?
- Is the presence of uranium correlated to the presence of other elements?
- What is the overall geographic distribution of uranium in the area?

The data for the groundwater uranium study are in the file *GDWTRDM*. Measurements were recorded for the following variables:

Variable	Description
SAMPLE	The ID of the groundwater sample
LATITUDE	Latitude at which the sample was taken
LONGTUDE	Longitude at which the sample was taken
HORIZON$	Initials of producing horizon
HORIZON	ID of producing horizon
URANIUM	Uranium level in groundwater
ARSENIC	Arsenic level in groundwater
BORON	Boron level in groundwater
BARIUM	Barium level in groundwater
MOLYBDEN	Molybdenum level in groundwater
SELENIUM	Selenium level in groundwater
VANADIUM	Vanadium level in groundwater
SULFATE	Sulfate level in groundwater
TOT_ALK	Alkalinity of groundwater
BICARBON	Bicarbonate level in groundwater
CONDUCT	Conductivity of groundwater
PH	pH of groundwater
URANLOG	Log of uranium level in groundwater
MOLYLOG	Log of molybdenum level in groundwater

Potential Analyses

The following kinds of analyses may be useful in analyzing the groundwater data:

- Basic Statistics
- Transformations
- ANOVA
- Nonparametric tests
- Regression
- Correlation
- Cluster analysis
- Discriminant analysis

- Spatial statistics
- Smoothing techniques such as kriging
- Contour plotting

In these examples, we will show you descriptive graphs, ANOVA, nonparametric tests, smoothing and contour plotting.

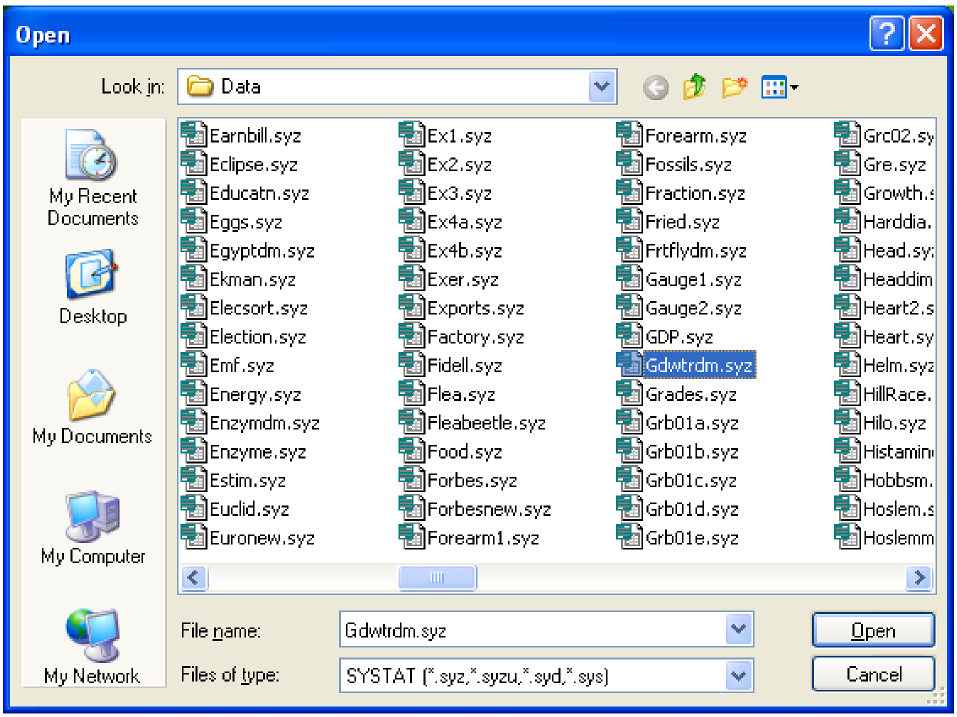
The Groundwater Data File

The data for this analysis are in the file *GDWTRDM.*

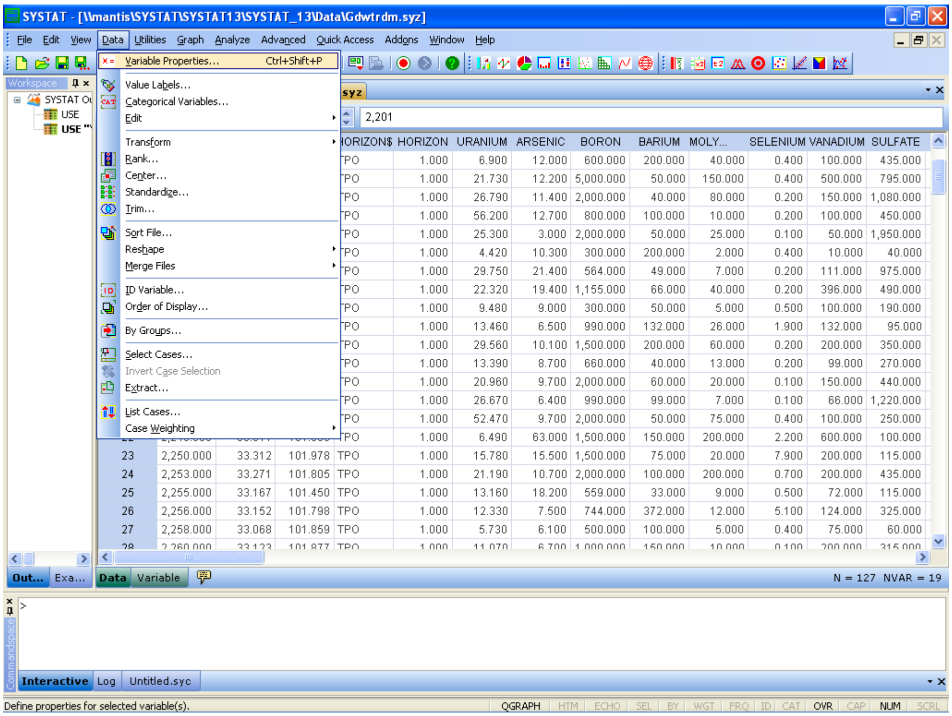
- To open the file, from the menus choose:

File
  Open
    Data...

- Select *GDWTRDM*, and click Open.



Data files that are opened or imported can be viewed and edited in the Data editor. You can also see the results of transform variables, select cases and so forth in the Data editor. In this example, measurements were taken of the levels of uranium and various other elements in the groundwater at each producing horizon. The measurements for each variable can be viewed and manipulated directly in the Data editor.



Graphics

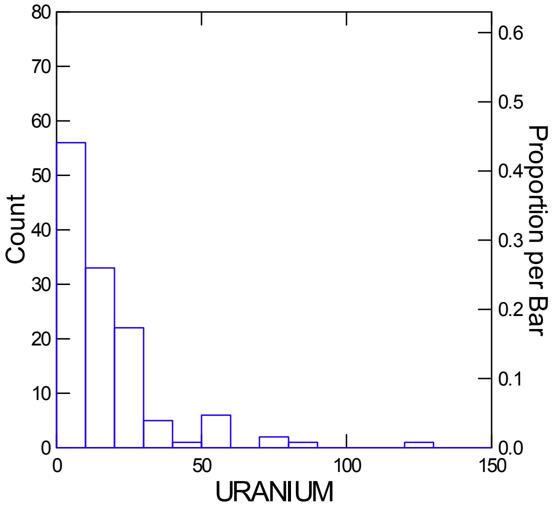
Distribution Plot

Since we will be looking extensively at uranium levels, it is a good idea to take a look at the distribution of this variable and make sure it meets assumptions for future analyses.

To plot a histogram of *URANIUM*:

- Click the Histogram icon  in the Graph Toolbars.
- Choose *URANIUM* and add it to the X-variable(s) list.
- Click OK.

SYSTAT displays the following plot in the Graph editor:

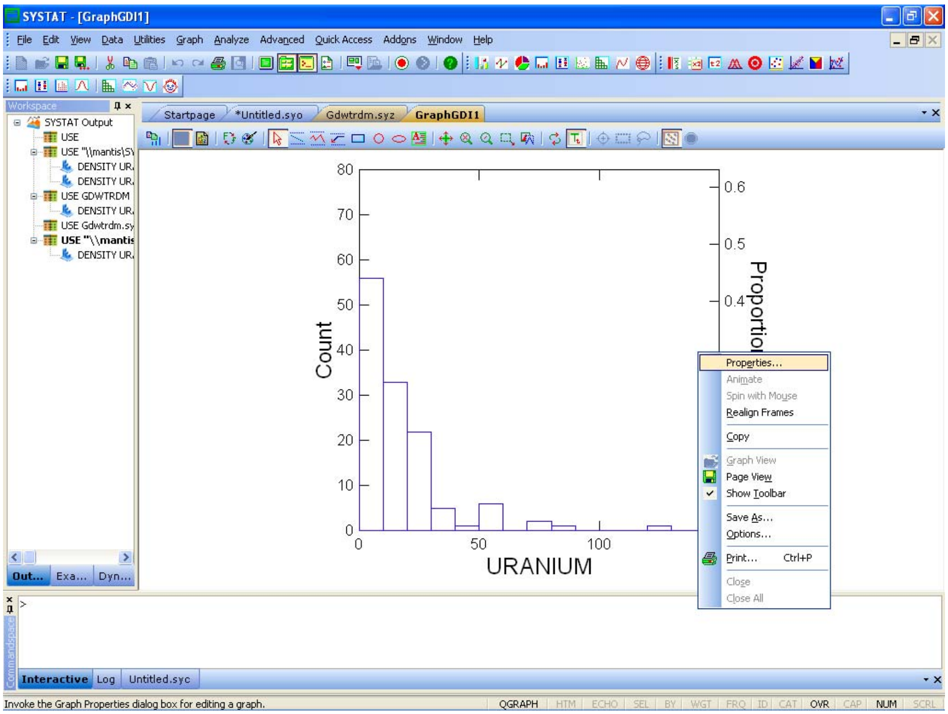


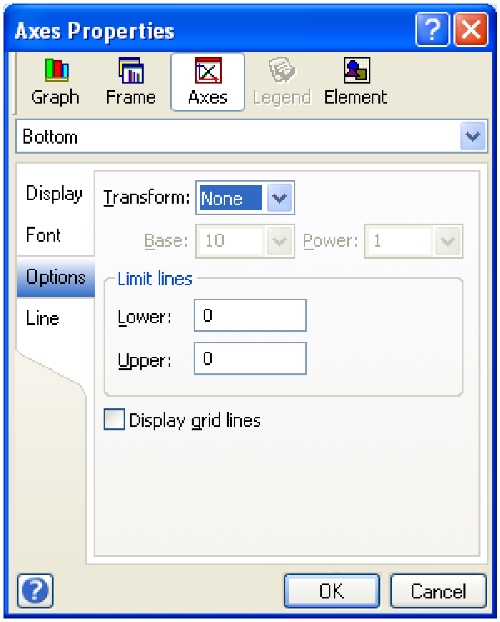
We can see that the distribution of *URANIUM* is skewed. To properly apply most statistical analyses, the histogram should show a bell-shaped, normal distribution.

Exploring the Groundwater Data Interactively

The Graph Properties dialog box is a tool that allows you to explore data interactively, increasing the efficiency of your analysis. It can be used to modify features of a graph or frame or elements of the graph.

■ To open the Graph Properties dialog box right-click on the graph. And click the Properties option to open the Graph Properties dialog box.



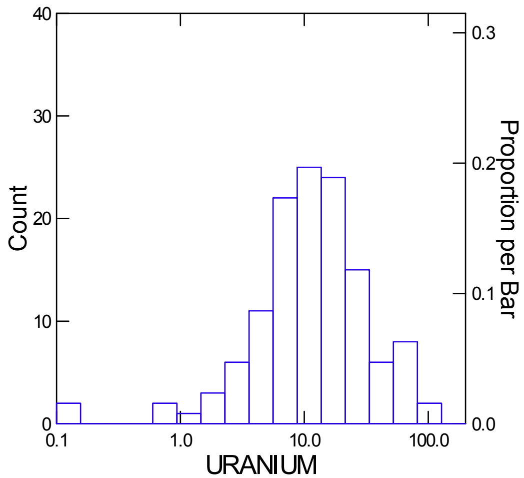


- Click the Axes tab in the Graph Properties dialog box and then select the Options tab. Select Power in the Transform combo box. This will enable the power combo box.
- Use the down arrow key in the keyboard to change the power value of the X-axis until the graph becomes a bell-shaped curve.

As you do this, SYSTAT is automatically calculating the power data transformation of the form *URANIUM* (power). A power of 0.5 is a square root transformation. A power of 0.333 is a cube root transformation.

Transformed Graph

At a power of 0, SYSTAT automatically performs a logarithmic transformation- for example, log (*URANIUM*). The log transformation appears to produce a very good bell-shaped curve. But this judgment is subjective and it is possible to use more formal and objective methods to examine the normality of the transformed data.



Normally, once the proper transformation has been identified using the Graph Properties dialog box, you create the transformed variable using the Data editor. We have already performed the transformation and included the variable *URANLOG* in the data file for further statistical analysis.

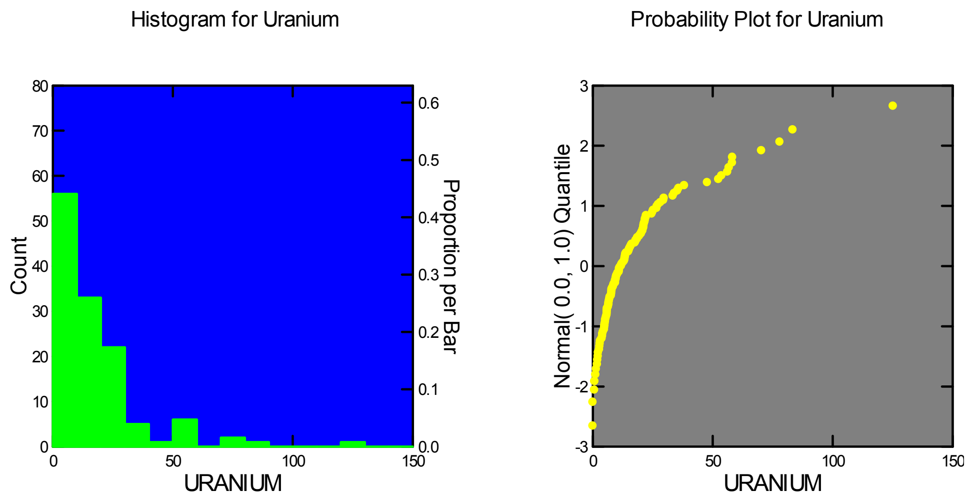
Histograms and Probability Plots

Let us take another look at the *URANIUM* distribution. We are going to plot two graphs, a histogram and a probability plot, by using commands. From the menus, submit the command file *GDWTR1DM*. For this:

- From the menus choose:

File
 Submit
  File...

- Select *GDWTR1DM* from the 'Miscellaneous' subfolder of the 'command' directory and click Open.
- The following graphs are displayed in the Output editor of the Viewspace:



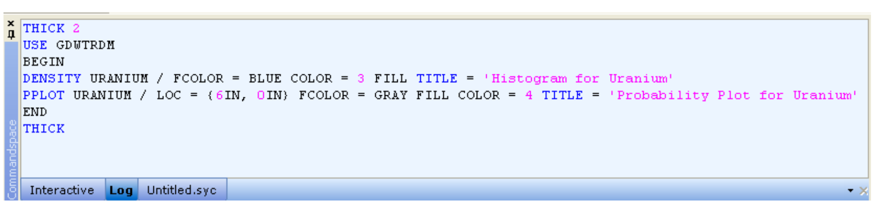
In this plot, we begin to glimpse SYSTAT's color and overlay capabilities. This command file created a side-by-side overlay of a histogram and a probability plot of the *URANIUM* variable.

SYSTAT Windows and Commands

SYSTAT gives you the flexibility to perform your analysis the way you want:

- Windows interface: icons, menus, and dialog boxes.
- Typed commands: typing commands at the Commandspace.
- Batch (Untitled) command files: submitting files directly or from the Commandspace.

Additionally, all menu actions can be optionally echoed to the Output editor, allowing you to perform initial analyses using the menus, and then to cut and paste the commands into the Untitled tab of the Commandspace for repeated use.



```
THICK 2
USE GDWTRDM
BEGIN
DENSITY URANIUM / FCOLOR = BLUE COLOR = 3 FILL TITLE = 'Histogram for Uranium'
PPLOT URANIUM / LOC = {6IN, 0IN} FCOLOR = GRAY FILL COLOR = 4 TITLE = 'Probability Plot for Uranium'
END
THICK
```

Plotting Several Graphs Using Commands

The commands in the file *GDWTR1DM* are:

```
THICK 2
USE GDWTRDM
BEGIN
   DENS URANIUM / HIST, FCOLOR = BLUE,
                  COLOR = GREEN, FILL,
                  TITLE='Histogram for Uranium'
   PPLOT URANIUM / LOC = {6in,0in}, FCOLOR = gray,
                   FILL, COLOR = YELLOW,
                   TITLE = 'Probability Plot for Uranium'
END
THICK 1
```

The DENS and PPLOT commands create the histogram and the probability plot, respectively. Between the BEGIN and END statements, we can change the data file in use and plot an unlimited number of graphs. Each graph can have its own attributes, such as location and color.

Plotting Several Graphs Using Menus

Plotting more than one graph can be accomplished directly from SYSTAT's menu.

- From the menus choose:

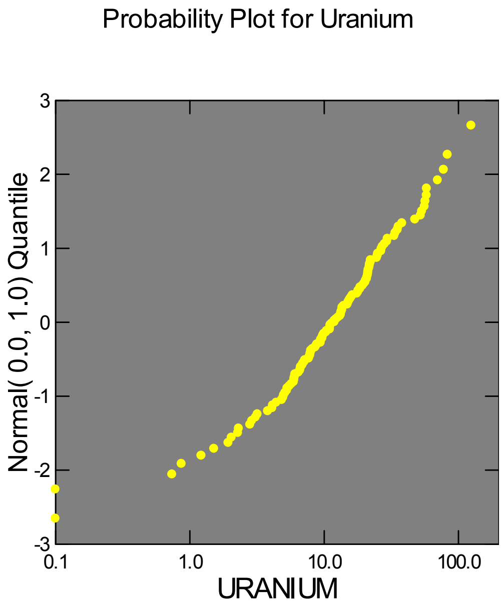
Graph
  Begin Overlay Mode

- Choose graphs and options from menus and dialog boxes. You can choose locations for the graphs in the Layout tab, unless you want them overlaid on top of one another.
- Then, from the menus choose:

Graph
  End Overlay Mode (Display)

Transforming Data and Selecting Cases

In the Commandspace, select and submit the line beginning with PPLOT. Using the Graph Properties dialog box in the Workspace, transform the *URANIUM* variable by clicking the down arrow of X-Power until 0 is reached, yielding a log transformation.



Notice that the probability plot is much more linear.

Using SYSTAT's lassoing capability, you can isolate outliers.

■ Click the Lasso icon 

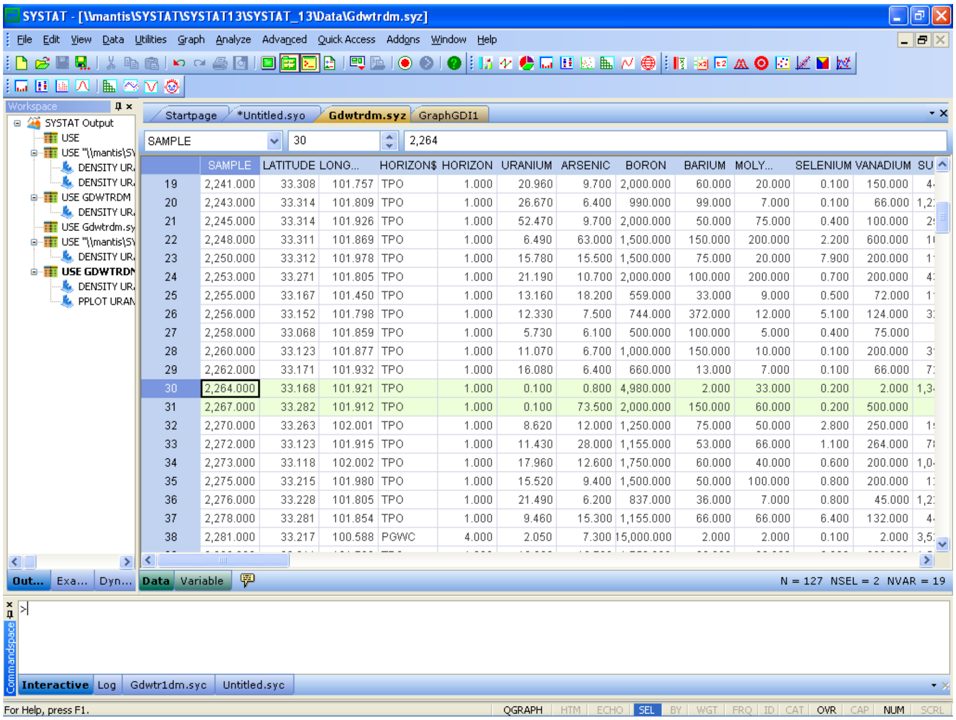
and lasso the two outliers on the lower left of the graph by holding down the left mouse button and circling them.

■ Click the Show Selection icon 

to highlight the selected cases.

Dynamically Highlighted Cases

Cases selected by the Lasso tool are highlighted in the Data editor. Click on the Data Editor to see these cases, 30 and 31, directly.



SYSTAT dynamically links data across graphs and the Data editor. These cases are now selected. If you were to run a statistical analysis or plot another graph at this point, it would use only these two cases. As pointed out earlier, SYSTAT manages data and graphics globally.

Make sure you deselect the data before continuing. Otherwise the remainder of the analyses will be done only on the selected observations. To deselect the cases, use the Lasso tool to select an area of the graph that contains no data points.

Connections between Graphs and the Data Editor

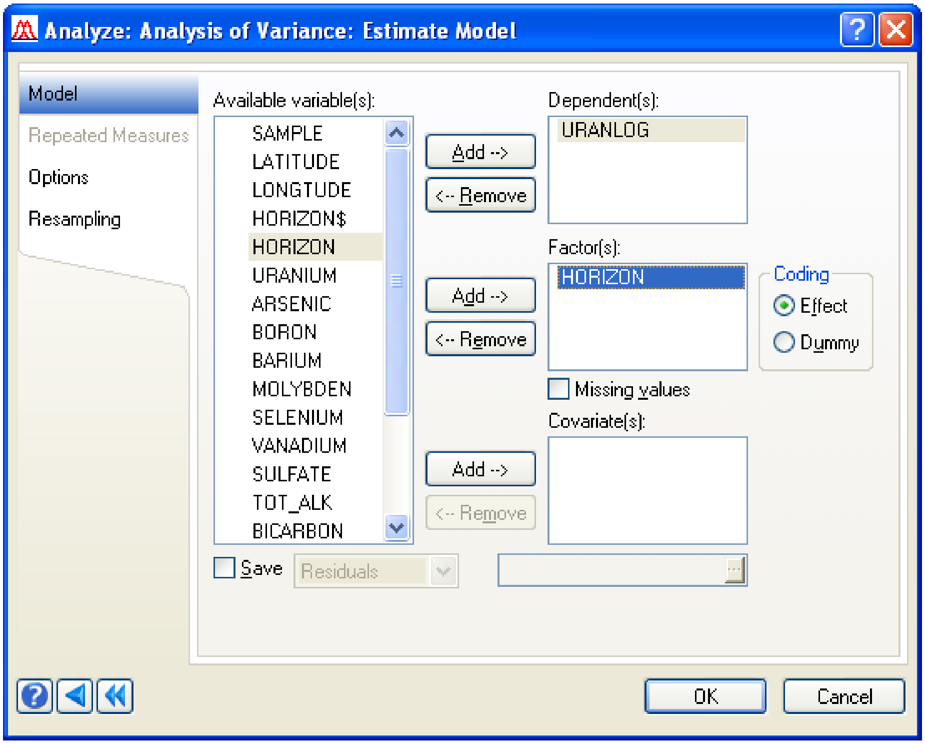
For those of you with a technical inclination, here is the explanation of the connection between the graphs and the Data editor:

- Graphs have their own data, allowing the real-time transformations of the Graph Properties dialog box and the ability to save and reload them without the original data file.
- When a graph is plotted, the data in the graph are linked to the Data editor, allowing lassoing.
- The Data editor and the program kernel share the same data set, so all data are "live," and what you see is what you get. For example, if you select data in the Graph editor and then run a regression, the regression applies only to the selected data.

Statistics

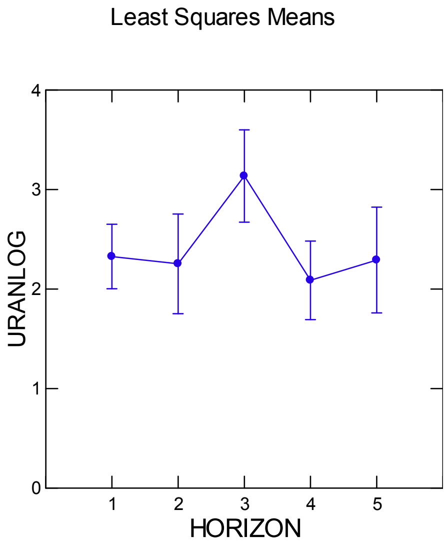
This part of the tour introduces SYSTAT's statistics capability. Here, we explore the question of whether the five producing horizons have varying levels of uranium by performing an ANOVA of *URANLOG* (the log of *URANIUM*) versus *HORIZON*. This analysis is being done based on the visual judgment that the normal distribution for log(*URANIUM*) is a valid model.

- In the SYSTAT window, click the ANOVA icon on the Statistics toolbar.
- Select *URANLOG* as the dependent variable and *HORIZON* as the factor.
- Click on Options tab.
- Check Shapiro-Wilk option.
- Click OK.



Graph of Mean Uranium Levels

Along with numeric output, SYSTAT produces a Quick Graph: a line-connected plot of mean uranium levels and confidence intervals for the different producing horizons.



Most of SYSTAT's statistical procedures have associated Quick Graphs. Quick Graphs speed up analysis by providing immediate visual feedback on results. In this Quick Graph, it is easily seen that the third group, Quartermaster, has a much higher level of uranium.

Output for ANOVA

The numeric output of the ANOVA appears in the Output editor.

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
HORIZON	14.978	4	3.744	3.252	0.014
Error	140.484	122	1.152		

In the Analysis of Variance table, the *F* test has a *p-value* of 0.014, meaning that there is only a 1.4% chance that these data would be measured if the individual producing horizons have the same average level of uranium—that is, the uranium level differs significantly by producing horizon. We saw this immediately in the Quick Graph. In fact, in the Quick Graph we also saw that producing horizon 3, the Quartermaster horizon, differs the most.

Outliers and Diagnostics

The Output editor also has warnings about outliers.

```
*** WARNING *** :

Case  30 is an Outlier        (Studentized Residual : -4.732)
Case  31 is an Outlier        (Studentized Residual : -4.732)

Test for Normality
                 ¦ Test Statistic   p-value
-----------------+------------------------
Shapiro-Wilk Test ¦         0.913     0.000

Durbin-Watson D Statistic   ¦ 1.305
First Order Autocorrelation ¦ 0.345
```

There are two outliers in the data: cases 30 and 31. These are the same two that we lassoed earlier in the probability plot.

SYSTAT performs diagnostics to verify that the data meet the underlying assumptions for ANOVA, Linear Regression, and General Linear Models (GLM). Diagnostics speed up the analysis and help to produce more accurate results by alerting you to problems with the data. Both the Durbin-Watson *D* statistic and the first-order autocorrelation appear by default and these are parts of such diagnostics.

The Options tab provided in the ANOVA dialog box performs diagnostics. The Shapiro-Wilk option performs the test for normality of residuals. From the above output of *Test for Normality*, the *p-value* is an indication (as in any hypothesis testing results) of whether the hypothesis being tested (in this case the normality of the residuals) is to be accepted or rejected.The smaller the *p-value* the stronger is the evidence against the hypothesis. Since in this case the value is near 0 (0 up to 3 places of decimal) the normality hypothesis of residuals is rejected. When the assumption of normal residuals cannot be justified even for a transformed variable, we may consider nonparametric methods, which do not depend on such assumptions.

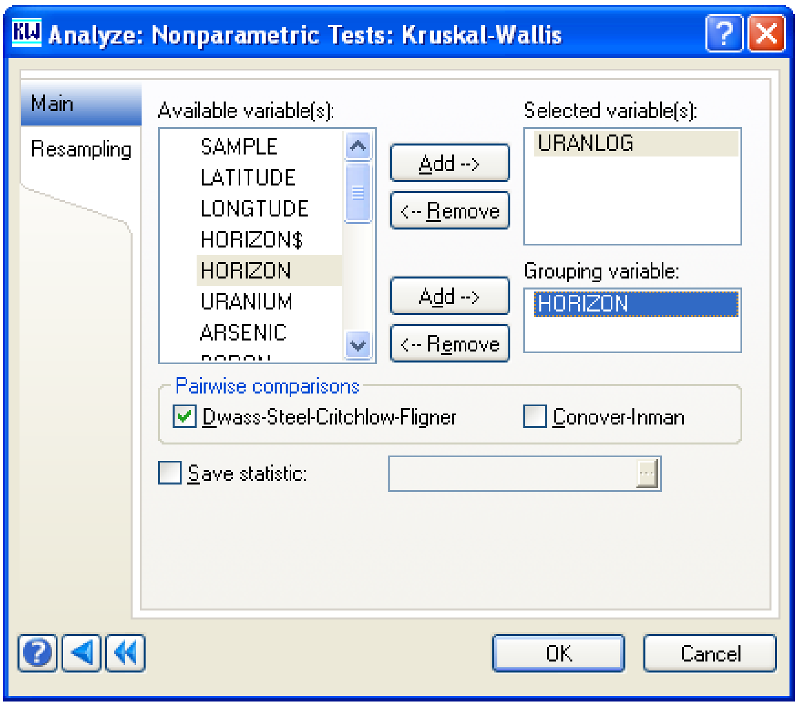
Nonparametric Tests

Now we see how the question earlier answered by using ANOVA (with normality assumption on residuals) can be answered by a nonparametric test, which does not make this assumption. Now you might ask: Why then bother with ANOVA at all? The answer is: If the normality assumption actually holds, then ANOVA is a more powerful method, but it is not valid when the assumption fails. If we do not have a good distribution model for *URANLOG* or a transformed variable, then it is safer to use a distribution-free (nonparametric) method, even if it is not powerful. For a nonparametric test for the equality of *URANLOG* levels at various horizons:

From the menus choose:

Analyze
  Nonparametric Tests
    Kruskal-Wallis…

- Select *URANLOG* as the Selected variable(s) and *HORIZON* as the Grouping variable.



Output from Kruskal-Wallis Test

```
Kruskal-Wallis One-way Analysis of Variance for 127 Cases

Categorical values encountered during processing are

  Variables          ¦                     Levels
 --------------------+-------------------------------------------------
  HORIZON (5 levels) ¦  1.000     2.000     3.000     4.000     5.000

  Dependent Variable  ¦ URANLOG
  Grouping Variable   ¦ HORIZON

  Group   Count   Rank Sum
 -------------------------
  1          43   2851.500
  2          18    986.000
  3          21   1880.500
  4          29   1455.000
  5          16    955.000

Kruskal-Wallis Test Statistic : 15.731
p-value is 0.003 assuming Chi-square Distribution with 4 df
```

From the Kruskal-Wallis One-way Analysis of Variance table, the chi-square test has a *p-value* 0.003, meaning that there is only 0.3% chance that these data would show this much difference between the groups if the individual producing horizons have the same average level of uranium. Thus we conclude that the uranium level differs significantly for producing horizons. We arrived at the same qualitative conclusion from ANOVA and its Quick Graph, but it was quantitatively different. The *p-value* in ANOVA was 0.014; here it is 0.003.

Advanced Graphics

This part of the tour explores SYSTAT's advanced graphics capabilities, including 3-D rotation, animation, zooming using the Dynamic Explorer, smoothers, contour plots, and Page view. (The graphics in this section are best viewed in 16-bit or 32-bit true color on a high-resolution monitor.)

From the preceding statistical analysis, we can conclude that there are differences in the uranium level between the producing horizons. However, we also have the latitude and longitude for each sample, so we can perform a geographic analysis to better pinpoint the variations in uranium. To accomplish this, we will apply a smoothing technique called "kriging" (pronounced *kree*-ging) to fit a 3-D scatterplot of uranium by latitude and longitude. Kriging is a smoothing technique often used in geostatistics. It uses local information around points to extrapolate complex and irregular geographic patterns.

Kriging Smoother

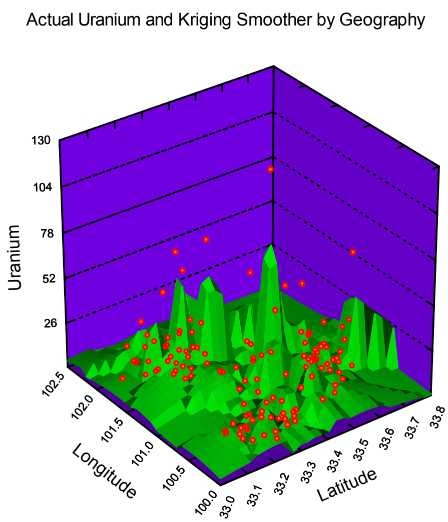
From the menus, submit the file *GDWTR2DM.*

- From the menus choose:

File
  Submit
    File...

- Select the file *GDWTR2DM* from the 'Miscellaneous' subfolder of the 'command' directory and click Open.

The following graph is displayed in the Output editor:



This plot shows the level of uranium against latitude and longitude (the data points) and the kriging smoother (the surface). The plot provides us with a topography of the uranium level, and we can see immediately that there is a pronounced peak near the center of the sampling area.

Rotation

If you look at the Dynamic Explorer, the rotation arrows have been activated. The rotation arrows can be used interactively to rotate the plot in three dimensions, allowing you to examine your data from all angles. Try pressing each of the four rotation keys to examine how the plot changes.

Notable features include:

- True graphical rotation with automatic recalculation of the graph upon each rotation. (SYSTAT does not just rotate a picture or bitmap, it physically transforms the graph data and replots the graph and all of its elements in real time with each rotation.)
- Realistic 3-D lighting to increase the volume effect.
- Notable 3-D fonts on each axis that rotate along with the graph.
- The ability to view from all angles, including above and below.
- Closer data points look larger and more distant points look smaller.

Smoothers

SYSTAT offers 126 nonparametric smoothers for exploratory analysis. In addition, nineteen smoothers can be directly added to graphical output. The smoothing options available for scatterplots are:

None	LOWESS	Inverse	Andrews
Linear	DWLS	Mean	Bisquare
Quadratic	Spline	Median	Huber
Log	Step	Mode	Trimmed
Power	NEXPO	Midrange	Kriging

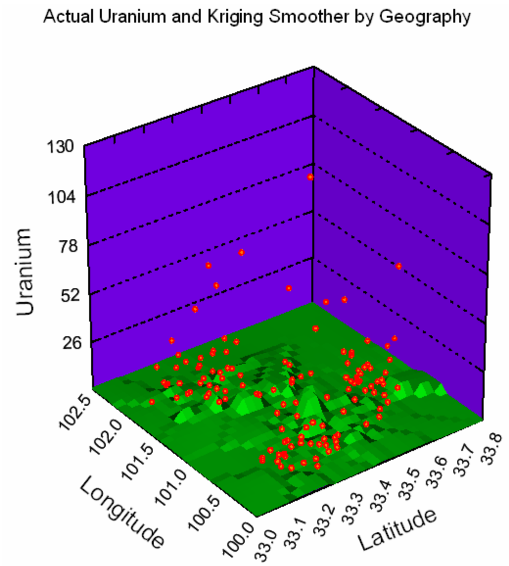
Smoothers help you view your data in unique and informative ways. In this case, we are using kriging because it is especially designed for examining spatial distributions such as mineral deposits.

Tension of Smoothers

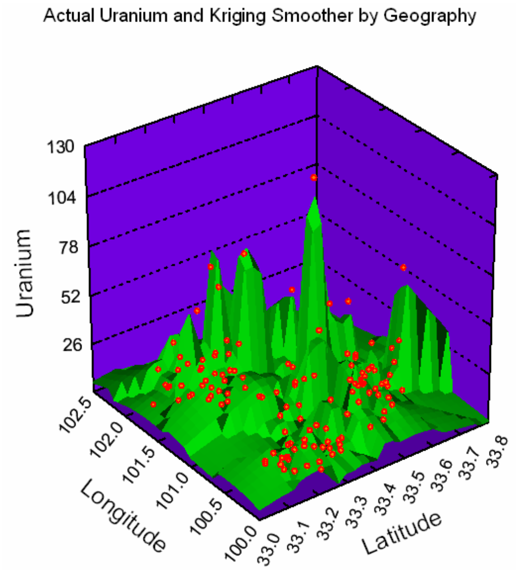
Each smoother has a tension associated with it. If you consider the smoother to be a string or membrane loosely attached to each data point, then the higher the tension on the ends of the string, the less influence any individual point has and the smoother averages across them all. The lower the tension on the ends of the string, the greater the influence of the individual data points, and the smoother approaches a path that passes through each point.

In addition to rotation, with the help of Graph Properties dialog box, you can also alter the tension of the kriging smoother.

- To open the Graph Properties dialog box, right-click on the graph editor and select Properties.
- Click the Graph tab in the Graph Properties dialog box.
- Use the up arrow key in the keyboard to select the graph as "Actual Uranium and Kriging Smoother by Geography".
- Now, click on the Element tab and select the Smoother tab.
- Select Kriging from the Method combo box.
- Use the down arrow key to change the tension value from 0.35 to 0.90 in Tension combo box.



Notice how the surface becomes flatter and lower -- recall from the histogram that most samples have a low value for the uranium level. Decrease the tension from 0.90 to 0.10.



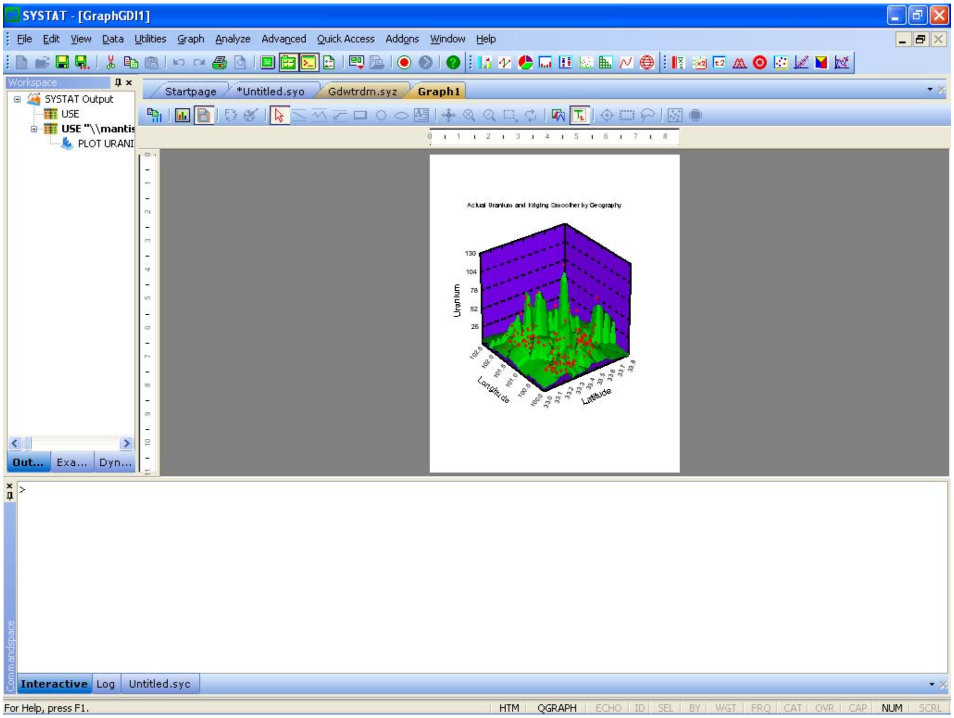
Notice how the surface reaches out to each individual point.

Page View

If at this point you switch to the Page view by selecting from the menu,

View
  Page View...

You can see that you have the capabilities from the Dynamic Explorer (rotation, animation, and zoom) available in Page as in Graph view. In addition, you can position the chart by dragging it around on the page.



Contour Plot of the Kriging Smoother

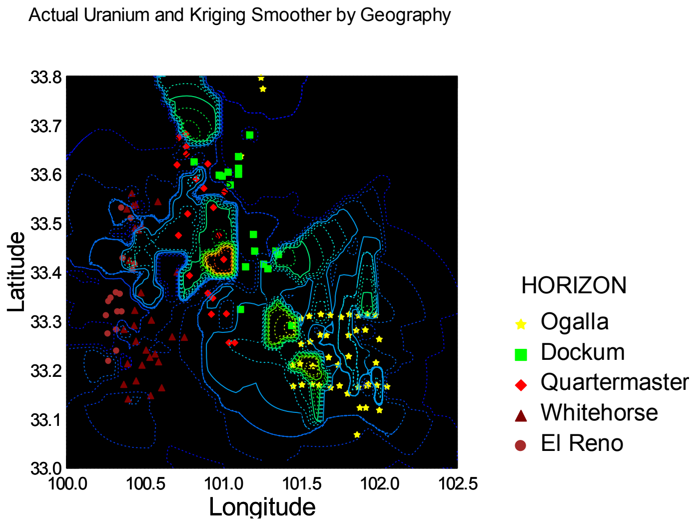
So far we have looked at this data by producing horizon and by latitude and longitude. SYSTAT allows us to combine these two pieces of information by tailoring and coloring symbols. As a final analysis, we will use another advanced graphing technique: a contour plot of the kriging smoother. This final plot consists of successive vertical slices through the surface of the kriging smoother overlaid on the data coded by producing horizon. From the menus, submit the file *GDWTR3DM*.

- From the menus choose:

File
  Submit
    File...

- Select *GDWTR3DM* from the 'Miscellaneous' subfolder of the 'command' directory and click Open.

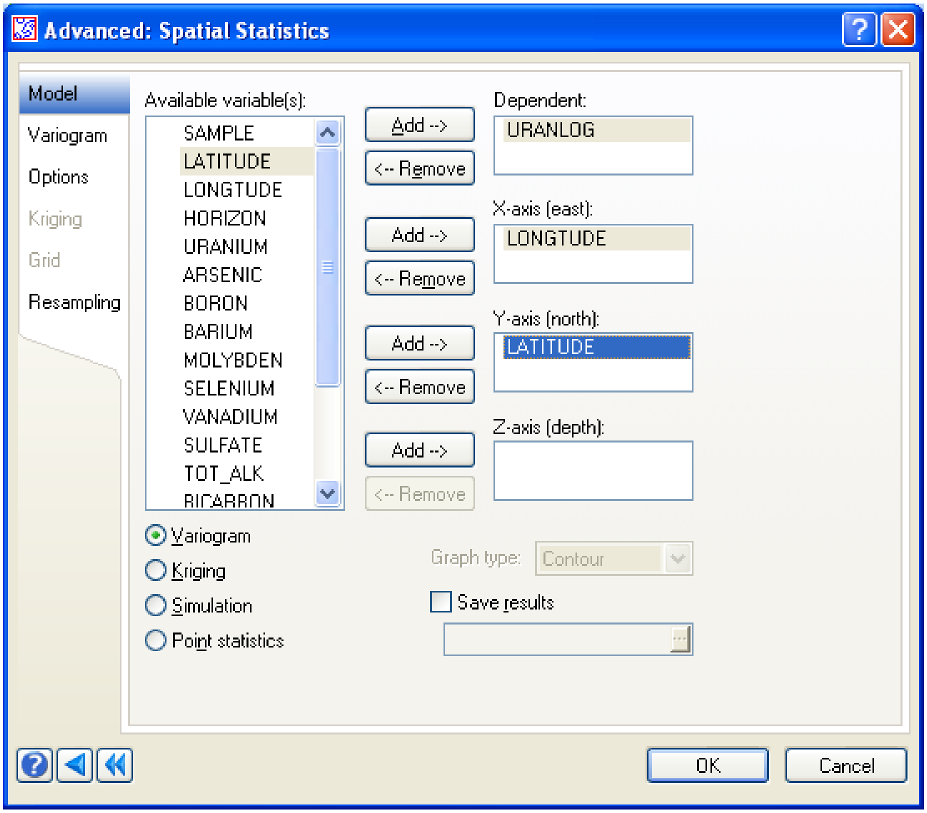
The following graph is displayed:



The plot is simply a different view of the 3-D plot, but now we can use the contours to pinpoint the high levels of uranium with respect to the producing horizons. The peaks of the kriging smoother are represented by tighter, brighter yellow and red contours, while the valleys are represented by dashed blue and green contours. The actual data points are distinguished in color and symbol by producing horizon. Notice how the peak is in the middle of the Quartermaster group; this is why it had the highest value in the earlier ANOVA. We can also see that the uranium level is not uniformly higher throughout this producing horizon but is highly localized.

Advanced Statistics

The kriging smoother provided a quick geographic visualization of uranium concentrations. SYSTAT also provides a comprehensive spatial statistics procedure for analyzing and modeling geographic data. You can create variograms and perform stochastic simulation or kriging.



Summary

At this point, we have made some significant discoveries about the groundwater data: we know exactly where the uranium is geographically concentrated both in terms of producing horizon and latitude and longitude. We also have some very high-quality graphics to communicate our findings in print or in a presentation. SYSTAT has taken us from data to discovery.

By the way, this groundwater application has many other areas to explore other than the few that we have examined in this tour. For example, we have not even looked at the relationships between uranium and the other elements in the data set. You are encouraged to explore the power of SYSTAT further through this application, beginning with any of the other potential analyses mentioned earlier.

Alternatively, examine any of the other 16 applications provided with SYSTAT. You can access them either through the Application Gallery in the Help system Table of Contents or through the chapter "Applications" on p. 247 in the *Getting Started* manual.

References for Groundwater Data

The groundwater data used in these examples were obtained from the following sources:

Original Source. Nichols, C. E., Kane, V. E., Browning, M. T., and Cagle, G. W. (1976). National Uranium Resource Evaluation, *Northwest Texas Pilot Geochemical Survey,* Union Carbide Corporation, Nuclear Division, Oak Ridge Gaseous Diffusion Plant, Oak Ridge, Tenn., K/UR-1, U.S. Department of Energy, Grand Junction, Colo., GJBX-60(76), 231.

Data Reference. Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, 123–126. Springer-Verlag, New York.

Chapter

5

Command Language

(*Revised by Rajashree Kamath*)

Most SYSTAT commands are accessible from the menus and dialog boxes. When you make selections, SYSTAT generates the corresponding commands. Some users, however, may prefer to bypass the menus and type the commands directly at the command prompt. This is particularly useful because some options are available only by using commands, not by selecting from menus or dialog boxes. Whenever you run an analysis--whether you use the menus or type the commands--SYSTAT stores the processed commands in the command log.

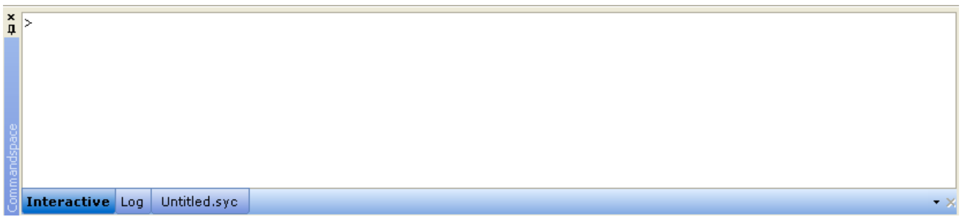
A command file is simply a text file that contains SYSTAT commands. Saving your analysis in a command file allows you to repeat it at a later date. Many government agencies, for example, require that command files be submitted with reports that contain computer-generated results. SYSTAT provides you with a command file editor in its Commandspace.

You can also create command templates. A template allows customized, repeatable analyses by allowing the user to specify characteristics of the analysis as SYSTAT processes the commands. For example, you can select the data file and variables to use on each submission of the template. This flexibility makes templates particularly useful for analyses that you perform often on different data files, or for combining analytical procedures and graphs.

Commandspace

Some of the functionality provided by SYSTAT's command language may not be available in the dialog box interface. Moreover, using the command language enables you to save sets of commands you use on a routine basis.

Commands are run in the Commandspace of the SYSTAT window. The Commandspace has three tabs, each of which allows you to access a different functionality of the command language.



Interactive tab. Selecting the Interactive tab enables you to enter the commands in the interactive mode. Type commands at the command prompt (>) and issue them by pressing the Enter key. You can save the contents of the tab (SYSTAT excludes the prompt), and then use the file as a batch file.

Log tab. Selecting the Log tab enables you to examine the read-only log of the commands that you have run during your session. You can save the command log or submit all or part of it.

Batch (Untitled) tab(s). Selecting a Untitled tab enables you to operate in batch mode. You can open any number of existing command files, and edit or submit any of these files. You can also type an entire set of commands and submit the content of the tab or portions of it. This tab is labeled Untitled until its content is saved. The name that you specify while saving the content replaces the caption 'Untitled' on the tab.

When the Commandspace is active, you can cycle through its tabs using the following keyboard shortcuts:

- CTRL+ALT+TAB. Shifts focus one tab to the right.
- CTRL+ALT+SHIFT+TAB. Shifts focus one tab to the left.

Although each tab provides a unique function, you can save the contents of any Commandspace tab to a command file for subsequent submission to SYSTAT.

What Do Commands Look Like?

Here are some examples of SYSTAT commands:

```
1  XTAB
2       USE food
3       PLENGTH NONE/ LIST
4       TAB food$ brand$ diet$
5       CSTAT
6       BY diet$
7       CSTAT / MEDIAN  MIN  MAX  MEAN  CI
8       BY
9  CORR
10      PEARSON calories fat protein cost / BONF
11      SPLOM   calories fat protein cost
12      PLOT calories * protein / LABEL=brand$
```

The CSTAT command on line 5 produces a set of descriptive statistics for all seven numeric variables in the *FOOD* data file. Line 7 asks for the median, minimum, maximum, means, and confidence intervals for all of the variables.

SYSTAT commands are made up of keywords meaningful to the function that they perform on execution. As far as possible, all meaningful words associated with a given function are applicable. For example, CSTATISTICS, CSTATS, and STATISTICS will all give you descriptive statistics. Likewise, PLENGTH or DISPLAY will both allow you to specify the length of output produced by a given command. A keyword will typically be made of letters of the alphabet, and sometimes numbers. All other characters like the hyphen and underscore are avoided; a space and some other characters like the plus (+), minus(-), asterisk (*), hash (#) and exclamation mark (!), are not used as they may be used in other parts of a command.

Interactive Command Entry

Commands can be issued automatically when the Interactive tab is selected in the Commandspace. To issue a command, type the command and press the Enter key. SYSTAT's commands can be categorized into four broad categories: general commands, data related commands, graph related commands, and statistical commands. The statistical commands are in turn grouped by module. While the other commands are available for use at any time, the statistical commands will only function after you enter or, in other words, "load" the relevant module. The modules are as follows:

ANOVA	BAYESIAN	BETACORR	BLOGIT
CFA	CLOGIT	CLUSTER	CONJOINT
CORAN	CORR	DCLOGIT	DESIGN
DISCRIM	EXACT	FACTOR	FITDIST
GAUGE	GLM	IIDMC	LOGLIN
MANOVA	MCMC	MDS	MISSING
MIX	MIXED	MLOGIT	MSIGMA
NETWORK	NONLIN	NPAR	PERMAP
PLS	POLY	POSAC	POWER
PROBIT	QC	RAMONA	RANDSAMP
RANKREG	RDISCRIM	REGRESS	RIDGEREG
ROBREG	RSM	SAVING	SERIES
SETCOR	SIGNAL	SMOOTH	SPATIAL
SURVIVAL	TESTAT	TESTING	TLOSS
TREES	TSLS	VC	XTAB

Note:

1. There are three other modules in SYSTAT that are not listed above, viz. BASIC, MATRIX and STATS. Commands related to these modules will work directly without having to load the modules. In other words, they function just like the general commands.

2. Some of these modules are available only as add-ons.

- To enter a module, type its name after the prompt, and press the Enter key. For example, type:

```
XTAB
```

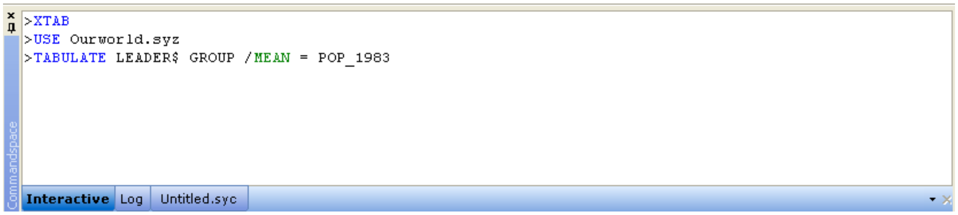
- Next, identify which data to use. For example, type

```
USE ourworld
```

and press the Enter key.

- Now type a command line:

```
TABULATE leader$ group/MEAN= pop_1983
```



■ Press the Enter key to obtain output.

To create graphs, type the desired graph command followed by the variables to use. Specify optional settings to customize the resulting display. Valid graph commands include:

BAR	CONE	CYLINDER	DENSITY	DOUGHNUT
DOT	DRAW	FOURIER	FPLOT	ICON
LINE	MAP	PARALLEL	PIE	PLOT
PPLOT	PROFILE	PYRAMID	QPLOT	SPLOM
WRITE				

Note:

SYSTAT can use one of two modes for drawing graphs. One is the DirectX mode, and the other is the classic mode. The options CONE, CYLINDER, DOUGHNUT are available only in the DirectX mode. By default, SYSTAT uses the classic mode. You can run RENDER DIRECTX to switch to the DirectX mode.

Refer the *Language Reference* volume for details regarding general and data related commands.

Command Syntax

Most SYSTAT commands have three parts: a command, an argument(s), and options.

```
command argument / options
```

Each module name or command must start on a new line. A command must be separated from its argument by a space (the equal sign is not allowed except in a few specific cases) and options must be separated from commands by a slash (/). For example:

```
CSTATISTICS urban babymort / MEAN SEM MEDIAN
```

- The command specifies the task--in this case, to display statistics.
- The arguments are the names of the variables, *URBAN* and *BABYMORT*, for which statistics will be computed.
- The options (following the slash) specify which statistics you want to see. If you do not specify any options, SYSTAT displays a default set of statistics.

In general, the argument may be one or more variables, numbers or strings separated by a space or comma, variable lists separated by the asterisk (*), file names, folder names, a specific keyword that may or may not be equated to a number, an expression, an equation or an inequality. Each option is a keyword that may or may not be equated to an option value (the equal sign is compulsory). The option value has the same possibilities as the argument.

Hot versus Cold Commands

Some commands execute a task immediately, while others do not. We call these *hot* and *cold* commands, respectively.

Hot commands. These commands initiate immediate action. For example, if you type LIST and press the Enter key, SYSTAT lists cases for all variables in the current data file.

Cold commands. These commands set formats or specify conditions. For example, PAGE WIDE specifies the format for subsequent output, but output is not actually produced until you issue further commands. Similarly, the SAVE command in modules specifies the file to save results and data to, but does not in itself trigger the saving of results; the next HOT command does that.

Command Syntax Rules

Upper or lower case. Commands are not case sensitive. You can type commands in upper or lower case or both:

```
CSTATISTICS or cstatistics or CStatistics
```

The only time SYSTAT distinguishes between upper and lower case is in the values of string variables. In other words, for a variable named *SEX$*, SYSTAT considers the text values “male” and “MALE” to be different.

Abbreviating commands. You can shorten commands and options to the first two to seven letters, as long as the resulting abbreviation is unique and the largest expansion sounds “nice” (commonly used). For e.g., COV, COVA, and COVAR, will all be permissible abbreviations of COVARIANCE. For commands, abbreviations till the full word (even beyond 7 characters) will be supported. For example:

- CSTATISTICS can be shortened as CSTA or CST.
- DENSITY *var* can be shortened as DEN *var*.
- HELP *phrase* can be shortened as HE *phrase*.

In the case of commands within a module, the abbreviation needs to be unique within the module. For example, STAR, STAN, STE and STO will be interpreted as START, STANDARDIZE, STEP and STOP respectively within the GLM module. Outside GLM, STAN will be treated as STANDARDIZE - the command to standardize variables.

Note: BASIC commands, module and variable names must be typed in full; they cannot be abbreviated.

Interpreting common commands. Some commands like STANDARDIZE perform different functions within and outside modules. Such commands will be interpreted based on a certain priority order: BASIC commands, commands related to the module currently loaded if any, and then the rest of the commands. If you want to use a global command - a command that is globally available irrespective of the module loaded - when a module is loaded, then you have to issue EXIT to exit from the module.

Retrieving commands. SYSTAT holds the most recently processed command lines in memory. From the Interactive tab of the Commandspace, use the Up arrow or F9 key to scroll through the commands. Press Up arrow or F9 once to recall the previous command, press it again to see the command before that, and so on. To define the number and source of commands to retain in memory, set Command buffer options in the General tab of the Edit:Options dialog box.

Continuing long commands onto a second line. To continue a command onto another line, type a comma at the end of the line. For example, typing

```
CSTAT urban babymort pop_1990 / MEAN SEM MEDIAN
```

is the same as:

```
CSTAT urban babymort,
              pop_1990 / MEAN SEM,
                MEDIAN
```

Do not use a comma at the end of the last line of a command; this will cause SYSTAT to wait for the rest of the command. Also one word cannot be typed into two lines for example:

```
USE OUR,
WORLD
or
US,
E OURWORLD
```

are invalid shortcuts, whereas the following is a valid one:

```
USE,
OURWORLD
```

Commas and spaces. Except when used to continue a command from one line to the next, and in the case of functions, commas and spaces are interchangeable as delimiters. For example, the following are equivalent:

```
CSTAT urban babymort pop_1990

CSTAT urban, babymort, pop_1990

CSTAT urban,babymort, pop_1990
```

Quotation marks. You must put quotation marks around any character (string) data that belongs to a string variable, a string that needs to be case sensitive, or contains spaces.

For example, type:

```
NOTE ‘Statistical Analysis'
```

to display a note in the output in title case and on a single line. If your data file has a string variable for country names written in title case, the following command will select the case corresponding to Sweden:

```
SELECT country$ = ‘Sweden’
```

You can use either double (" ") or single (' ') quotes. If you are using a dialog box to generate commands involving strings, in general, you may not need to specify quotation marks.

In certain commands that involve values taken by string variables, if you do not use quotes around a value, SYSTAT looks for the value written in uppercase. For example, SPECIFY gov$[Democracy] + urban$[city] will be interpreted as SPECIFY gov$[DEMOCRACY] + urban$[CITY] whereas SPECIFY gov$['Democracy'] + urban$['city'] will be interpreted as SPECIFY gov$[Democracy] + urban$[city].

Furthermore, for any command involving filenames (such as USE and SAVE), filenames and file paths containing spaces require quotation marks around them.

Braces. If an option takes more than one value, then the option values should be enclosed in braces ({}). For example:

```
CSTATISTICS urban babymort / MEAN SEM MEDIAN,
                             ROWS = {row(1), row(2), row(3)}
```

Specifying matrices. Some commands and options accept matrices as their arguments. Enclose the elements in brackets ([ ]), and indicate the end of rows by semicolons (except the last row). Each row may be written on a separate line. The following are two possibilities:

```
AMATRIX [1 0 0 -1; 0 1 0 -1; 0 0 1 -1]
```

Or

```
AMATRIX [1, 0, 0 -1;
         0 1 0 -1;
         0 0 1 -1 ]
```

SYSTAT functions. A typical SYSTAT function has the syntax FUN(par1, par2, …) where *par1*, *par2*, … are the parameters of the function *FUN*. When the number of parameters is more than one, the parameters have to be separated by commas (a space cannot be used as a delimiter). The parameters are optional for many functions (default values will be used) in which case the function has to be written as *FUN()*. For instance, *ZRN()* will generate random numbers from the standard normal distribution.

Unit of measurement. Certain commands and options related to graphs allow you to specify the unit of measurement. The available units of measurement are inches, centimeters and points that can be indicated using the keywords IN, CM and PT respectively. When used in the arguments of commands, you should separate the number from the unit by a space. For example, *DEPTH 2 CM* sets the depth of a graph to 2 centimeters. In the case of option values, a number can be suffixed by the unit of measurement with or without a space. For example, the option *HEIGHT = 200PT* sets the height of a graph to 200 points.

Reserved Keywords. The following commands from the BASIC module are reserved keywords in SYSTAT. You cannot use these words as variable names.

```
LET, FOR, IF, THEN, ELSE, ARRAY, DIM, PRINT
```

Barring these keywords, you may name a file variable, matrix, array, or user-defined function by any string that you so desire. However, SYSTAT may encounter some name conflicts in certain commands. In order to resolve such conflicts, we will use a precedence rule.

Precedence. The SYSTAT namespace, which consists of all its possible module names, commands, arguments, options and option values, has the following precedence structure (highest to lowest):

- **Class 0.** SYSTAT module names, commands, options, and option values where such values are fixed keywords
- **Class 1.** Built-in function names
- **Class 2.** User-defined function, matrix, and array variable names
- **Class 3.** File variable names (in the currently active data file)

When SYSTAT encounters a potential conflict in a command line, it will use the precedence rule to resolve the conflict. Depending on the context, a name will be treated as coming from the lowest-numbered class possible. For example, consider the following commands used to draw a bar chart of the INCOME variable in the SURVEY2 data file:

```
USE SURVEY2

BAR INCOME / COLOR = BLUE
```

In general, the COLOR option accepts either a color name (like RED, BLUE, YELLOW, and so on) or a variable name as option value. Incidentally, BLUE is also a variable in the data file. As color names belong to Class 0 in the above precedence rule, whereas file variable names belong to Class 3, SYSTAT interprets BLUE as the color name. If you need to set COLOR to the variable name BLUE, rename the variable and then use it as the option value of COLOR. The command script to do this is as follows:

```
USE SURVEY2

LET BLUE2 = BLUE

BAR INCOME / COLOR = BLUE2
```

Shortcuts

There are some shortcuts you can use when typing commands.

Listing consecutive variables. When you want to specify more than two variables that are consecutive in the data file, you can type the first and last variable and separate them with two periods (..) instead of typing the entire list. This shortcut will be referred to as the ellipsis. For example, instead of typing

```
CSTAT babymort life_exp gnp_82 gnp_86 gdp_cap
```

you can type:

```
CSTAT babymort .. gdp_cap
```

You can type combinations of variable names and lists of consecutive variables using the ellipsis.

Multiple transformations: the @ sign. When you want to perform the same transformation on several variables, you can use the @ sign instead of typing a separate line for each transformation. For example,

```
LET gdp_cap = L10(gdp_cap)

LET mil = L10(mil)

LET gnp_86 = L10(gnp_86)
```

is the same as:

```
LET (gdp_cap, mil, gnp_86) = L10(@)
```

The @ sign acts as a placeholder for the variable names. The variable names must be separated by commas and enclosed within parentheses ( ).

Autocomplete commands

As you begin typing commands in the Interactive or batch (Untitled) tab of the Commandspace, you will be prompted with the possible command keywords, available data files, or available variables. When a letter is typed, all commands beginning with that letter will appear in a dropdown list. Select the desired command or continue typing. For a command involving file names, on pressing space and then any letter, the files of the relevant folder as specified in the File Locations tab of the Edit: Options dialog box, beginning with that letter, will be listed. For a command involving variable names, if a data file is open, all available variable names, beginning with that letter, will appear in a drop down list. When you type expressions, the relevant function names will be shown. In general, for any given letter that you type, the relevant arguments, options, and option values will be listed.

If you do not know the exact syntax of a particular command, press Ctrl + Spacebar to get a list of all available commands/arguments/options/option values. Press the Esc key to close the drop down list.

Command autocompletion is enabled by default. You can turn it off by unchecking Autocomplete commands in the General tab of the Edit: Options dialog box, or by clicking on AUTO in the Status Bar..

Command Coloring

The commands, variable names, numbers, strings and comments (REM statements) that you type will be colored in distinguishing colors. The colors are as follows:

`Commands`	Blue
`Command options, comments`	Green
`Arguments,option values`	Purple
`File, variable names`	Black
`Numbers, strings in quotes`	Pink

Coloring makes it easy for you to identify the various components of a command line thereby reducing the risk of making syntax errors. Command coloring is enabled by default. You can turn it off by unchecking Color command keywords in the General tab of Edit: Options dialog box.

Online Help for Commands

SYSTAT's online help system provides easy access to information about SYSTAT commands. At the command prompt, type HELP followed by the name of a module or command for which you want help.

For example, you can access help on the CORR module by typing:

```
HELP CORR
```

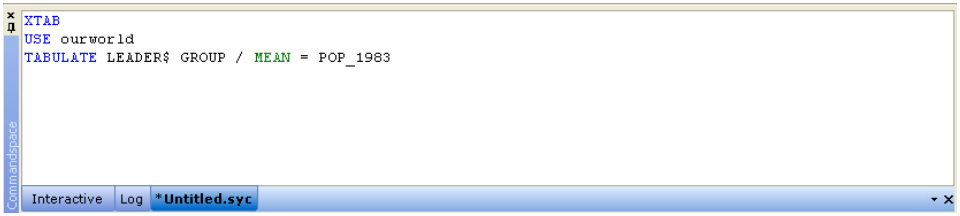
If you are already in the CORR module, you can type just HELP to get a list of commands available within CORR, HELP followed by the name of a command that you know belongs to the CORR module (for example, HELP PEARSON) or HELP followed by the name of any other module or global command (for example, HELP CLUSTER).

You can also start help by choosing Index from the Help menu and selecting the desired command from the list. Yet another alternative is to type the command in any tab of the Commandspace, and either clicking on it and pressing Ctrl+F1, or right-clicking on it and selecting the HELP command.

Command Files

A command file is a text file, in Unicode or ANSI format, that contains SYSTAT commands. Saving your analyses in a command file allows you to repeat them at a later date.

You can create a command file by selecting the batch (Untitled) tab in the Commandspace. This tab corresponds to a simple text editor; type the desired commands line by line. When you are done, save the commands to a file or submit them to SYSTAT for processing. In contrast to the Interactive tab, no interactive prompt (>) appears on the batch tab; commands are not processed until the resulting command file is submitted to SYSTAT.




```
XTAB
USE OURWORLD
TABULATE leader$ group/MEAN = pop_1983
```

If you find any of the SYSTAT examples relevant to your analysis, you can open this example command file in the SYSTAT Command folder, edit it to suit your data and save it under a different filename. You can in fact simultaneously create or open any number of command files, copy/paste among them, edit any of them, and submit any of them.

To create a new command file

- From the menus, choose

File
  New
    Command...

Or

click in the Commandspace and press the New toolbar button on the Standard toolbar.

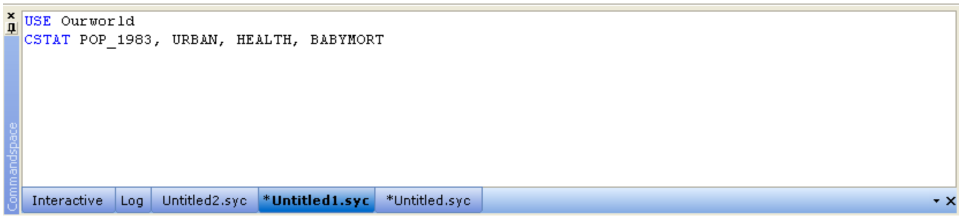
Or

double-click on the empty space beside the last tab in the Commandspace.

Or

right-click on a batch (Untitled) tab and select New.

- Type SYSTAT commands in the batch (Untitled) tab. For more information on SYSTAT commands, see SYSTAT *Language reference*.




```
USE OURWORLD
CSTAT pop_1983, urban, health, babymort
```

- To save the command file, click the corresponding tab and, from the menus choose:

File
  Save Active File...

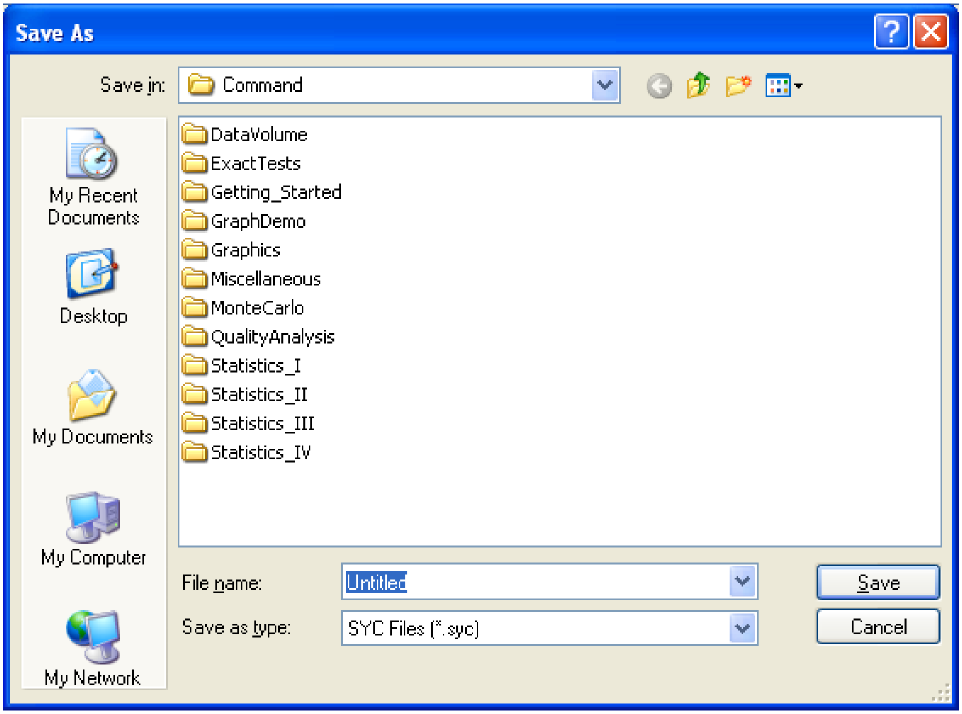
Or

  Save As...

Or

right-click on the corresponding tab and click Save.

- In the Save in field, select the appropriate drive or folder to save to.
- Type a suitable filename or select an existing file from the list if you want to overwrite.
- The default format is unicode. If you want to save the command file in ANSI format, select "SYC Files (ANSI) (*.syc)" in the Files of type field. Select "All files" if you want to use a different file extension.
- Press Save.



Note:

- To save a file under a different name, click Save As… from the File menu and specify the desired filename and path.
- To change the default command file format, check ANSI under Default command file format in the General tab of the Edit: Options dialog box.
- To save all unsaved files, click from the File menu Save All and specify appropriate filenames for each.
- Instead of typing commands, you can perform the corresponding actions through menus and dialogs, and select Save or Save As with the Log tab active.
- The commands that you type line-by-line in the Interactive tab can also be saved to a command file, by selecting Save or Save As with the Interactive tab active.

To open a command file

- From the menus, choose

File
  Open
    Command...

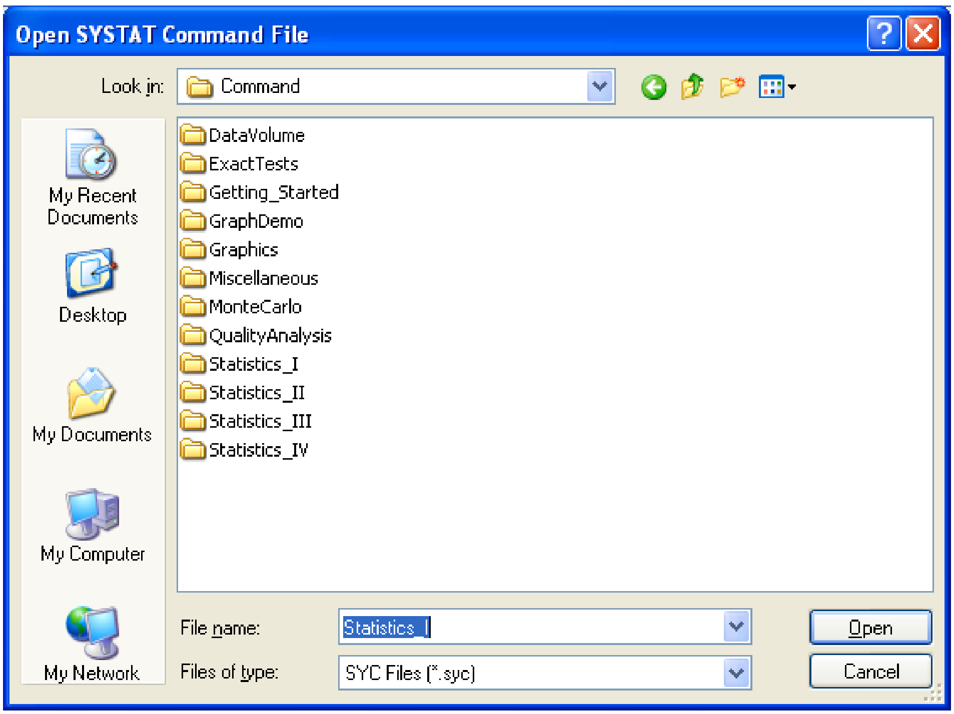
Or

click the batch (Untitled) tab and press the Open toolbar button on the Standard toolbar.

Or

right-click on any batch (Untitled) tab and click Open.

- In the Look in field, click the drive or folder that contains the command file you want to open.
- Double-click the folder that contains the command file you want to open.
- Click the command file name from the list that is displayed, and press Open.



Note:

- If you do not see the command file you are looking for, you can choose a different file type in the Files of type field.
- You can also open a command file you used recently by clicking its name in the Recent Files quadrant of the Startpage or in the Recent Command item of the File menu.

Working with Text

- To undo your last action, from the menus, choose

Edit
 Undo…

Or

`press Ctrl+Z.`

Or

`press Alt + Backspace.`

- To cancel your last undo action, from the menus, choose

Edit
 Redo…

Or

`press Ctrl + Y.`

- To search for text, from the menus, choose

Edit
 Find

Or,

`press Ctrl+F.`

In the Find what field, enter the text you want to search for, and then press Find Next.

To find additional instances of the same text, continue to press Find Next.

Or, from the menus, choose

Edit
 Find Next

Or

```
press F3.
```

You can search for whole words alone, do a case-sensitive search, or search backwards.

- To replace a text, from the menus, choose

Edit
 Replace

Or

```
press Ctrl + H.
```

Find the desired text and press Replace or Replace All as desired.

Printing Command Files

Currently, the facility to print command files is not available in SYSTAT. Open the command file in an alternative command editor, like Notepad, and use the Print option therein to print the command file.

Submitting Command Files

When you submit a command file, SYSTAT executes the commands as if they were typed line by line at the command prompt. For example, suppose you have a text file of SYSTAT commands named *TUTORIAL.SYC*. You can execute the commands in the file in eight different ways:

- Issue a SUBMIT command from any SYSTAT procedure:

```
SUBMIT tutorial
```

Note: Unless the command file is in the default directory, for commands in the File Locations tab of the Edit: Options dialog box, you have to define the path for the file. For information on Global Options, see Chapter 7*, Customization of the SYSTAT Environment*.

- In the SYSTAT window, from the menus choose:

File
  Submit
    File...

- Open the command file in the batch (Untitled) tab in the Commandspace using the File or context menu. From the Submit sub-menu of the File menu, you can then submit the entire file (Window), submit from the cursor's location till the end of the file (From Current Line to End), or submit just the current line (Current Line).
- From the menus choose:

Utilities
  User Menu
    Menu List...

and click on the item from the list. For information on creating menu items in the User Menu, refer Chapter 7, *Customization of the SYSTAT Environment*.

- Double-click the file after navigating to its location in the hard disk through Windows Explorer. The file opens in a new instance of the SYSTAT application. Right-click in the batch (Untitled) tab of the Commandspace and submit the file.
- Use the DOS command syntax to (open or) submit the file. The details of this method are explained later in this chapter.
- Create a link to the command file in the Examples tab of the Workspace using the Add Examples dialog box that opens on clicking Add Examples under the Utilities menu. Double-click the link, or right-click and select Run to execute the command file as it is. You can even use the context menu to open the command file in the batch tab, edit it and then execute it. Refer Chapter 7, *Customization of the SYSTAT Environment* to know more about adding examples.
- Open the command file in any external application like Notepad, copy some or all commands, right-click anywhere in the Commandspace, and select Submit Clipboard.

To submit a range of commands, select the commands and choose Submit Selection from the context menu. If the range includes the last command in the tab, use Submit From Current Line to End. If you choose either Submit Window or Submit From Current Line to End, SYSTAT prompts you to specify whether to submit the range or not.

Alternative Command Editors

Command files are ASCII text files having an SYC filename extension and containing command syntax. Hence, you can use any text editor to create command files. In your editor, type each command on a new line and save the resulting file as ASCII text. We recommend using the SYC extension when saving these files. Although any text file containing commands can be processed, using an SYC extension for these files allows maximal Windows functionality, such as double-clicking a file to automatically open it. In addition, you can use a text editor in conjunction with the Windows Clipboard to submit syntax for processing without creating command files or using the Commandspace. After typing the commands in your editor, select and copy them. In the processing environment, select Submit Clipboard from the File menu or the context menu of the Interactive/batch (Untitled) tabs of the Commandspace. The software processes the commands without changing any text in the Interactive or batch (Untitled) tabs of the Commandspace.

Using a text editor for command entry allows you to hide the Commandspace, creating more area in which to display the output. To hide/unhide, collapse or resize the Commandspace, see Commandspace Customization (cross refrerence) in Chapter 7: *Customization of the SYSTAT Environment*. As you change between the editing and processing environments, the currently active application appears in front of the other. Consequently, you can maximize the area for both the input and the output, switching between the two by toggling between the applications. You can also have multiple command files open, submitting commands from each of them using the 'Copy/Submit Clipboard' procedure. However, the Clipboard only accesses the last copied item. Be sure the most recently copied text corresponds to the commands to be submitted.

Because the Commandspace itself is a text editor, you can also copy commands from any of the tabs for subsequent submission via the Clipboard. However, other submission methods (Submit Window, Submit from Current Line to End, Submit Current Line and Submit Selection) offer the same functionality without replacing the contents of the Clipboard. Moreover, the command prompt (>) prevents successful submission of two or more command lines copied from the Interactive tab.

Comments in Command Files

The !! or REM command can be used for inserting comments in command files and for making a command inactive during the current run. All text following !! or REM on the same line is ignored.

```
REM Now we merge files side-by-side
REM MERGE file1 file2
MERGE file1 file3
```

The text following the first REM command remains in the command file. The MERGE statement in the second line is not invoked.

The !! command can also be used at the end of another command line. You can use this to append comments to a command line. The comments could indicate what the command line does, why it was written, which step of a procedure it is, or even the name of the person who has written it.

Tip: To add comments that appear in your output, use the NOTE command.

Commands to Control Output

SYSTAT provides a number of commands to save and print output, as well as to control its appearance. These commands may be particularly useful when creating command files.

OUTPUT command. Enables you to route subsequent plain text output to a file or a printer.

PAGE command. Enables you to specify a narrow (80 columns, the default) or wide format (132 columns) for output. You can also specify a title that appears at the top of each printed output page.

FORMAT command. Enables you to specify the number of character spaces per field displayed in data listings and matrix layouts, and the number of digits printed to the right of the decimal point. You can also display very small numbers in exponential notation (instead of being rounded to 0).

NOTE command. Enables you to add comments to your output. For example:

```
NOTE "THIS IS A COMMENT.",
     "This is the second line of comments."
     "It's the 'third' line here!"
```

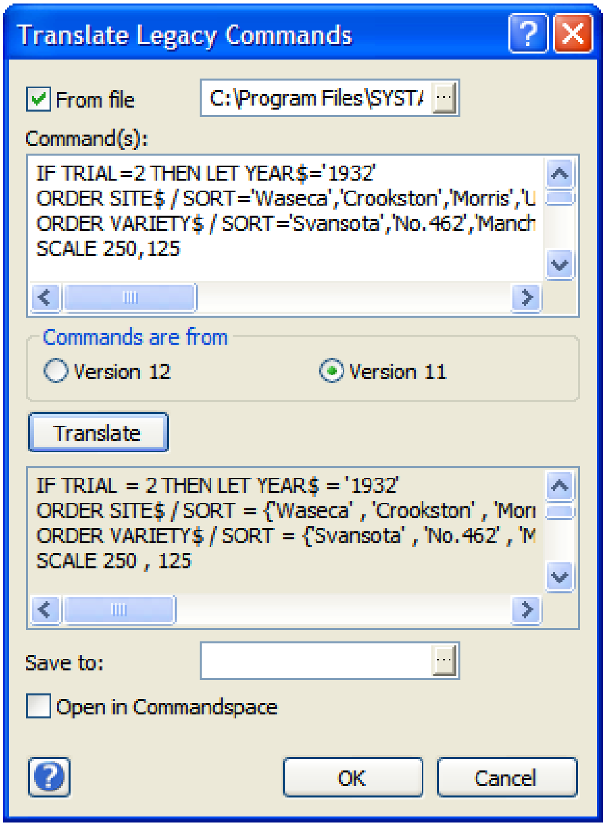
Each character string enclosed in either single or double quotation marks is printed on a separate line. A note can span any number of lines, and , and can contain ASCII codes to display the corresponding ASCII characters.

Translating Legacy Commands

SYSTAT provides a feature whereby you can translate legacy command files to the current command syntax supported by SYSTAT 13. You can either translate commands that are in a file, or directly type the commands to be translated. To translate legacy command files, from the menus choose:

Utilities
  Command
    Translate Legacy Command Files…

Alternatively, you can right-click in an untitled tab of the Commandspace, and select Translate Legacy Command Files. To translate just some selected commands, select the commands in the untitled tab, right-click on the selection, and then click Translate Legacy Command Files.



From file. Specify a file to read the legacy commands from. The contents of the file are displayed in the box below.

Command(s). You can type the legacy commands that you want to translate, in this box. If you have chosen a file to translate from, you can edit the contents shown in the box before you request a translation.

Commands are from. Select the version of your commands/command file.

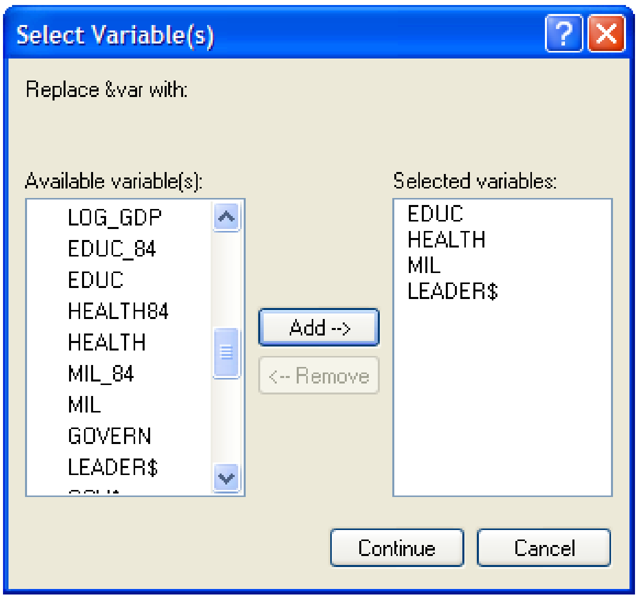
Translate. Press Translate to translate the commands. The translated commands are displayed in the box below. You can select and copy a part or the whole of the translated commands for pasting to the desired location.

Save to. You should save the translated commands to a SYSTAT command file.

Open in Commandspace. You can request that the translated commands be opened in an untitled tab in the Commandspace.

SYSTAT Command Translator

In addition to the Translate Legacy Command Files dialog box, SYSTAT provides a SYSTAT Command Translator application. To access it, click Command Translator in the SYSTAT 13 program group of the Windows Start Menu.



Add Files. Add any number of files that you need to translate to SYSTAT 13 syntax. In the File: Open dialog box, you will be able to click and drag the mouse, or use the Ctrl/Shift keys to select multiple files simultaneously. When you press Open, the files will get listed in the box beneath. Click on any command file to view its content in the box beneath.

Translate from. The following options are available. Check one of the following:

- Version 12 to Version 13
- Version 11 to Version 13
- Version 11 to Version 12

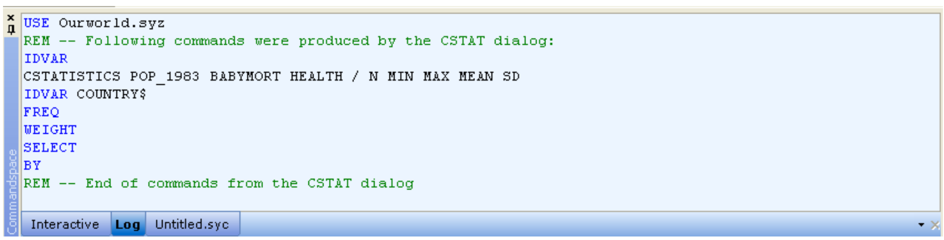
Translate. When you press this button, all the selected files will be translated so as to be suitable for execution in the specified version of SYSTAT.

Save. Specify the folder to save the translated command files to. They will be saved with an "_Trans" suffixed to the original filenames. Check Retain original filename(s) to avoid the suffix.

Close. Closes the application.

Command Log

SYSTAT records the commands you specify during your current session in a temporary file called the command log. Select the Log tab in the Commandspace to view the command log. You can view, copy, submit, and save all of the commands stored in the command log at any time during a session. However, because the log serves as a command recorder, you cannot edit commands using the Log tab.



After selecting the Log tab, you can submit commands directly from the command log in four ways:

- Submit the entire log by choosing Submit Window from the File or context menus.
- Submit the most recently processed commands by moving the cursor to the desired starting point and choosing Submit From Current Line to End from the File or context menus.
- Submit a subset of commands by selecting the desired commands and choosing Submit Selection from the context menus.
- Submit the desired line by moving the cursor to the line and choosing Submit Current Line from the context menus.

To modify commands before submission, copy the log contents, paste the copied portion to the batch (Untitled) tab or Interactive tab, edit the pasted commands, and submit the resulting syntax.

Recording Scripts

SYSTAT provides you an option to reuse a part or whole of the log file of the current session. To start/stop recording the scripts:

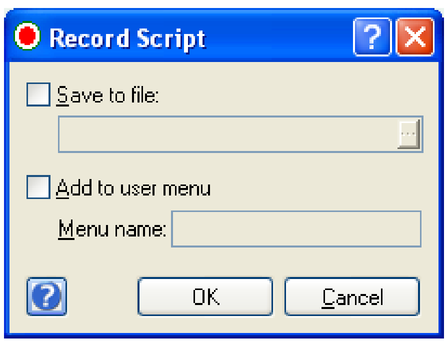
- From the menus choose:

Utilities
  Start/Stop Recording...

or

- Click on the Record Script tool provided in the Standard toolbar:

The Record Script dialog pops up when you stop the recording.



You can save the recorded script to a file and/or you can add it to the User Menu for use in subsequent sessions. For more information on the User Menu, see Chapter 7, *Customization of the SYSTAT Environment*. Quit the dialog by pressing Cancel if you do not want to save the recorded script.

There is also another way to reuse the recorded commands:

- From the menus choose:

Utilities
  Macro
    Play Recording...

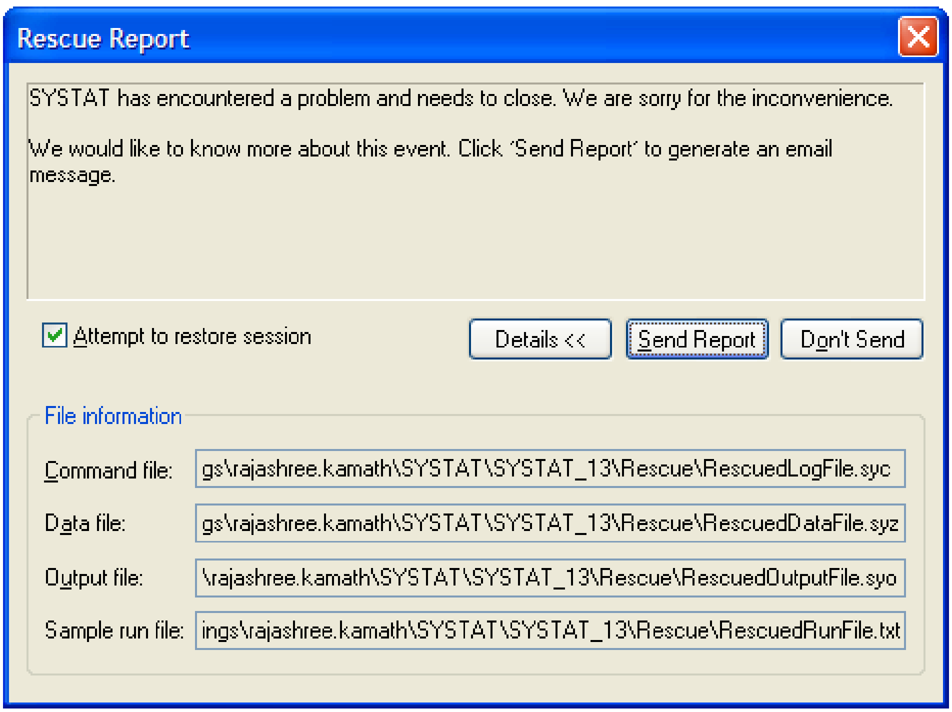
Or

- click on the Play Recording tool button. 

Note: The Play Recording option can only play the latest recording. So, a recording will be lost if you start recording another set of commands without saving it.

Rescuing Sessions

The command log records only the commands from your current session. You cannot use the command log to recover commands from a previous session unless you saved those commands in a command file before exiting SYSTAT. However, in the unfortunate event of a crash, SYSTAT attempts to recover the log, output, and data files of the session. These files are saved to the Rescue sub-folder within the SYSTAT user folder. Before closing, the Rescue Report dialog pops up:



Attempt to restore session. Opens the recovered files, if any, on restarting SYSTAT. You will be prompted to save the recovered data files.

Details. Displays the filename and location of the recovered files.

Send Report. Generates an email message with the recovered files attached.

Don't Send. Terminates the current session without generating the email message.

Working with DOS Commands

Some of the tasks that SYSTAT is capable of can be performed with minimum user intervention. For instance, there may be very large command files you want to execute, or command files that require a long time to produce output, or command files that produce a large number of graphs all of which you want to save. It is indeed possible to do all this and much more in the Windows environment. In fact, you can work with SYSTAT command files even without having to open the SYSTAT application manually. All you need to do is to invoke the MS-DOS Prompt from the Windows Start Menu, or the Windows Run dialog and type the following command line with appropriate command switches:

```
"filepath1\App\systat.exe" /switch(es) "filepath2\filename.xxx"
```

where *filepath1* is the SYSTAT installation folder path, *filepath2* is the location of the file on which SYSTAT will operate. (The quotes are required only if there are gaps in the file path or filename.) Depending on the switch(es) and .xxx you give, the tasks described below can be automated:

Switch	.xxx	Description	Example command
/x	.syc or .cmd	Opens SYSTAT and submits filename.syc	Systat /x c:\data\name1.syc
/c	.syc or .cmd	Opens SYSTAT and loads filename.xxx onto the Untitled tab of the Command-space	Systat /c "c:\my data\name2.cmd"
/e /x	.syc or .cmd	Opens SYSTAT, submits filename.xxx, and exits the application if file-not-found errors are encountered.	Systat /e /x c:\data\name3.syc
/gscgm	.cgm	Opens SYSTAT, executes any commands the user may give, and on exit, automatically saves (in CGM format) all graphs in the Output Editor.	Systat /gscgm "c:\graphs\my graph.cgm"

/elog	.dat	Opens SYSTAT, and stores all error messages encountered during command execution, into filename.xxx.	systat /elog c:\data\prompt\ErrorLog.dat
/gexit /x	.syc	Opens SYSTAT, submits filename.xxx, and exits the application if no graph is generated on running it.	Systat /gexit /x c:\data\prompt\name4.syc
/m	.xxx	Opens SYSTAT with its window minimized; you can include other keys with this.	Systat /m /x c:\data\name5.syc
/out	.dat	Opens SYSTAT, executes any commands the user may give, and on exit, saves all the text output generated during the session into filename.xxx.	systat /out c:\data\prompt\testN.dat
/x /mht	.mht	Opens SYSTAT, executes the command file given with /x, and saves the output in the MHT format to filename.mht.	systat /x c:\data\prompt\name6.syc /mht c:\data\prompt\outfile6.mht
/q	.xxx	Opens SYSTAT, performs the actions stipulated by any other switches specified, and quits SYSTAT.	systat /x c:\data\prompt\name6.syc /mht c:\data\prompt\outfile6.mht /q systat /x c:\data\prompt\name7.syc /out c:\data\prompt\outfile6.txt /q

Note: In the command file you submit, any GSAVE, OSAVE, and EXPORT commands, will save the graph, output and data respectively, into a filename of your choice, which can be later used for further processing by SYSTAT or other programs, after this session of SYSTAT has quit.

Environment Variables

SYSTAT provides environment variables in the STATS module. These are variables that contain the computed values of various statistics for a given session, a given data file, and given variables. The following environment variables are available:

```
%<statistic>_<by group>_<variable name>%
```

where <statistic> is as follows:

Name of Statistic	<statistic>
N	NU
Minimum	MI
Maximum	MA
Sum	SU
Median	MD
Arithmetic Mean	ME
Standard Deviation	SD
Variance	VA
Shapiro Wilk Statistic	WS
Shapiro Wilk p-value	WP
Cleveland Percentile for # percentile	PTILE1_#

Following this naming convention, the environment variable name for the 66th Cleveland percentile for the 3rd BY group for a variable VAR(32) would be:

```
%PTILE1_66_3_VAR(32)%
```

Example: Computing Mean Using Environmental Variables

Sometimes, the data that we need to analyze is not available in a single file but scattered across different files, say, in different locations. One approach to analyze such data is to append all the files and do the analysis. In this example, we illustrate an alternative approach whereby basic statistics are computed for the individual data files and the final statistic is computed using environmental variables. We generate a random sample of size 200,000 from the normal distribution, split it into two sub-samples, and compute the mean of the entire sample using the environment variables of the sub-samples.

The input is:

```
RANDSAMP
UNIVARIATE ZRN(5, 1) /SIZE = 200000 NSAMP = 1 RSEED = 100
DSAVE rannormal

SELECT CASE() <= 100000
EXTRACT rannormal1

USE rannormal
SELECT CASE() > 100000
EXTRACT rannormal2

USE rannormal1
CSTATISTICS S1/SUM N
TEMP sum~=%su_s1%
TEMP n~=%nu_s1%

USE rannormal2.syz
CSTATISTICS S1/SUM N
TEMP sum~=sum~ + %su_s1%
TEMP n~=n~ + %nu_s1%
TEMP mean~=sum~/n~
PRINT "The mean of the variable S1 is: ", mean~
```

```
The output is:

           ¦          S1
-----------+------------
N of Cases ¦     100,000
Sum        ¦ 499,915.318


SYSTAT created a temp variable "sum~".
SYSTAT created a temp variable "n~".

           ¦          S1
-----------+------------
N of Cases ¦     100,000
Sum        ¦ 500,062.493

SYSTAT created a temp variable "mean~".
The mean of the variable S1 is: 5.000
```

Command Templates

Command files provide a method for repeating analyses across SYSTAT sessions. Output produced by a particular command file will be identical to output produced by any subsequent runs of the same command file (assuming the data do not change). If, however, we change the data file in use or replace the variables used for a graph or statistical analysis, the results will vary from the original output but still retain the same structure. Command templates provide a method for achieving this customizability.

A command template provides a skeletal framework for graph creation, statistical analysis, or file management. The template has the appearance of a standard command file, but uses tokens in place of filenames, variables, numbers, or strings. Tokens serve as substitution markers; a value must be substituted for the token for command processing to continue. Every time you submit the command template, you can substitute a different value for each token.

For example, suppose we were to create a template for simple linear regression. This model requires a response variable and a predictor variable. We define the model with placeholders for these two variables. Substituting empirical variables for these placeholders yields regression output for that model. Either or both of these variables could be replaced to generate new output using the same general model for different data.

The ampersand character denotes tokens. The text immediately following an &' corresponds to a token name. Token names may contain any number of characters, numbers, underscores, and dollar signs, but the first character after the ampersand must be a letter or number. Dollar signs do not denote strings and may appear anywhere in the token name. As with variable names, token names are not case sensitive. The names *&tokn*, *&tOKn*, *&ToKn*, and *&TOKN* are equivalent; if all of these names appear in a template, substituting a value for one of them also substitutes that value for the others.

In some instances, ampersands should not be treated as token indicators. For example, the command

```
USE JUNE&JULY
```

accesses the data file *JUNE&JULY.* However, SYSTAT interprets the & as a token indicator and prompts the user for replacement text for *&JULY.* Two methods exist for avoiding this problematic behavior:

- If the command file does not involve any token substitution, turn token processing off by including the line TOKEN / OFF at the beginning of the command file or by using the General tab of the Global Options dialog. Use TOKEN / ON to reactivate token processing for subsequent command submissions.
- If some ampersands denote tokens but others do not, suppress token processing wherever needed by doubling the ampersand character. For example, replace *JUNE&JULY* with *JUNE&&JULY*. SYSTAT interprets two consecutive ampersands as a single character rather than a token indicator.

As SYSTAT processes commands, token substitution occurs either automatically or interactively. In automatic substitution, information supplied *in the template* replaces placeholders as they are encountered. Interactive substitution, on the other hand, involves *prompting the user* for placeholder replacement information. Command processing halts until valid information is supplied.

Automatic Token Substitution

Define tokens for automatic substitution by specifying:

```
TOKEN &tok = value
```

When SYSTAT encounters *&tok* during command submission, the defined value replaces the token automatically.

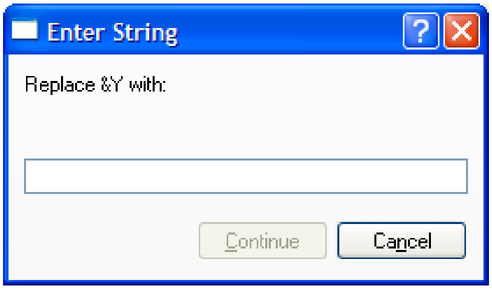
Quotes around token values are NOT included in the replacement value of the token. For example:

```
TOKEN &str1 = 'Depression'
LABEL dscore / 1 = &str1
BAR dscore / XLAB=&str1  TITLE='Bar graph of &str1'
```

defines the token *&str1* to have a value of *Depression*. In the bar graph, *Depression* appears entirely in capital letters for the tick label corresponding to 1 label but not for the title. Because the token value does not include the quotes, the value can be incorporated into other strings, as in this graph title. Without quotes, labels appear in upper case, as in this tick label. If quotes around the token are desired in the command file, explicitly include them in the command lines.

Interactive Token Substitution

To prompt the user for a token substitution value, precede the token text with an ampersand in the command file. During processing, when SYSTAT initially encounters the token, a dialog prompts for a replacement value.



Entering a value and pressing the Continue button allows processing to continue. Pressing the Cancel button halts further submission of the command file.

If subsequent commands use a token which has already been assigned a value, SYSTAT substitutes that value automatically. For example, the command:

```
PLOT &y*&x
```

results in dialog prompting for the tokens *&y* and *&x*. Suppose the current file has variables named *AGE* and *DEPRESS*. If we assign *DEPRESS* to *&y* and *AGE* to *&x*, the resulting graph plots depression score versus age. If the command file continues with:

```
REGRESS
   MODEL &Y = CONSTANT + &X
   ESTIMATE
```

SYSTAT computes the regression of depression score on age without prompting for substitution values.

Validating Input. The Token Substitution dialog accepts any value supplied by the user. However, commands typically require numbers, strings, or filenames to execute correctly. To impose restrictions on token replacement values, define tokens using the TOKEN command with the TYPE option, as follows:

```
TOKEN &tok1 / TYPE = tokentype
```

Valid tokentype values include: MESSAGE, OPEN, SAVE, VARIABLE, NVARIABLE, CVARIABLE, MULTIVAR, NMULTIVAR, CMULTIVAR, STRING, NUMBER, INTEGER and CHOICE.

During processing, when a token is encountered, SYSTAT scans for a definition. If SYSTAT finds an associated TOKEN definition, a dialog consistent with the token type appears. Otherwise, a default dialog prompts the user for information.

Resetting Tokens. Tokens can be reset individually or globally. To clear all tokens, use TOKEN without arguments or options. Any tokens used in subsequent command lines result in prompting for replacement values.

To reset an individual token, redefine the token using a new TOKEN command. For example,

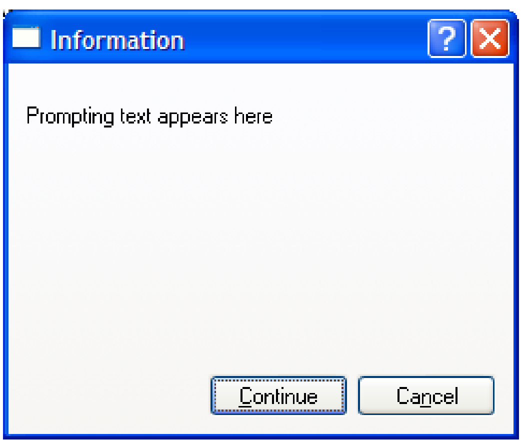
```
BAR &y*&x
TOKEN &x
DOT &y*&x
```

initially prompts for two token values. DOT, however, only prompts for a value for *&x*, the token reset between the BAR and DOT commands.

Message Tokens

In contrast to all other token types, message tokens do not function as substitution markers. Instead, the message token yields a dialog designed to provide the user with information about the template. To define a message token, include a command line having the following form in your command file:

```
TOKEN &msg/ TYPE=MESSAGE PROMPT="Prompting text appears here."
```



Common information to include in the prompting text includes:

- the result of running the template file.
- changes to the data file, if any.
- state of SYSTAT when template processing completes.

When command processing begins, SYSTAT immediately displays the prompting text for a message token in a dialog. Based on this text, the user can elect to continue or cancel processing. Pressing Cancel halts processing with no other commands in the template being executed.

If you exclude &msg in the above command, you will get a smaller message pop-up.

Filename Tokens

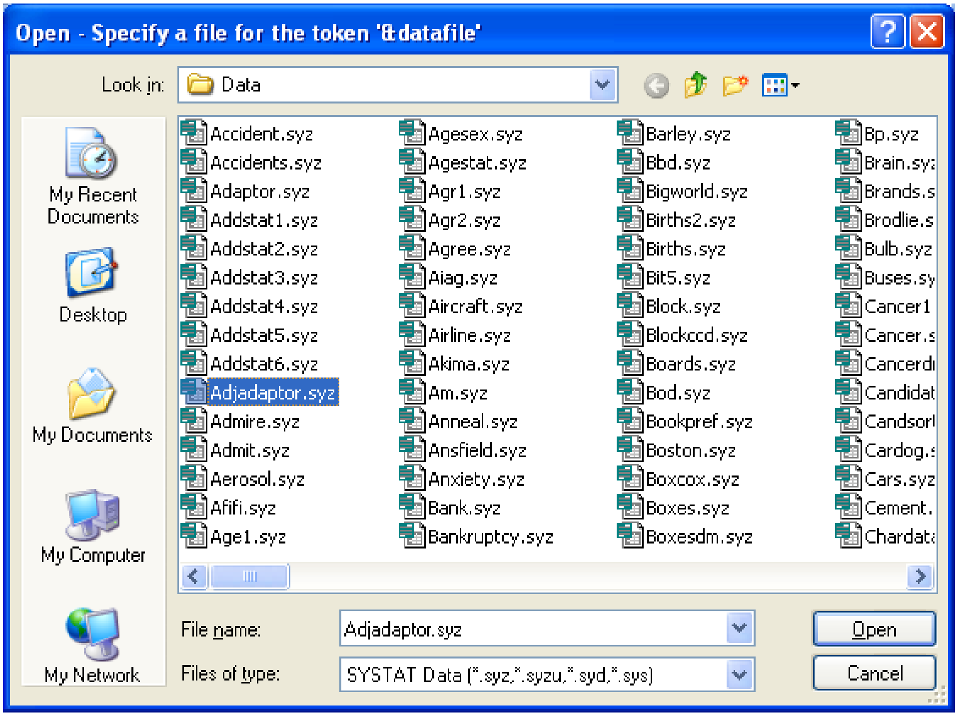
Filename tokens represent any file that SYSTAT can open or save, including data files, command files, and output files. To substitute a filename for a token, specify one of the following:

```
TOKEN &file / TYPE=OPEN
TOKEN &file / TYPE=SAVE
```

When SYSTAT encounters the token *&file* in the command file, a dialog prompting the user for a filename appears. SYSTAT substitutes the name of and path to the selected file for the corresponding token.

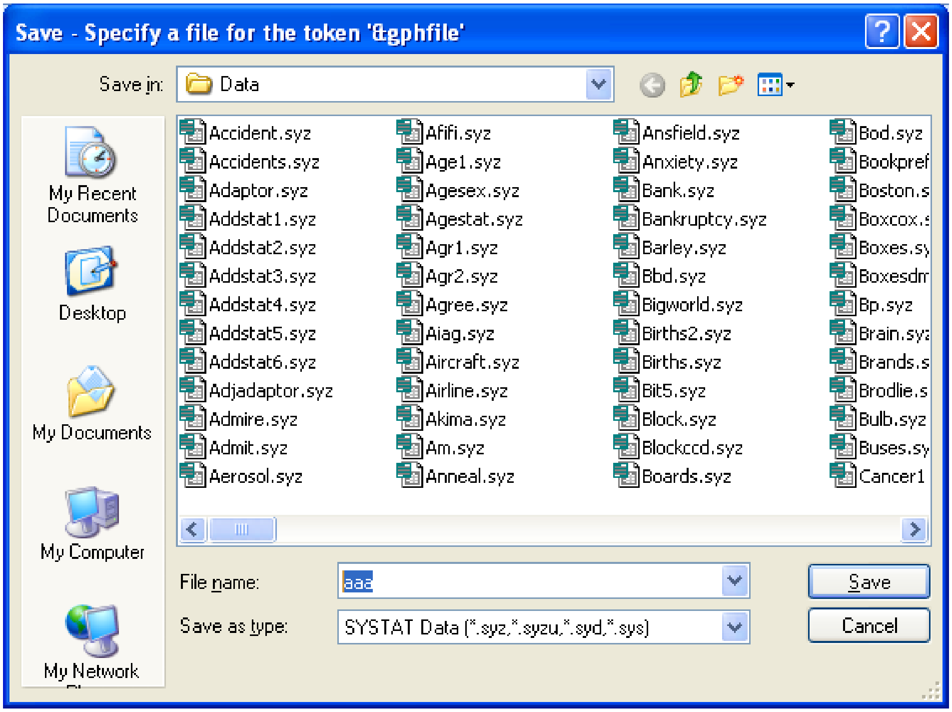
The OPEN type should be used when opening data files or for submitting command files. For example:

```
TOKEN &datafile / TYPE=OPEN
TOKEN &cmdfile / TYPE=OPEN
USE &datafile
SUBMIT &cmdfile
```



Use the SAVE type for saving output, data, or graphs to files. For example:

```
TOKEN &gphfile / TYPE=SAVE
TOKEN &outfile / TYPE=SAVE
PLOT Y*X
GSAVE &gphfile / BMP
OSAVE &outfile / HTML
```



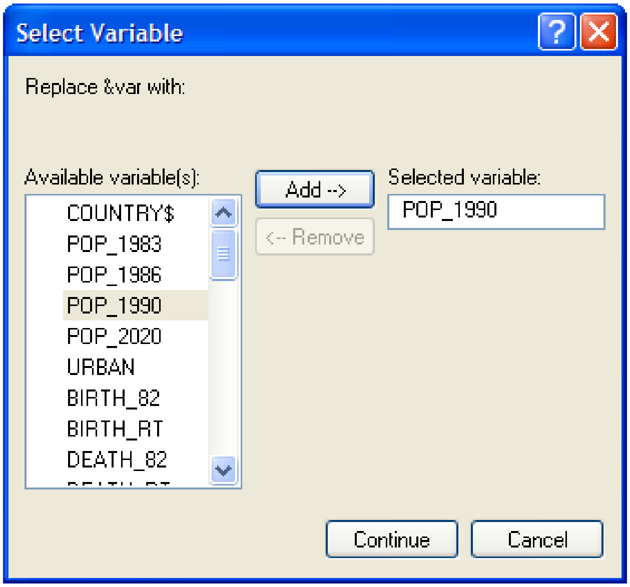
To add an instructional title to the dialog, use the PROMPT option. The specified prompt text appears in the title bar of the dialog. Ensure that the length of the text is limited to that of the title bar.

Single Variable Tokens

To substitute a single variable for a token, specify one of the following:

```
TOKEN &var / TYPE=VARIABLE
TOKEN &var / TYPE=CVARIABLE
TOKEN &var / TYPE=NVARIABLE
```

When SYSTAT encounters the token *&var* in the command file, a dialog prompting the user to select a variable appears. If no data file is currently open, SYSTAT prompts the user to open a file before proceeding to variable selection.



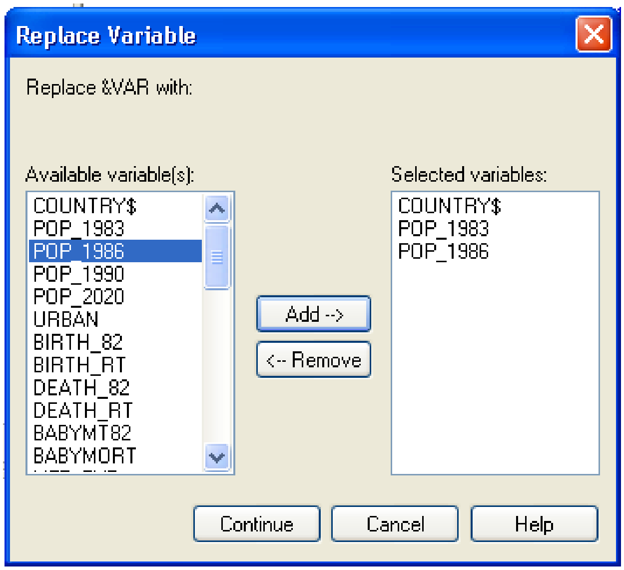
Select a variable and click Add. Click Continue to continue command processing. The list of available variables corresponds to the dialog type. The variable list contains only string variables if the token type equals CVARIABLE. The NVARIABLE type lists numeric variables for token substitution. To list all variables, use TYPE=VARIABLE.

Multiple Variable Tokens

To substitute multiple variables for a single token, specify one of the following:

```
TOKEN &var / TYPE=MULTIVAR
TOKEN &var / TYPE=CMULTIVAR
TOKEN &var / TYPE=NMULTIVAR
```

When SYSTAT encounters the token *&var* in the command file, a dialog prompting the user to select multiple variables appears. If no data file is currently open, SYSTAT prompts the user to open a file before proceeding to variable selection.



Select one or more variables and click Add to include the variable(s) in the token replacement set. To select multiple, consecutive variables, hold down the Shift key and click the first and last variables in the desired set. To select multiple, nonconsecutive variables, hold down the Ctrl key and click each variable before clicking Add. Click Continue to continue command processing.

The list of available variables corresponds to the dialog type. To list all variables, use TYPE=MULTIVAR. The variable list contains only string variables if the token type equals CMULTIVAR. The NMULTIVAR type lists numeric variables for token substitution.

By default, during multiple variable substitution, SYSTAT inserts a space between the selected variables. To specify an alternative character, use the SEPARATOR option of the TOKEN command.

```
TOKEN &var / TYPE=NMULTIVAR SEPARATOR='char'
```

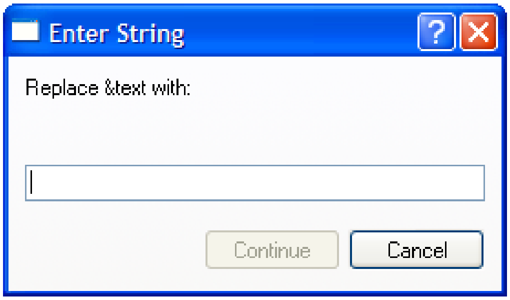
Replace *char* with the desired single character separator. SYSTAT truncates separators longer than one character to the first character. The designated character does not appear before the first variable or after the last variable.

String Tokens

To substitute a text string for a token, specify:

```
TOKEN &text / TYPE=STRING
```

When SYSTAT encounters the token *&text* in the command file, a dialog prompting the user for a string appears.



Type the desired text string. The entire string, including any quotes entered as part of the string, replaces the token. For instance, if a plot command contains a string token as an option:

```
PLOT Y*X / &text
```

you can enter a list of options such as

```
XLAB='X Variable' YLAB='Y Variable' SYMBOL=2
```

as replacement text for the token. Alternatively, to prompt for each option setting, assign each to a separate token:

```
NOTE 'Analysis of &str1 data'
```

```
NOTE &str1
```

Notice the tokens for the strings in the preceding command line. For the first note, quotes enclose the token. In this arrangement, the token replacement value should not include quotes, but should only contain the text used to label the axis. In contrast, for the second note, the token is not enclosed in quotes. The appearance of this note depends on whether the quotes are included in the token replacement value:

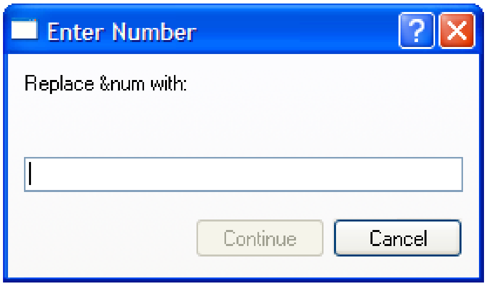
- Typing *Response* results in a note of *RESPONSE*. Without using quotes, SYSTAT displays labels in upper case.
- Typing '*Response*' results in a label of *Response*. Because the command line does not include quotes around the token for the second note, quotes must be included in the replacement value for the note to match the case of the supplied text string.

Numeric Tokens

To substitute a numeric value for a token, specify one of the following:

```
TOKEN &num / TYPE=NUMBER
TOKEN &num / TYPE=INTEGER
```

When SYSTAT encounters the token *&num* in the command file, a dialog prompting the user for a number or integer appears.



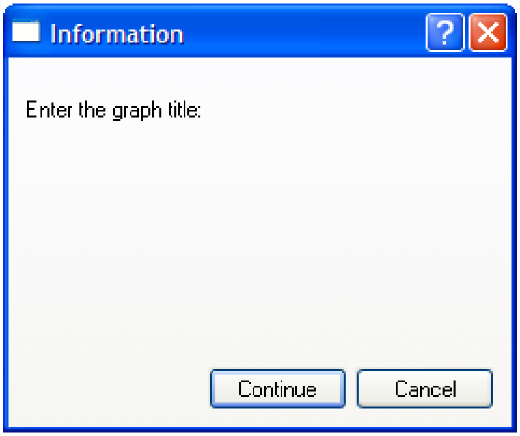
After entering a value, press Continue. If the value is not numeric, an error occurs and the user is prompted again. Likewise, attempts to input a decimal value for an integer result in re-prompting. The prompting dialog continues to appear until a valid value is entered or the Cancel button is pressed.

Custom Prompts

By default, the instruction appearing in substitution dialogs states "*Replace &tok with:*". To assist the user in entering valid information for a token, replace the default instruction with a custom prompt using the PROMPT option of the TOKEN command. For example, to prompt the user for a graph title, use

```
TOKEN &title1 / PROMPT='Enter the graph title:'
```

When SYSTAT encounters *&title1*, the following dialog appears:



Custom prompts can include carriage returns in the prompting text, allowing you to define the text appearing on each line of a multi-line prompt. For example:

```
TOKEN &var1/ TYPE=VARIABLE,
             PROMPT='This is the first line,
                     this is the second, and,
                     this is the third'
```

results in a three-line prompt. In the absence of carriage returns, SYSTAT automatically wraps prompting text to fit the dialog. Although the dialogs for string, number, and integer replacement have no practical limit on the number of lines that can be used as a prompt, the dialogs for variable selection limit custom prompts to three lines of text.

Choice Tokens

In contrast to all other token types except message tokens, choice tokens do not have a value. Instead, the choice token submits command files based on the choice given by the user. To define a choice token, specify:

```
TOKEN /TYPE=CHOICE "choice1"= "filename1.syc"  "choice2"=
"filename2.syc" … "choiceN"= "filenameN.syc"
```

Choice tokens are executed immediately. You may specify between 2 to 10 choices.

Dialog Sequences

Processing of command files begins at the first line of the file and continues to the last line. SYSTAT does not prompt for token replacement values until the token being defined is encountered in a command line, unless the IMMEDIATE option is specified. This can result in undesirable sequences of prompting dialogs. Consider the following set of commands:

```
TOKEN &xvar / TYPE=VARIABLE
TOKEN &xvarlabel / TYPE=STRING
TOKEN &yvar / TYPE=VARIABLE
TOKEN &yvarlabel / TYPE=STRING
PLOT &yvar*&xvar / YLAB=&yvarlabel XLAB=&xvarlabel
```

First, SYSTAT prompts for *&yvar*, the y-variable in the scatterplot. Next, a prompt for the x-variable appears. Prompting continues by asking for a label for the y-axis and finally for a label for the x-axis. Notice that the dialog sequence does not correspond to the order of the TOKEN statements, but instead corresponds to the ordering of the actual tokens in the PLOT command.

Rather than prompting in the order that the tokens are encountered, you can define a sequence for dialog prompting using the IMMEDIATE option. Instead of prompting when encountering the token, the prompting dialog appears when SYSTAT processes the TOKEN statement. For example, to prompt for the y-variable, the y-axis label, the x-variable, and the x-axis label, in that order, specify the following:

```
TOKEN &yvar / TYPE=VARIABLE IMMEDIATE
TOKEN &yvarlabel / TYPE=STRING IMMEDIATE
TOKEN &xvar / TYPE=VARIABLE IMMEDIATE
TOKEN &xvarlabel / TYPE=STRING IMMEDIATE
PLOT &yvar*&xvar / YLAB=&yvarlabel XLAB=&xvarlabel
```

In this case, SYSTAT prompts for information in the order of the TOKEN statements, rather than in the order that the tokens themselves appear.

Note: SYSTAT always processes MESSAGE tokens first; these tokens do not require the IMMEDIATE option.

Viewing Tokens

As you develop your own library of templates, it may become useful to have one template file submit another template file. However, if tokens have the same name in the two files, undesired output can result. To help correct any token 'conflicts', you can list all current tokens with their defining characteristics by specifying

```
TOKEN / LIST
```

You will get a list of predefined tokens, as well as user defined tokens. For each token, SYSTAT displays:

- the token
- the type
- the current assigned value
- text appearing in the prompting dialog

Generating this listing for each template identifies tokens common to both files. Differences should be examined closely; two tokens sharing a name but defined as different types are likely to yield odd behavior.

Predefined tokens

SYSTAT has default file locations for opening and saving files, which can be set through the File Locations tab of the Edit: Options dialog box or the FPATH command. When a command like USE filename or SUBMIT filename is executed without an explicit file path, SYSTAT looks for the file in the corresponding locations. The default file locations are assigned to built-in tokens as follows:

Token Name	**Token Value**
&EXPORT	Folder to which data will be exported
&GET	Folder containing ASCII data for import by BASIC
&GSAVE	Folder to which graphs will be saved
&IMPORT	Folder from which data will be imported
&OSAVE	Folder to which SYSTAT output will be saved
&OUTPUT	Folder to which ASCII output will be saved
&PUT	Folder to which ASCII data will be exported by BASIC
&ROOT	Folder to which SYSTAT is installed
&SAVE	Folder in which SYSTAT data files will be saved
&SUBMIT	Folder from which SYSTAT comand files will be submitted
&USE	Folder from which SYSTAT data files will be opened
&WORK	Folder to which temporary SYSTAT data files will be saved
&HTML	Folder to which HTML or MHT output will be saved
&TEMPDIR	Folder to which temporary files created by SYSTAT are saved
&RTF	Folder to which RTF output will be saved

Most of the built-in tokens are directly associated with the corresponding SYSTAT commands. You can use these appropriately in your command scripts so that files are opened from or saved to paths other than the assigned ones, without changing the default path. For example, the command:

```
SUBMIT &WORK\filename1.syc
```

submits filename1.syc from the path assigned to &WORK without changing the path specified in &SUBMIT.

In the case of the USE command, SYSTAT first searches in the path assigned to &SAVE. If the file is not found there, then it searches in the &USE path.

Now, there may be occasions where files with the same name exist in both these paths but you specifically need to open one of them.

For example, suppose a file named MYDATA exists in the &USE path and you issue the following commands:

```
USE mydata
DSAVE mydata
```

This saves a copy of the data file MYDATA in the default &SAVE path. Suppose a file by name MYDATA also exists in the &USE path. Now, if you need to open the original file that is in the &USE path, you will either have to issue the USE command with the full path or:

```
USE &USE\mydata
```

Refer Chapter 7, *Customization of the SYSTAT environment*, for details about SYSTAT's file locations.

Examples

The examples presented here illustrate some practical implementations of token substitution. For more examples, examine the command files used in the Graph Gallery.

Example 1
Automatic Substitution in Exploratory Analysis

In this example, automatic token substitution defines the input file to use. SYSTAT then prompts for a variable and creates a bar graph.

```
TOKEN &infile = survey2
TOKEN &catvar / TYPE=VARIABLE,
       PROMPT='Select the variable appearing in the bar graph.'
USE '&infile' / NONAMES
NOTE 'File in use = &infile'
CATEGORY &catvar
BAR &catvar
CATEGORY &catvar/OFF
```

The path to the file contains spaces and must therefore be enclosed in quotes when defining the token. However, the quotes appearing in the token definition are not included in the token value. To direct SYSTAT to the correct path, we use quotes around the token in the USE command. Without those quotes, the program would look for a file named *program* and would return an error.

Repeated submissions of this template allow rapid creation of exploratory bar charts to study the distributions of variables in the *SURVEY2* file. Due to the automatic substitution, we are not prompted for a data file on each submission. To change data files, replace the path and the file in the first TOKEN command in the template. The note appearing in the output automatically updates to reflect the new file.

Example 2
Token Substitution for Variables and Strings

Variable substitution allows templates to be used for any data file. The resulting output has the same general structure, but varies in its content. String, number, and integer substitution allows customization, giving output from different files unique features.

Here, we create a three-dimensional scatterplot.The string tokens provide custom labels and a title to help differentiate the plot from other 3D plots generated from other submissions of this template.

```
TOKEN &xvar / TYPE=NVARIABLE IMMEDIATE,
      PROMPT='Select a variable for the x-axis.'
TOKEN &xvarlab / TYPE=STRING IMMEDIATE,
      PROMPT='Enter a label for the x-axis:'
TOKEN &yvar / TYPE=NVARIABLE IMMEDIATE,
      PROMPT='Select a variable for the y-axis.'
TOKEN &yvarlab / TYPE=STRING IMMEDIATE,
      PROMPT='Enter a label for the y-axis:'
TOKEN &zvar / TYPE=NVARIABLE IMMEDIATE,
      PROMPT='Select a variable for the z-axis.'
TOKEN &zvarlab / TYPE=STRING IMMEDIATE,
      PROMPT='Enter a label for the z-axis:'
TOKEN &pltitle / TYPE=STRING,
      PROMPT='Enter a title for the plot:'
TOKEN &symlabel / TYPE=CVARIABLE,
      PROMPT='Select a variable to use for labeling the plot points.'
TOKEN &symsize / TYPE=NVARIABLE,
      PROMPT='Select a variable to use for sizing the plot points.'
PLOT &zvar*&yvar*&xvar / SIZE=&symsize LABEL=&symlabel,
         TITLE='&pltitle',
         XLAB='&xvarlab' YLAB='&yvarlab' ZLAB='&zvarlab'
```

We use the IMMEDIATE option to ensure that the axis labeling prompts occur immediately after the corresponding axis assignment.

In the PLOT command, we enclose the string tokens in quotation marks. Doing so preserves the case of the entered value and prevents potential syntax errors resulting from spaces in the replacement text.

Variable Creation

The VARIABLE, NVARIABLE, CVARIABLE, MULTIVAR, NMULTIVAR, and CMULTIVAR types of the TOKEN command allows the user to select a variable or variables from those found in the current data file. These types cannot be used to create new variables. Instead, use the STRING type for variable creation.

In this example, we create ten new variables. Each variable contains 100 cases drawn randomly from a standard normal distribution.

```
TOKEN &v / TYPE=STRING,
    PROMPT='Enter a name for the new variables.,
            Names should be 256 characters long or less.'
  NEW
  DIM &v(10)
  REPEAT 100
  FOR i=1 TO 10
    LET &v(i)=ZRN ()
  NEXT
```

The DIM statement reserves memory for ten subscripted variables, assigning a root name supplied by the user. REPEAT generates 100 cases. The FOR..NEXT loop assigns standard normal deviates to each of the ten variables.

Notice that although we are dealing with variables, the VARIABLE type refers to *existing* variables and thus cannot be used for our purposes, namely to create *new* variables.

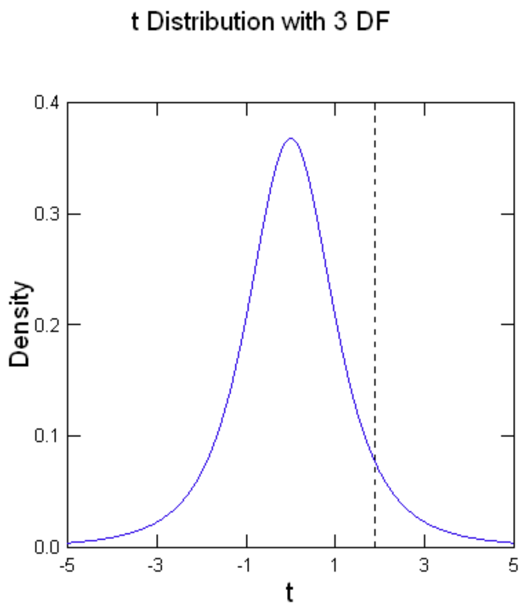
Example 3
Token Substitution for Numbers and Integers

The following commands generate a t-distribution with a reference line at a specified location. The output includes the cumulative area up to, and the probability of obtaining a value as extreme as, the given value.

```
TOKEN &df / TYPE=INTEGER PROMPT='Enter the degrees of freedom,
for the t-distribution.'
TOKEN &tval / TYPE=NUMBER PROMPT='Enter a t value.'
FPLOT Y=TDF(t, &df);  XLIMIT=&tval  XLAB='t'  YLAB='Density',
TITLE='t Distribution with &df DF'
TEMP tarea~ = TCF(&tval, &df)
PRINT "Area to the left of &tval = ", tarea~
IF (&tval >= 0) then TEMP pval~ = (2*(1-tarea~))
IF (&tval < 0) then TEMP pval~ = (2*tarea~)
PRINT "Two-tailed p-value =    " , pval~
```

The degrees of freedom for a t-distribution must be an integer so we restrict the corresponding token to accept values of this type. However, t-values can be decimal numbers so we only restrict our t-value token to be a number instead of a character.

The template uses the two tokens to compute the desired statistics. In addition, the *&df* token is used to generate a function plot and to title the plot. The other token, *&tval*, appears as a reference line in the function plot and in the output messages. The output using a value of 1.88 for a t-distribution having 3 degrees of freedom follows:




```
Area to the left of 1.88 = 0.922
Two-tailed p-value = 0.157
```

Example 4
Normal Random Deviates Using Tokens

No other distribution has received more attention or been used more often than the normal. In keeping with this trend, we use tokens to generate random deviates from a normal distribution with a user-specified mean and standard deviation. The user also indicates the number of deviates to create. The final command plots the normal distribution.

```
TOKEN &num / TYPE=INTEGER,
 PROMPT='How many standard normal random observations should be,
 generated?'
```

```
TOKEN &mean / TYPE=NUMBER,
    PROMPT='What is the mean for the normal distribution?'
TOKEN &stdev / TYPE=NUMBER,
    PROMPT='What is the standard deviation for the normal,
distribution?'
  NEW
  REPEAT &num
  LET nrd=ZRN(&mean,&stdev)
  DENSITY nrd / NORMAL
```

This template writes the generated deviates to a new variable named *NRD*. Alternatively, you could use another token to prompt the user to specify a name for the new variable.

Example 5
Random Number Generation Using Tokens

In this example, we combine interactive and automatic token substitution to generate random deviates from one of four distributions: Uniform, Normal, Exponential, or Logistic.

```
TOKEN &rndnum='rndnum'
TOKEN &RN='RN()'
TOKEN &dist / TYPE=STRING IMMEDIATE,
    PROMPT='Select a distribution by entering a letter.,
            (U=Uniform; Z=Normal; E=Exponential; L=Logistic),
            Default parameter values = (0,1)'
TOKEN &num / TYPE=INTEGER,
    PROMPT='How many random observations should be generated?'

  NEW
  REPEAT &num
  LET &dist&rndnum=&dist&RN
DENSITY &dist&rndnum / FILL=.5
```

The *&dist* token yields a dialog prompting for a single letter. We use the IMMEDIATE option to prevent the prompt for the number of observations from appearing first.

The LET statement combines three tokens to yield one transformation statement. A closer examination of this statement reveals some of the subtleties of token processing:

- First, we need a replacement value for *&dist*. Due to the IMMEDIATE option, this token already has a replacement value (U, Z, E, or L) so processing continues. Suppose the entered value equals *U*.
- Next, we encounter the *&rndnum* token. The first TOKEN statement assigns this token a value of *rndnum*. As a result, the left side of the LET statement becomes

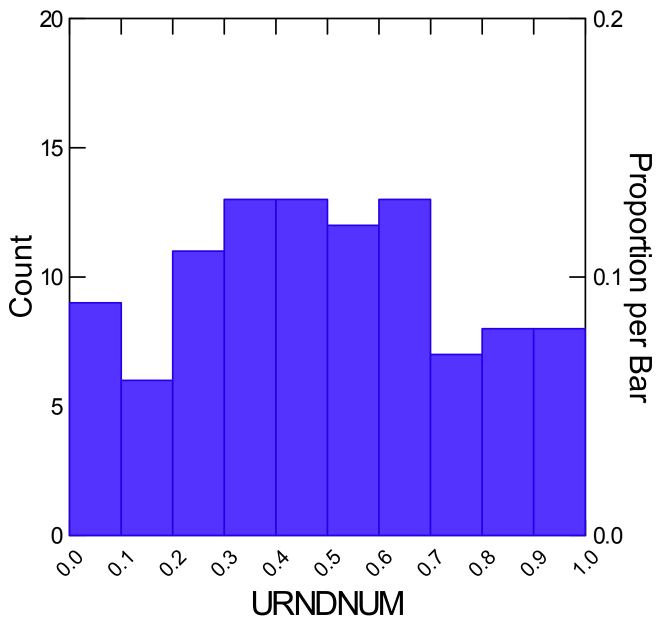
```
LET Urndnum
```

- After the equals sign, we again find the *&dist* token, which has a value of *U*.
- The final token on this line, *&RN*, has an assigned value of '*RN ()*', resulting in the following valid transformation statement (after token substitution):

```
LET Urndnum = URN()
```

The template creates a new variable with a seven-character name. The first character of the name denotes the distribution used to generate the values, and the final six indicate that the entries correspond to random numbers.

The output after randomly generating 100 observations from a uniform distribution follows:



Example 6
Multiple Variable Substitution

The number of variables analyzed often varies across applications of a particular technique. For instance, one regression model may include two variables, but another may include four. We can create a template for each model as follows:

```
TOKEN / ON
TOKEN &file/TYPE= OPEN PROMPT = "Choose a file to run regression"
TOKEN &resp/TYPE= variable prompt= ”Select the dependent varaible,
of the model”
TOKEN &v1/TYPE = variable prompt = “Select a variable”
TOKEN &v2/TYPE = variable prompt = “select a variable”
TOKEN &v3/TYPE = variable prompt = “select a variable”
TOKEN &v4/TYPE = variable prompt = “select a variable”

USE &file
REM Two predictors
REGRESS
MODEL &resp = CONSTANT + &v1 + &v2
ESTIMATE

REM Four predictors
REGRESS
MODEL &resp = CONSTANT + &v1 + &v2 +,
                           &v3 + &v4
ESTIMATE
```

Unfortunately, although these templates apply linear regression to user-specified variables, these templates only apply to models involving two and four predictors, respectively.

To create templates allowing for a varying number of variables, use the MULTIVAR, NMULTIVAR, and CMULTIVAR token types. Here, we create a linear regression template allowing any number of predictors and generate hypothesis tests to determine whether coefficients equal zero.

```
TOKEN &resp / TYPE = NVARIABLE,
     PROMPT = 'Select the response variable.'
TOKEN &predictors / TYPE = NMULTIVAR SEPARATOR = '+',
     PROMPT = 'Select the predictor variables,
               for the multiple regression model.'
```

```
TOKEN &hypeff / TYPE = NMULTIVAR SEPARATOR = '&',
PROMPT='Select predictors whose coefficients,
             you wish to test for differences from 0.'

REGRESS
MODEL &resp = CONSTANT + &predictors
ESTIMATE
HYPOTHESIS
ALL
TEST
HYPOTHESIS
EFFECT hypeff
TEST
TOKEN /OFF
```

The *&predictors* token represents all predictors in the model. The user selects the variables to include and SYSTAT generates the token value by inserting a '+' between them, yielding a valid MODEL statement.

The first HYPOTHESIS command generates a test for each coefficient in the model. The second HYPOTHESIS omits the selected variables from the regression model and compares the result with the original model. The EFFECT statement for this test requires an ampersand between terms, so we define the separator for this token to be '&'.

Example 7
Graph Option Template

The Graph tab of the Global Options dialog defines several appearance features for subsequently created graphs. As an alternative, the following template prompts for scaling percentages, line thickness, and character size before submitting a command file. As a result, all graphs created by the specified file use common values for these three global graph characteristics.

```
TOKEN / ON
TOKEN &xyscale /TYPE=INTEGER,
    PROMPT='Enter the % reduction or enlargement for graphs.,
            Values below 100 result in reduction.,
            Values above 100 result in enlargement.'
TOKEN &charsize / TYPE=NUMBER,
```

```
   PROMPT='Enter the factor by which to scale graph characters.,
            A value of 2 doubles the character size.,
            A value of .5 halves the character size.'
TOKEN &linethickness / TYPE=NUMBER,
    PROMPT='Enter the factor by which to scale line thickness.,
            A value of 2 doubles the line thickness.,
            A value of .5 halves the line thickness.'

TOKEN &cmdfile / TYPE=OPEN,
     PROMPT='Open a command file for creating graphs'
SCALE &xyscale &xyscale
CSIZ &charsize
THICK &linethickness
SUBMIT &cmdfile
SCALE
CSIZE
THICK
```

The final three commands return the global options to their default settings.

Example 8
Combining Analyses -- Two-Way ANOVA

Menus and dialogs offer a prescribed set of options resulting in a variety of statistics and graphs. When performing a series of analyses or including graphs with statistical output, using token substitution simplifies the process considerably. For example, multidimensional scaling requires a matrix input. You could generate this matrix from a rectangular file using the CORR procedure before running MDS. You could then save the final configuration for custom plotting. Instead of running each procedure separately, however, we can automate the entire process using a template. You can apply the template to any data to generate output customized to your needs.

In this example, we focus on two-way ANOVA. Using four tokens, we generate:

- box plots displaying the distribution of the dependent variable for every level of each factor.
- analysis of variance results.
- post-hoc tests for main and interaction effects.

- an interaction plot displaying the dependent variable mean in each cross-classification of the two factors.
- a residual plot.
- a stem-and-leaf-plot of the residuals.

```
USE OURWORLD
     TOKEN / ON
     TOKEN &outfile / TYPE=SAVE PROMPT='Save ANOVA Statistics'
     TOKEN &factor1 / TYPE=variable,
                      PROMPT='What is the first factor?'
     TOKEN &factor2 / TYPE=variable,
                      PROMPT='What is the second factor?'
     TOKEN &dep / TYPE=variable,
                  PROMPT='What is the dependent variable?'
     NOTE 'Two-way Analysis of Variance of'
     NOTE '&dep using &factor1 and &factor2 as factors'
     DENSITY &dep * &factor1 &factor2 / BOX
     ANOVA
        CATEGORY &factor1 &factor2/REPLACE
        DEPEND &dep
        SAVE &outfile / RESID DATA
        ESTIMATE

     HYPOTHESIS
     POST &factor1/ SCHEFFE
     TEST
     HYPOTHESIS
     POST &factor2/ SCHEFFE
     TEST
     HYPOTHESIS
     POST &factor1*&factor2/ SCHEFFE
     TEST
     USE &outfile
     CATEGORY &factor1 &factor2 / REPLACE
     LINE ESTIMATE*&factor1 / OVERLAY GROUP=&factor2,
                              TITLE='Least Squares Means',
                              YLAB=&dep
     CATEGORY &factor1 &factor2 / OFF
     PLOT student*estimate / SYM=1 FILL=1
     STEM student
```

To create the same output without a template requires the following dialogs:

- Box Plot
- ANOVA: Estimate Model
- Three uses of GLM: Pairwise Comparisons invoked thrice
- Line Chart
- Scatterplot
- Stem-and-Leaf

For every dialog, variable selection must occur. Creating a command file does automate these analyses, but command files do not generalize across data files.

By using this template, we replace the eight dialogs (and the necessary specifications for those dialogs) with four simple prompts. In addition, the resulting template can generate results for any specified data file.

Example 9
Chi-Square Test Using Choice Tokens

In this example, we perform chi-square test by offering four different choices for specifying expected frequencies. You will be prompted to open a data file and select the variable on which the test is to be performed. The computations are performed based on your choice of the form and way in which the expected frequencies are stored.

```
TOKEN
TOKEN / ON
TOKEN &filename / TYPE = OPEN PROMPT = "Select the file
to use." IMMEDIATE
USE &filename
TOKEN &ovar / TYPE = NVARIABLE IMMEDIATE PROMPT = "Select
the variable you want to analyze." IMMEDIATE
TOKEN / TYPE = CHOICE PROMPT = "Specify the form and way
in which you want to input expected frequencies.",
"Equal expected frequencies" =
"Miscellaneous\choice1.syc",
"Equal expected frequencies with missing values as a
separate category" = "Miscellaneous\choice2.syc",
"Unequal expected frequencies specified in a data file"
= "Miscellaneous\choice3.syc",
"Unequal expected frequencies specified through the
keyboard" = "Miscellaneous\choice4.syc"
```

If you select the first choice, chi-square test is performed using a one-way crosstabulation of the input variable by assuming equal expected frequencies across cells. The second choice does the same while treating missing values of the input variable as a separate category. The third choice accepts unequal expected frequencies in the form of a column in the data file. The fourth allows you to enter the expected frequencies using the keyboard.

For another illustration of choice tokens, try the Simple Correspondence Analysis Plot in the Graph: Graph Gallery dialog box.

Chapter

6

Working with Output

Lou Ross
(revised by Poornima Holla)

All of SYSTAT's output appears in the Output editor, with corresponding entries appearing in the Output Organizer. You can save and print your results using the File menu. Using these options, you can:

- Reorganize and reformat output.
- Save data and output in text files.
- Save charts in a number of graphics formats.
- Print data, output, and charts.
- Save output from statistical and graphical procedures in SYSTAT output (SYO) files, Rich Text Format (RTF) files, Rich Text Format (Wordpad compatible) (RTF) files, HyperText Markup Language (HTML) files, or (MHT) files.

You can open SYSTAT output in word processing and other applications by saving them in a format that other software recognize. SYSTAT offers a number of output and graph formats that are compatible with most Windows applications.

Often, the easiest way to transfer results to other applications is by copying and pasting using the Windows clipboard. This works well for charts, tables, and text, although the results vary depending on the type of data and the target application.

Output Editor

The Output editor displays statistical output and graphics. You can activate the Output Editor by clicking on the tab, or selecting

View
  Output Editor

Using the Output editor, you can reorganize output and insert formatted text to achieve any desired appearance. In addition, paragraphs or table cells can be left-, center-, or right-aligned.

Tables. Several procedures produce tabular output. You can format text in selected cells to have a particular font, color, or style. To further customize the appearance of the table (borders, shading, and so on), copy and paste the table into a word processing program.

Collapsible links. Output from statistical procedures appears in the form of collapsible links. You can collapse/expand these links to hide/view certain parts of the output.

Graphs. Double-clicking on a graph opens the Graph in the Graph tab. When the Output editor contains more than one graph, the Graph tab contains the last graph.

Note: The Output editor supports opening and editing output files of .SYO format created in previous versions of SYSTAT. Such output files, however, cannot be saved.

Format

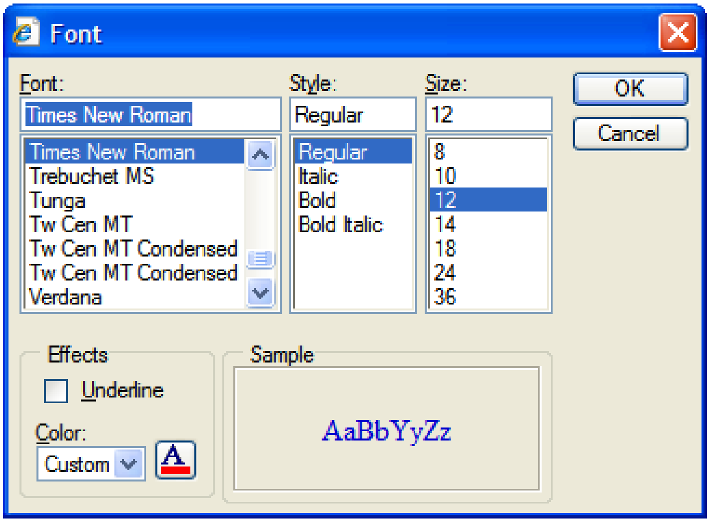
SYSTAT displays different formatting tools. To change the formats of the outputs, go to

Edit
  Format...

and then apply different formatting tools. Common formatting tools also appear on the toolbar in Customize... in the View menu, and in the toolbar in the Output editor.

Fonts. SYSTAT displays output in an Arial font by default. Select Font dialog box from

Edit
 Format
  Font...



Use different options of the Font dialog box to change the appearance of any selected output text. You can select the desired font type, style and size. You can also select effects like Underline and font color to be used.

Font Style. You can change the selected output text to Bold, Italicized, and Underlined typefaces and also change the font color of the selected output text by selecting these options from Format in the Edit menu.

Alignment. You can align the selected output text to the left, right or centre by selecting those options in Format.

Bullets and Numbering. Any selected text can be formatted as a Numbered list or a Bulleted list from the options in Format. You can also reduce the indentation of the text or indent text by selecting Outdent or Indent respectively.

Inserting Image. You can insert an image in the desired location of the Output editor by selecting the Insert Image option in Format.

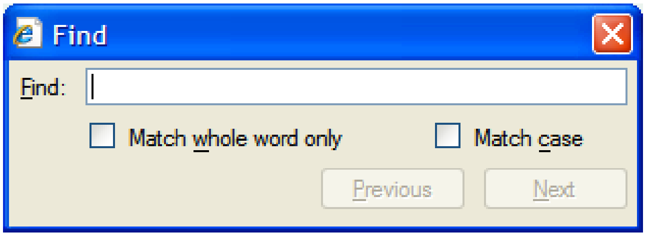
Collapsible Links. By selecting the Expand All option in Format you can expand all the links in the output; you can collapse all those expanded links by selecting the Collapse All option in Format.

Find

You can search for specific numbers or text in the Output editor.

To open the Find dialog box, from the menus choose:

Edit
  Find…



Search strings contain either complete or partial text. SYSTAT searches the specified direction (up or down) from the current location. A string search may consist of only letters or letters with numbers and punctuation. For any search involving letters, you can impose a case restriction. For example, selecting Match case prevents a search for *median* from finding *Median*.

Note: SYSTAT operates in the active space. Click the Output editor to make it active. If the Commandspace is active, SYSTAT searches in the active tab of the Commandspace.

Output Editor Right-Click Menu

Right-clicking in the Output editor provides standard editing features. These are:

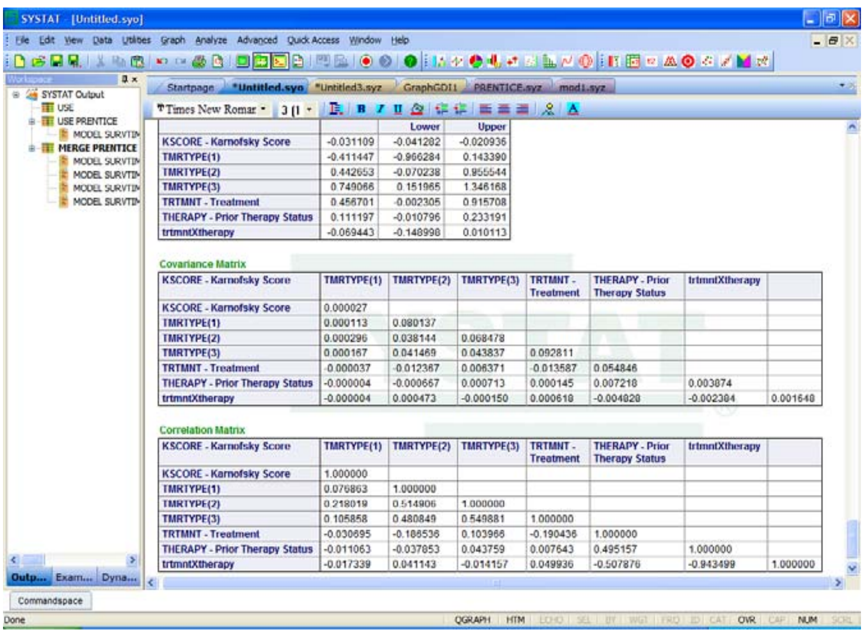
- **Cut.** Cut the selection and place it in the clipboard for pasting at the desired location(s).
- **Copy.** Copy the selection and place it in the clipboard for pasting at the desired location(s).
- **Paste.** Paste previously cut or copied output.
- **Delete.** Delete the selections in the active tab.
- **Copy All.** Copy all the content in the Output editor.
- **View Source.** View the HTML source code.

- **Refresh.** Refresh the content being viewed in the Output editor.
- **Print Preview.** Display the file in the active tab as it would appear when printed. You can view multiple pages at a time, scroll through and zoom in or out of pages.
- **Collapse All /Expand All.** Collapse/Expand all the links in the Output editor.
- **Show Toolbar.** Show or hide the Format Bar.
- **New Output.** Open a new output file in the Output editor where further output will appear. If an output file is already open, it is closed with the option of saving it.
- **Save As.** Save the file in the active tab, as a separate file. You will be prompted to specify the name and location of your choice.
- **Options.** Set SYSTAT's global options according to your preferences.

Note: Cut, Copy, and Delete are available only when a selection has been made.

Output Organizer

The Output Organizer serves primarily as a table of contents for the Output editor. Use it to jump to any location in the Output editor without having to scroll through long statistical or graphical results.



Each data file opened during a session, creates a new tree folder in the Output Organizer. Within each tree folder, each procedure generates entries -- one for text results and one for every graph. If there is no data file open, the entry is created under the last tree folder. Clicking an entry scrolls the Output editor to the corresponding output. Double-clicking on a graph entry opens the corresponding graph in the Graph tab. When the Graph tab is active, clicking a graph entry dynamically changes the graph that is displayed in the Graph tab.

You can close folder icons by clicking the "-" to the immediate left. Clicking a "+" opens the corresponding folder. In case of the SYSTAT output tree, you can also close (open) it by selecting Collapse Tree (Expand Tree) from the Edit menu. However, opening and closing folders in the Organizer does not affect the Output editor.

A second use of the Output Organizer is to reorganize the results in the Output editor. Cutting, copying, or pasting in the Organizer yields parallel results in the Output editor. For example, clicking an icon in the Output Organizer selects that entry. Clicking a folder icon selects all entries contained in that folder. With the Organizer entry selected, copying (via the Edit menu or right-clicking) results in the output corresponding to the selection being copied to the clipboard. Select a new entry and paste to insert the copied output at the new location. Note that although the Organizer represents an outline of what will be copied from the Output editor, the Output editor itself does not show the selection.

Transformations. Because transformations do not produce output, they do not generate Output Organizer entries. To note when transformations occur, echo the commands or add notes to the output. However, echoed commands still do not yield an entry in the Organizer.

To Move Output Organizer Entries

You can reorganize SYSTAT's output simply by selecting and dragging Organizer entries to new locations. Use the Shift key to select a range of entries or the Ctrl key to select multiple but nonconsecutive entries. Selecting a folder entry causes all items within the folder to be selected. The Organizer places selected items immediately after, and at the same level as, the location to which you drag them.

If you select items at differing levels and drag them to a new location, SYSTAT places the entries at the level of the target location.

To Insert Tree Folder

SYSTAT generates Output Organizer entries for all statistical and graphical procedures. You can also create customized tree folders. Use customized trees to place output from several procedures in one location.

To insert a new tree folder, from the menus choose:

Edit
  Output Organizer
    Insert Tree Folder...

Alternatively, you can right-click on the Output Organizer, and select 'Insert Tree Folder'. SYSTAT creates a folder named 'New Folder'. To rename it, select the folder and go to

Edit
  Output Organizer
    Rename...

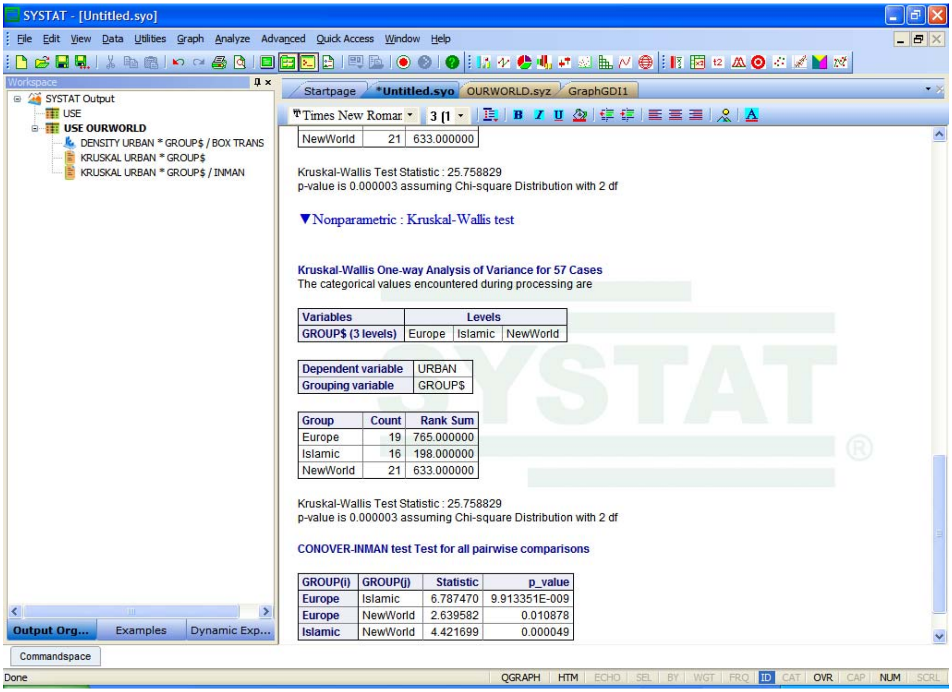
Alternatively, right-click on the folder and select Rename. Headings appear just below and at the same level as the selected Organizer entry.

You can rename any Output Organizer entry, collapse/expand all trees from Output Organizer in the Edit menu or from the right-click menu of Output Organizer. You can also view a data from the right-click menu of Output Organizer.

Configuring the Output Organizer

Output Organizer headings are often truncated at the right edge of the pane. To view the entire heading, move the mouse over the heading.

Alternatively, you can resize the Workspace by dragging the boundary between the Viewspace and Workspace to new locations. Position the pointer of your mouse over the boundary until a double-headed arrow appears. Click your left mouse button and hold it down while you drag the pane edge to the desired location.



You can hide (or view) the entire Output Organizer without resizing it by selecting

View
  Workspace...

Although the Output Organizer may be hidden, the subsequent output still generates entries in the tree. Consequently, you can jump quickly to a specific output by reopening the Workspace and clicking on the entries.

Workspace settings persist across SYSTAT sessions. For example, if you hide the Workspace and close SYSTAT, the next SYSTAT session begins with the Workspace hidden.

To view the entire Viewspace in the full screen mode, from the menus choose:

View
  Full Screen Viewspace...

Output Organizer Right-Click Menu

Right-clicking in the Output Organizer provides some important features. These are:

- **Rename.** Rename the selected tree folder.
- **Expand All/ Collapse All.** Expand/Collapse the Output Organizer tree without affecting the output in the Output editor.
- **Insert Tree Folder.** Insert a new tree folder under the active Output Organizer data node. You can drag and drop Output Organizer text and graph nodes and other tree folders into this tree folder.
- **Set as Active Data File.** Set the data file as active. With more than one data file open in the Output Organizer, this gives you the option to work with any previously opened data file as active.
- **View Data.** View the data file corresponding to the selected data file node.
- **New Output.** Open a new output file in the Output editor where further output will appear. If an output file is already open, it is closed with the option of saving it.
- **Clear Output.** Clear all the output generated in the Output editor so far.
- **View Graph.** View the graph corresponding to the selected graph node, in the Graph tab.
- **Save As.** Save the file that is in focus as a separate file. You will be prompted to specify the name and location of your choice.
- **Show Detailed Captions.** Show the underlying SYSTAT commands as Output Organizer node captions.

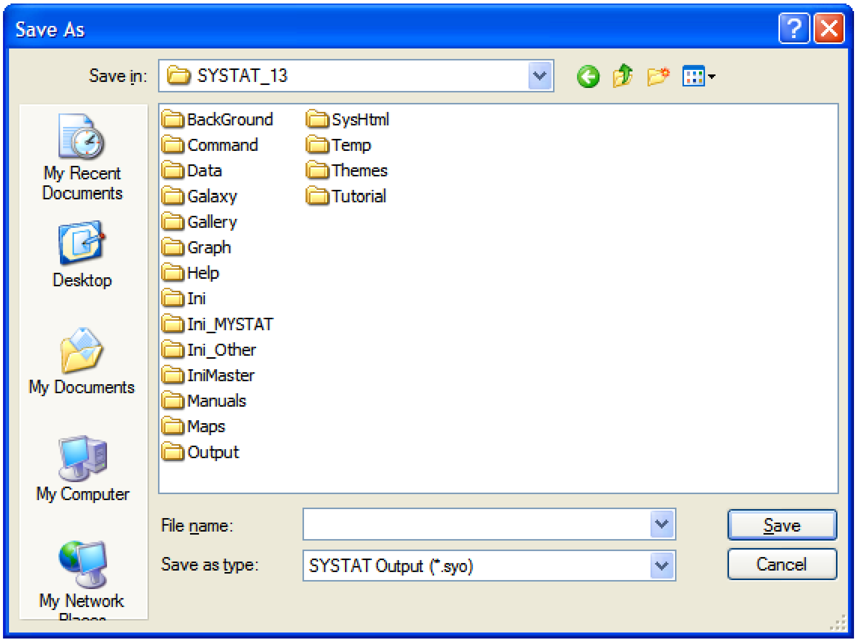
Saving Output and Graphs

You can save the contents of the active tab or pane in a file. SYSTAT saves combined statistical and graphical output in four file types. In addition, individual graphs can be saved in a number of graphic formats.

When you choose Save Active File from the File menu, what is saved depends on which pane is active. If either the Output Organizer or the Output editor is active, the entire contents of both panes are saved. When you choose Save All from the File menu, the current output, data file, and the current file of the commandspace are all saved.

To Save Output

SYSTAT displays statistical and graphical output in the Output editor. Click the Output Organizer or Output editor and choose Save As from the File menu to save the contents of the pane. You can save Data, Command, Output, Graph, or Log using Save from File menu.



Select a directory and specify a name and file type for the output. Output can be saved in SYSTAT Output (*.*SYO*), Rich Text Format (*.*RTF*) , Rich Text Format (Wordpad compatible) (*.*RTF*), Hyper Text Markup Language (**.HTM)* or (*.*MHT*) format.

Note: Unlike output saved in *SYO* or *RTF* format, output saved in HTM or MHT format preserves some properties:

- HTML or MHT outputs are not editable.
- As HTML or MHT underlies web page creation, presenting the resulting output on the Internet involves simply creating a link from a web page to the filename.htm or .mht file. In addition, HTML or MHT output allows sharing your results with colleagues who do not (yet) have SYSTAT, but do have a browser, by simply supplying the .htm or .mht file.

Using Commands

To save output, enter the following:

```
OSAVE FILENAME / SYO or RTF OR HTML or MHT
```

Omitting SYO or RTF or HTML or MHT saves the output as a SYSTAT output file with an *.SYO* extension.

To Direct Output to a File or Printer

You can use commands to send output directly to a file or the printer:

```
OUTPUT <FILENAME> | VIDEO or * | PRINTER or @ |
                    [ /COMMANDS, ERRORS, WARNINGS ]
```

For example, the commands below send a listing of cases, including commands, to the text file *MYFILE.DAT*. The OUTPUT * command at the end closes the text file so that subsequent output is sent to the screen only.

```
USE OURWORLD
OUTPUT MYFILE /COMMANDS
LIST COUNTRY$ HEALTH
OUTPUT *
```

To Save Results from Statistical Analyses

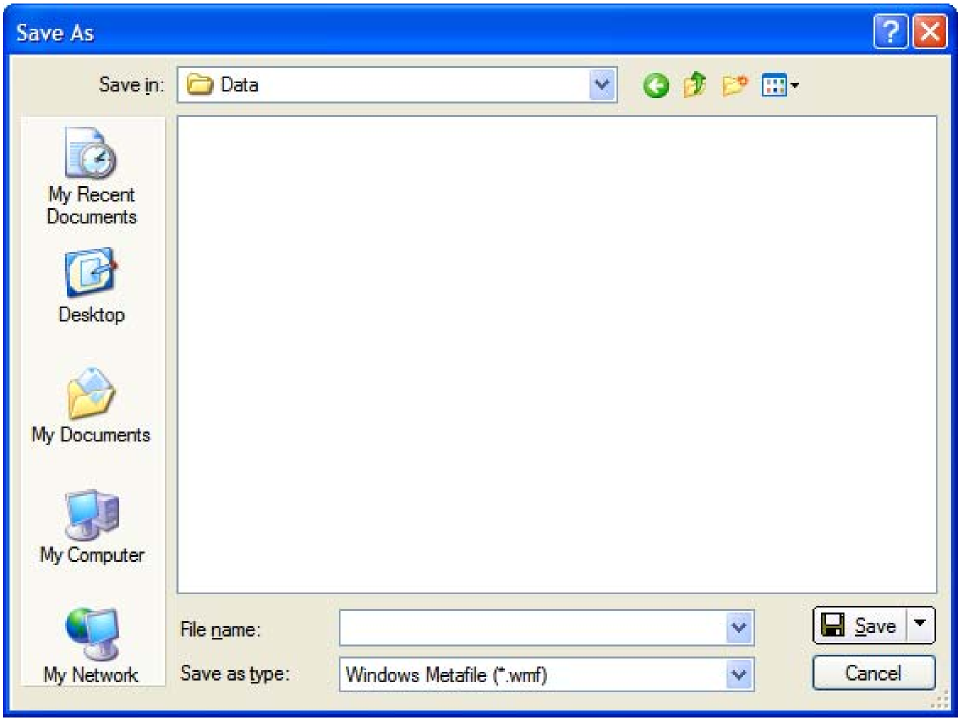
Many procedures include an option such as Save or Save File that saves the results of the analysis in a SYSTAT data file. The contents of the file depend on the analysis. For example:

- Correlations can save Pearson and Spearman correlations.
- Factor Analysis can save factor scores, residuals, and a number of other statistics.
- Linear Regression can save residuals and diagnostics for each case.
- Basic Statistics can save selected statistics for each level of one or more grouping variables.
- Crosstabs can save the count in each cell for later use as table input.

Check each procedure to see what is saved.

To Save Graphs

SYSTAT displays graphs in the Output editor of the Viewspace. You can save the graphs along with the output by using the Save on the File menu. To save an individual graph, double-click the graph to activate the Graph tab and use Save As on the File menu.



By default, the file is saved as a Windows Metafile (*.*WMF*). You can select a different file type from the drop-down list. Available formats include:

- Windows Metafile (*.*WMF*)
- Windows Enhanced Metafile (*.*EMF*)
- Encapsulated Postscript (*.*EPS*)
- PostScript (*.*PS*)
- JPEG (*.*JPG*)
- Windows Bitmap (*.*BMP*)

- Computer Graphics Metafile: binary or clear text (*.*CGM*)
- Tagged Image File Format (*.*TIFF*)
- Graphics Interchange Format (*.*GIF*)
- Portable Network Graphics (*.*PNG*)

Depending on the graphic format, you can select from a number of options when saving the file. See the online help for details.

Using Commands

To save an individual graph, enter the following:

```
GSAVE FILENAME / FILETYPE
```

For *FILETYPE*, enter one of the following: WMF, EMF, EPS, PS, JPG, BMP, TIFF, GIF, or PNG. SYSTAT saves the most recently created graph as *FILENAME*. Issuing multiple, consecutive GSAVE commands results in multiple graphs being saved. SYSTAT saves the most recent first, the graph created before the most recent graph second, and so on. However, issuing any other command after a GSAVE command resets the internal index for the next GSAVE to the most recent graph.

To save all graphs in the Output Editor, use:

```
GSAVE ROOT / ALL FILETYPE
```

When naming the resulting files, the software appends consecutive integers beginning with 1 to *ROOT*.

To Export Results to Other Applications

You can open your saved output and charts in word processing and other applications. In SYSTAT, save the file in a format that the other application can handle; then open or import the file in that application. SYSTAT offers a number of graph formats that are compatible with most Windows applications. For example, you can save a SYSTAT graph as a Windows Metafile (*.*WMF*) and then insert or import the metafile into most Windows word processing applications. See the target application's documentation for specific information.

To Export Results Using the Clipboard

Often, the easiest way to transfer results to other applications is to copy and paste using the Windows clipboard. This works for charts as well as text, although results vary depending on the target application.

In SYSTAT, select the output or chart.

From the menus choose:

Edit
  Copy

- In the other application, position the cursor where you want the output to appear.

From the menus choose:

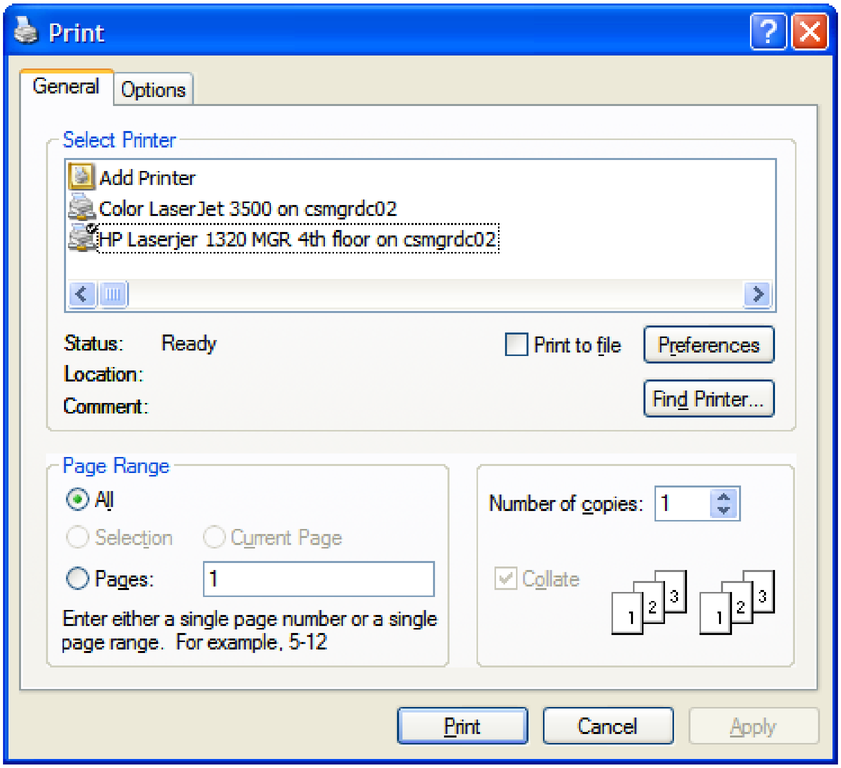
Edit
  Paste

Tips:

- If you have problems with Paste, try using Paste Special on the Edit menu in the target application. With Paste Special, you can specify whether you want to paste the clipboard contents as text or a Windows Metafile (graphic). (Note that Paste Special is not available in all applications.)
- For columns to line up properly, you must highlight text output after you paste it and apply a fixed-pitch font (for example, Courier or Courier New). Or, use Paste Special on the Edit menu to paste the text as a metafile graphic.

Printing

In any SYSTAT window, choose Print from the File menu to open the Print dialog box.



Select a printer and a print range. You can choose to print the current selection, the entire print range, or a specific page range.

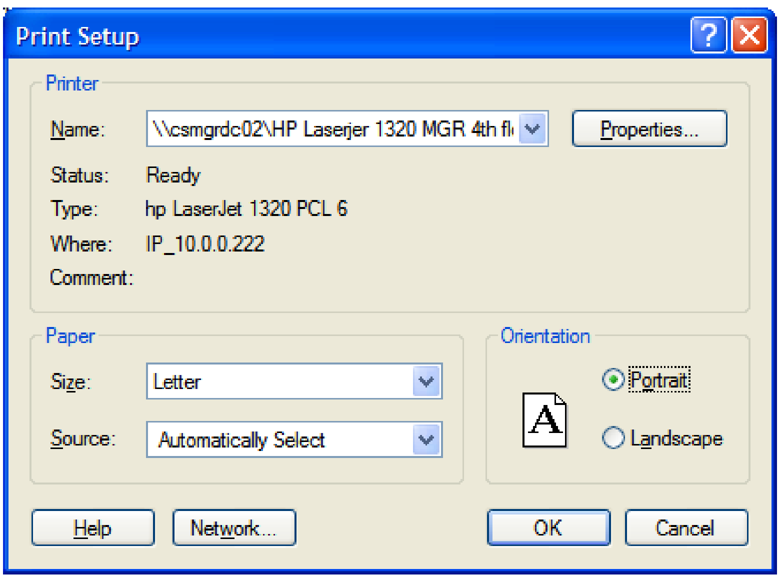
Use the Print Preview command in the File menu, to preview the content before printing it.

Print Preview

In any SYSTAT window, choose Print Preview from the File menu to display the active document as it would appear when printed. When you choose this command, the main window will be replaced with a print preview window in which one or two pages will be displayed in their printed format. The print preview toolbar offers you options to view either one or two pages at a time; move back and forth through the document; zoom in and out of pages; and initiate a print job.

Page Setup

To optimize printed output, you may need to adjust various page settings. The available options vary for different printers. To open the Page Setup dialog box, choose Page Setup from the File menu.



If more than one printer is installed on your system or network, you can choose which one to print to. You can also specify paper size and orientation--portrait (tall) or landscape (wide).

Printing Graphs Using Commands

You can print individual graphs by entering the following:

GPRINT / LANDSCAPE *or* PORTRAIT

SYSTAT automatically sends the most recently created graph to the default printer. In the absence of an orientation specification, the software uses the setting for the current printer. Issuing multiple, consecutive GPRINT commands results in multiple graphs being printed: SYSTAT prints the most recent graph first, the graph created before the most recent graph second, and so on. However, issuing any other command after a GPRINT command resets the internal index for the next GPRINT to the most recent graph.

Chapter

7

Customization of the SYSTAT Environment

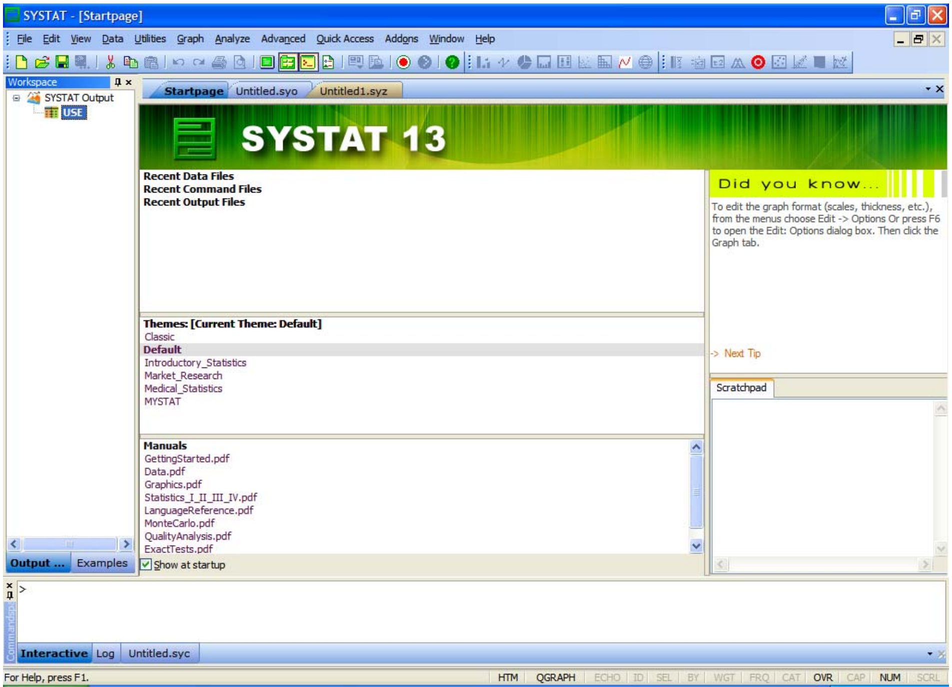
(*Revised by Rajashree Kamath*)

By default, the user interface contains, from top to bottom:

- Toolbars
- Workspace and Viewspace
- Commandspace
- Status Bar

However, as you work with SYSTAT, you may discover that an alternative window organization would better match the way you work.

The interface for SYSTAT can be completely restructured to create a comfortable, analytical environment in which you can be maximally productive.



You can:

- resize, hide, and reorganize windows and panes
- create, reposition, and modify toolbars
- assign sets of command files to a toolbar button, allowing quick submission of commonly used commands
- add menu items for frequently used commands and command files
- define settings for output, data, and graph appearance
- specify file locations for navigational ease
- define and set themes to suit your needs
- set the output to appear either based on data files used or in the order of execution of analyses

Commandspace Customization

Users who frequently use SYSTAT's command language may prefer a larger command area for viewing and editing of command files. To change the size of the Commandspace, hover the mouse on its upper boundary until the mouse cursor changes to a double-sided arrow , hold down the mouse and drag to a new location.

The output area is automatically resized to accommodate the resized Commandspace.

Alternatively, you can undock the Commandspace from the bottom edge of the user interface to increase the space available for displaying output. To do this:

- Click the upper boundary or sidebar of the Commandspace ensuring that the mouse pointer does not change appearance and drag the outline to a new location without releasing the mouse button. Hold down the Ctrl key as you drag, to prevent docking with the user interface. Release the mouse button and Ctrl key when the outline indicates the desired position.
- Double-click the upper boundary of a docked Commandspace to detach it into its last undocked position.

Similarly, you can dock the Commandspace to its original position:

- Click the title bar of the undocked Commandspace and drag the outline to a new location in the user interface without releasing the mouse button. Release the mouse button (do not press the Ctrl key while you do this) when the outline is at the desired position and touches either one of the edges of the user interface, or that of the Viewspace.
- Double-click the title bar of an undocked Commandspace to reattach it at its last docked position.

Hiding the Commandspace

An undocked Commandspace always appears in front of the rest of the user interface and may obscure output. In such a situation, it can be hidden until needed. Selecting Commandspace from the View menu, pressing Ctrl + W, right-clicking in the toolbar area and selecting Commandspace, or clicking the Close button after undocking it toggles the visibility of the Commandspace. Alternatively, you can hide the Commandspace and use a text editor like Notepad for command entry. The Commandspace can be collapsed by clicking the pin button.

Tip: Users who favor dialog use over typing commands should hide the Commandspace to maximize the area available for output.

Workspace Customization

The technique to customize the Workspace is analogous to that explained for the Commandspace. The Workspace can also be hidden either by invoking the View menu and selecting Workspace, by right-clicking on the toolbar area and selecting Workspace, or by clicking the Close  button after undocking the Workspace. You can collapse (auto hide) the Workspace by clicking on the auto hide pin .

Customizing the Output Organizer

You can customize the captioning of text nodes in the Output Organizer. By default, the caption is the title of the analysis that the node pertains to. The associated command appears as a tooltip on mouse hover. To see the tooltips themselves as node captions, from the menus choose:

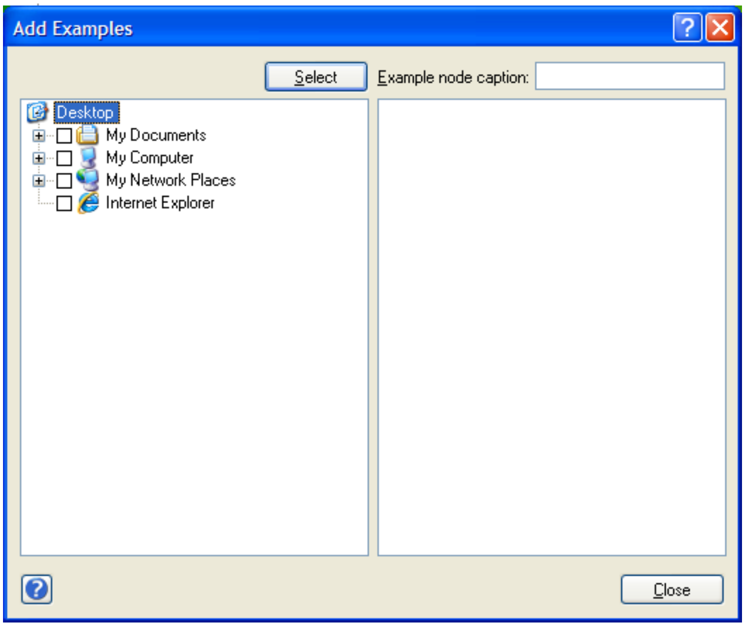
Edit
  Output Organizer
    Show Detailed Captions...

For a given analysis, the associated command is the most significant command related to that analysis; typically the HOT command. For example, for least-squares regression, the default node caption is 'OLS Regression' whereas the detailed node caption is the MODEL command line.

Adding Examples

The Examples tab in the Workspace contains a “SYSTAT Examples” tree that is organised by folders and nodes, the folders corresponding to volumes or chapters of the SYSTAT User Manual, and the nodes corresponding to the example command scripts therein. Double-clicking a node executes the underlying command script.You can add your own examples to this tree, organized according to the directory structure of your folder containing such examples.To add examples, from the menus choose:

Utilities
  Add Examples...



In the dialog that opens, the left hand side contains a box displaying all the drives, folders, sub-folders and files in your hard disk.There is a check box besides each item to indicate whether or not you want it to be included in the **Examples** tab.Click on the check box beside a folder twice if you want to include it along with all its sub-folders and files in the **Examples** tab.The check box changes to  when you do so. Click on it once if you want to include just the folder and the files in it. Click on a file once if you want to insert a node corresponding to the file in the Examples tree. Clicking again will allow you to uncheck an item.

When you check a folder, ensure that you have expanded all the nodes that belong to it so that all the filenames therein are seen. Once you have made your selections, enter an Example node caption. This caption will be set for the top-level folder that is to contain the links to your example command files. Then press **Select** so that the corresponding tree structure is displayed in the right hand side of the dialog box. You can review this tree and make any further changes if desired. Once you have finalized your selections, press **Close**. This will trigger the creation of an initialization file corresponding to your selections. Close the current session of SYSTAT and reopen it to see the newly added examples. If you need to replace an examples tree that you have created, specify the same Example node caption when you create the new tree.

Note: You can also customize the tree structure directly using the initialization files in the INI sub-folder of the SYSTAT program folder. Edit the "SycSamples.ini" file while maintaining the formatting of the content (described below). This initialization file expects the related command files to be in the SYSTAT Command folder. So you can add nodes for your own command files provided they are saved in the Command folder. Alternatively, you can save your command files in any desired location, create a new initialization file in the INI folder and enter the file path of the location suffixed by "\#*your cmdfiles.ini*" in the SysMaster.ini file that is in the INI folder. Use the following guidelines while creating the content of *your cmdfiles.ini*:

- Type the top level folder caption without indentation.
- Use a hash (#) at the end of a caption to define tree folders or nodes.
- Indent with the appropriate number of tab stops to create sub-folders or nodes within a given folder.
- If a caption relates to a node, type the filename (including the file extension) after the hash.You can even include a sub-folder name with the filename.You can also skip the caption in which case the filename will be used as the node caption.

Viewspace Customization

By default, the Data Editor and the Graph Editor tabs are in the Viewspace. However, users may want to view the Data Editor and the Graph Editor simultaneously. To do this, click the Window menu or right-click in the toolbar area and select Show Stacked or Show Side-by-Side. All the panes in the Viewspace get laid out in a tiled fashion. Click the Minimize  or Close  (if it is enabled) button of the panes that you do not want to see, and select Show Stacked or Show Side-by-Side again. The pane that is active will be placed first in the tiled layout. Using the Window menu (or context menu of the toolbar area), you can also Cascade windows or Arrange Icons that have been minimized. Double-click one of the title bars to dock the panes to their default or previously docked positions.

Maximizing the Viewspace

Almost every command and dialog box creates output, all of which appears in the Output Editor of the Viewspace. Occasionally, statistical output or graphs may be too large to be viewed in the Output Editor. Even data files will typically contain more number of rows than visible in one view. Although scrollbars allow control over the contents of the viewable area, displaying graphs or results in their entirety in a single pane simplifies interpretation.

The most obvious method for increasing the size of the Output Editor involves maximizing the user interface to fit the size of your monitor. You can close toolbars that you do not use frequently. You can resize or undock the Commandspace or Workspace to increase the viewable output region. You can also work with the Viewspace in the full screen mode. To set the Viewspace to the full screen mode

From the menus choose:

View
  Full Screen Viewspace...

Alternatively, right-click in the toolbar area and select Full Screen Viewspace.

However, some output may still require scrolling. When resizing alone cannot create an area large enough to view your output, consider hiding elements of the user interface, such as the Commandspace or the Workspace.

Startpage Customization

You can resize the partitions of the Startpage by positioning the mouse over any of the boundaries until the cursor changes to a double line | |, clicking and then dragging the boundary to the desired position.You can close the Startpage for the remainder of the session by clicking the View menu and selecting Startpage, by right-clicking on its tab and selecting Close, or by right-clicking on the toolbar area and selecting Startpage. You can even prevent the Startpage from appearing in subsequent sessions by unchecking the Show at startup check box in the Startpage.

Status Bar

The status bar appears at the bottom of the user interface.

When the mouse pauses on a toolbar button or menu entry (including right-click menus), the status bar displays a brief description of that item. These descriptions help guide you to the most appropriate procedure for a desired task. When the Graph Editor is active with a graph in it, the status bar displays the name of the graph element on which the mouse pointer is currently positioned.

The left side of the status bar will show the status of some output related options:

- **QGRAPH.** Displayed when statistical Quick Graphs are set to appear in the Output Editor. Toggle this mode on or off by clicking on it.
- **HTM.** Displayed when HTML based output is set to appear in the Output Editor. Click on this to toggle between HTML formatting and plain text formatting of the output.
- **PLENGTH.** NONE/SHORT/MEDIUM/LONG. NONE, SHORT, MEDIUM or LONG is displayed when the corresponding output length is set using the Global Options dialog or the PLENGTH command.
- **ECHO.** Displayed when the commands issued by the user are set to appear in the output. Click on it once if you do not want the commands to be echoed.
- **VDISP.** LABEL/NAMES/BOTH. LABEL/NAMES/BOTH is displayed depending on the global setting for display of variable labels or the VDISPLAY command.
- **LDISP.** LABEL/DATA/BOTH. LABEL/DATA/BOTH is displayed depending on the global setting for display of value labels or the LDISPLAY command.
- **NODE.** Displayed when detailed node captions are to be shown for the Output Organizer. Click on it once to display brief captions.
- **PAGE.** NONE/NARROW/WIDE. NONE/NARROW/WIDE is displayed depending on the global setting for page width or the PAGE command.

The middle portion of the status bar will show information about existing processing conditions on the data, and also allow you to edit them:

- **SEL.** Displayed when case selection is in effect. Pause the mouse on this to see the condition used for selection in the tooltip that appears. Click on SEL to invoke the Data: Select Cases dialog box and edit the condition or turn off selection.
- **BY.** Displayed when one or more grouping (By Groups) variables are declared. Pause the mouse on this to see the currently defined grouping variable(s) in the tooltip that appears. Click on BY to invoke the Data: By Groups dialog box and add/delete grouping variables, or turn off the By Groups declaration.
- **WGT.** Displayed when a weight variable is declared or exists in the data file. Pause the mouse on this to see the currently defined weight variable in the tooltip that appears. Click on WGT to invoke the Data: Case Weighting: By Weight dialog box and change the weight variable or turn off case weighting.
- **FRQ.** Displayed when a frequency variable is declared or exists in the data file. Pause the mouse on this to see the currently defined frequency variable in the tooltip that appears. Click on FRQ to invoke the Data: Case Weighting: By

Frequency dialog box and change the frequency variable or turn off frequency declaration.

- **ID.** Displayed when an ID variable is declared or exists in the data file. Pause the mouse on this to see the currently defined ID variable in the tooltip that appears. Click on ID to invoke the Data: ID Variable dialog box and change the ID variable or turn off ID variable declaration.
- **CAT.** Displayed when one or more categorical variables are declared or exist in the data file. Pause the mouse on this to see the currently defined categorical variable(s) in the tooltip that appears. Click on CAT to invoke the Data: Category dialog box and add/delete categorical variables, or turn off category declaration.

The right end of the status bar shows the current condition of the command autocompletion mode and four keyboard states:

- **AUTO.** Displayed when the Commandspace supports autocompletion of commands. Click on it to toggle this mode. See the Global Options section for details about this feature.
- **OVR.** Displayed when the keyboard is in overstrike mode. In this state, typed text *replaces* the text at the current location. This gets grayed out when the Insert key on your keyboard is pressed to set it to the insert mode. The insert mode allows insertion of new typed text at the current cursor location, shifting any existing text to the right.
- **CAP.** Displayed when Caps Lock is active. In this state, every typed letter appears in upper case. Use the Caps Lock key to toggle this state on and off.
- **NUM.** Displayed when Num Lock is active. With Num Lock on, the keyboard keypad enters numbers. With Num Lock off, the keypad moves the cursor in the current window. The Num Lock key toggles this state on and off.
- **SCRL.** Displayed when Scroll Lock is active. With Scroll Lock on, if the Data Editor is active and you use the arrow keys on the keyboard, the entire sheet will scroll. The Scroll Lock should be off, if you want to use the arrow keys for navigation around the Data Editor.

Status Bar Customization

Of the status bar items mentioned above, the QGRAPH, HTM, ECHO, SEL, BY, WGT, FRQ, ID, CAT and OVR items appear by default. You can add or remove items from the status bar by right-clicking on it. In the context menu that appears, check the items you want to keep and uncheck the items you do not use. You can get all the items to appear by selecting All Items; all the items will disappear if you select No Items. To revert to the default set of items, select Default Items. If you simply do not need the status bar or need more area available for a window, from the menus choose:

View
  Status Bar...

Repeat the above steps to bring back the status bar.

Customizing Menus and Toolbars in SYSTAT

Menu Customization

SYSTAT has a default organization for the menus and toolbars, based on similarity of features. However, users can customize these according to their needs and preferences using the Customize dialog box.

To open the Customize dialog box, from the menus choose:

View
  Customize…

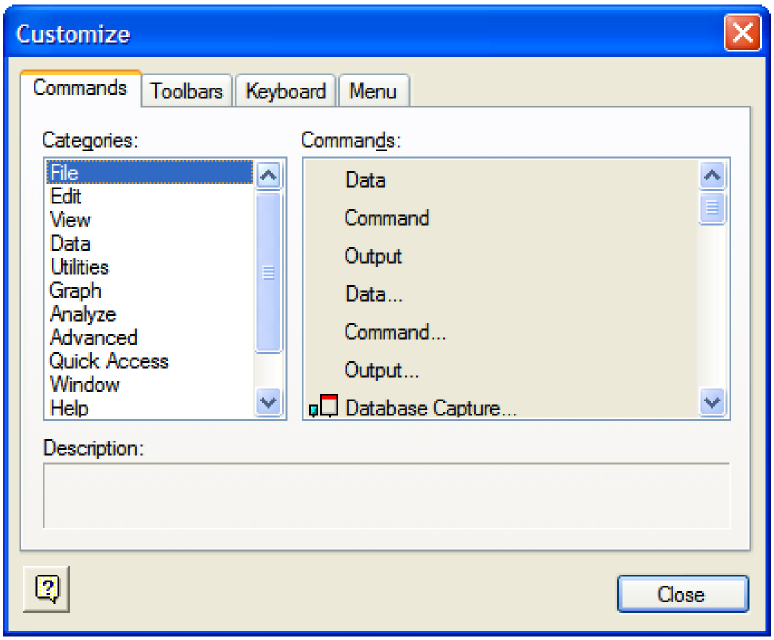
Alternatively, right-click in the Toolbar area and select Customize.

The four tabs in the Customize dialog box can be used to customize menus (including right-click or context menus), toolbars, and keyboard shortcuts. A context menu is also available to customize menu items and toolbar buttons, as long as this dialog is open.

Commands Customization

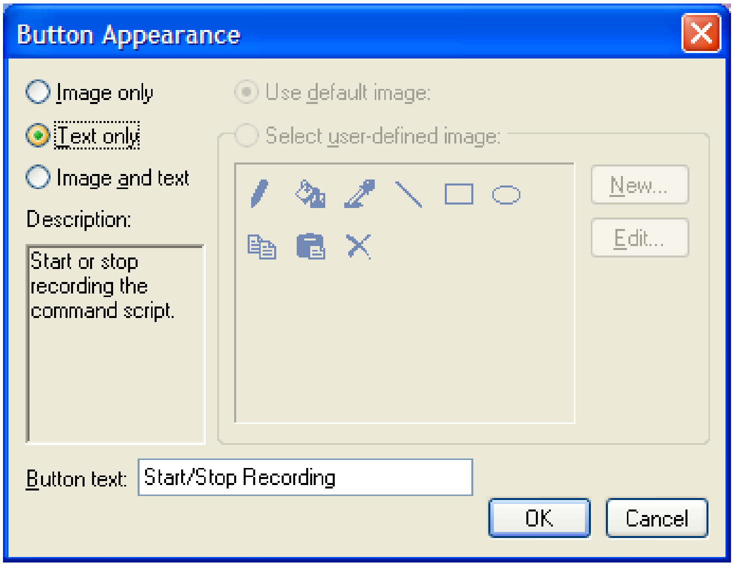
Any menu, menu item within it, or toolbar button can be moved from its default position to any other position either in the menu bar, any menu or in any toolbar. Keep the Customize dialog open or, in the case of toolbar buttons and terminal menu items, hold down the Alt key and drag and drop the item (there will be a border around the item while it is being dragged) to the desired position. To copy an item instead of moving it, hold down the Ctrl key as well. To completely remove an item, just drag it out of the menu and toolbar area. Dragging an item slightly to the right creates a separator before it, while dragging it slightly to the left removes the separator if any. All changes can be reset using the Reset and Reset All buttons in the Toolbar and Menu tabs of the Customize dialog, or the Default Settings link in the SYSTAT program group of the Windows Start Menu.

You can also create new menus, menu items or toolbar buttons by dragging and dropping items from the list of items in the Commands tab of Customize, into the desired menu or toolbar position.



The Categories list contains the names of all the menus and menu items. Clicking any of these displays the corresponding menu items in the Commands list. Now, all you need to do is to drag and drop items from this list to the desired position. If you are not sure what a particular item here corresponds to, select it to view a description of the item in the Description area.

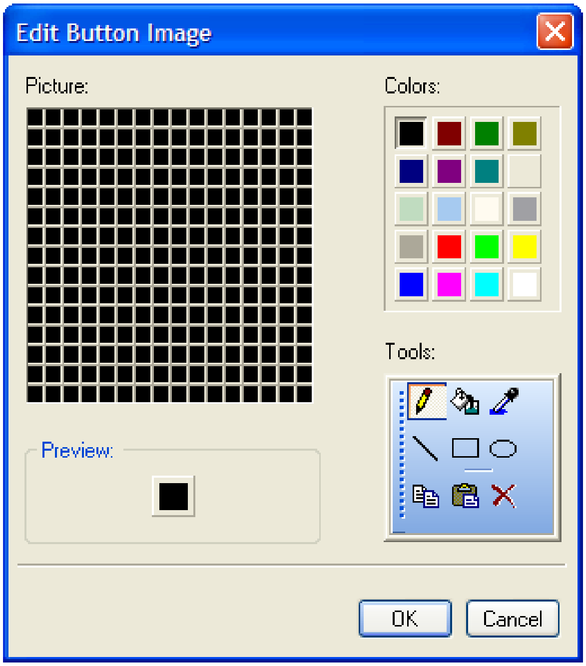
Items that have images preceding their names will be displayed as buttons with the images on them, whereas the Button Appearance dialog pops up when you drop items that do not.



Three choices are available:

- **Image only.** The image that you select from the Image area will be displayed.
- **Text only.** The button will only have a caption. Use the default button text that is displayed in the Button text area, or enter your own text.
- **Image and text.** Both the image that you select and the desired text will appear.

For the first and third options, you can also create your own image or edit an existing one in the Image area. Just press New or select an existing image and press Edit, to invoke the Edit Button Image dialog box.



Use any of the colors shown in the palette, and any of the tools in the Tools area, to create an image in the Picture area. The Picture area is split into pixels arranged in 16 rows by 15 columns. Clicking in the Picture area using any of the tools, colors the pixels in various ways:

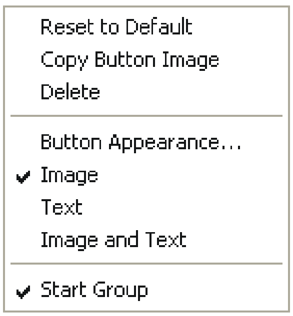
- **Pencil.** Fills any pixel that you click on, with the color selected in the Colors area.
- **Fill.** Fills the enclosed area (with an unbroken boundary made of a non-default color) in which you click, with the selected color.
- **Color selection.** Reads the color of the pixel that you click on, and automatically selects that color in the Color area.
- **Line.** Draws a line of the selected color along the pixels over which you press and drag the pointer.
- **Rectangle.** Draws a rectangle of the selected color, the line over which you press and drag the pointer being the diagonal.
- **Ellipse.** Draws an ellipse of the selected color, the line over which you press and drag the pointer being the diagonal.
- **Copy.** Copies the image in the Picture area to the clipboard.

- **Paste.** Pastes the image in the clipboard to the Picture area.
- **Delete.** Clears the image in the Picture area.

When you press OK, the image will be displayed in the User-defined image area. Press OK to use it, or press Edit to edit it further.

Button Customization

The option to edit button appearance is also available for items in the Commands list that have default images. In fact, you can edit the button appearance and also do a lot more for any menu, menu item or toolbar button. (A menu item is virtually a button with text.) Simply right-click on the desired button when the Customize dialog is open. The following context menu pops up:



Using this menu, you can:

- **Reset to Default.** Resets the button appearance to its default state. The default state for menu items without default images is the text displayed in the Commands list.
- **Copy Button Image.** Copies the button image to the clipboard. You can then paste this in the Picture area while creating new images.
- **Delete.** Deletes the button. Alternatively, you can simply drag a button out of the toolbar area to delete it. Note that, if you delete default buttons, you can only retrieve them by pressing the Reset or Reset All buttons in the Toolbar and Menu tabs of the Customize dialog.
- **Button Appearance.** Pops up the Button Appearance dialog. Use it as explained above to customize the selected button.
- **Image, Text or Image and Text.** Sets the button appearance to show the specified image alone, text alone or both image and text.

- **Start Group.** Inserts a separator before the selected button. This is equivalent to dragging the button slightly to the right.

Toolbars

SYSTAT offers over 250 buttons categorized into 32 default toolbars, to provide immediate access to most tasks. Since showing all of these buttons or toolbars would greatly diminish the area available for output and commands, only six default toolbars with functionality designed to appeal to most users are set up to show in the user interface during the installation of SYSTAT. The default buttons on each of the five default toolbars are:

- **Menu Bar.** File, Edit, View, Data, Utilities, Graph, Analyze, Advanced, Quick Access, Addons, Window, and Help.
- **Standard.** New, Open, Save, Save All, Cut, Copy, Paste, Undo, Redo, Print, Print Preview, Full Screen Viewspace, View/Hide Workspace, View/Hide Commandspace, Customize, Recent Dialogs, Submit from File List, Start/Stop Recording, Play Recording and Help.
- **Format Bar.** Font, Font Size, Block Format, Bold, Italic, Underline, Font Color, Outdent, Indent, Align Left, Align Center, Align Right, Insert Image and Font Dialog.
- **Data Edit Bar.** Variable name, Row number, and Value of the variable at that row.
- **Graph.** Bar Chart, Line Chart, Pie Chart, Histogram, Box Plot, Scatterplot, SPLOM, Function Plot, and Map.
- **Statistics.** Column Statistics, Two-Way Tables, Two Sample t-Test, ANOVA: Estimate Model, Design of Experiments Wizard, Correlations, Least-Squares Regression, Classical Discriminant Analysis, and Nonlinear: Estimate Model.

The Format Bar and two more toolbars, namely Data Edit Bar and Graph Editing, are embedded in the Output editor, Data editor and Graph editor tabs respectively. The Data and Graph Editing toolbars have the following buttons:

- **Data.** Variable Properties, Add Empty Rows, Insert Variable(s), Delete Variable(s), Insert Case(s), Delete Case(s), Find Variable, Go To, First Selected Case in Column, Previous Selected Case in Column, Next Selected Case in Column, Last Selected Case in Column, and Invert Case Selection.
- **Graph Editing (Classic mode).** Copy Graph, Graph View, Page View, Text Tool Font, Drawing Attributes, Pointer Tool, Draw Line, Draw Polyline, Draw Arrow,

Draw Rectangle, Draw Circle, Draw Ellipse, Text Tool, Pan, Zoom In, Zoom Out, Zoom Selection, Reset Graph, Realign Frames, Graph Tooltips, Highlight Point, Region Selection, Lasso Selection, Show Selection and Invert Case Selection.

- **Graph Editing (DirectX mode).** Copy Graph, Graph View, Page View, Format Painter, Pointer Tool, Pan, Zoom In, Zoom Out, Reset Graph, Realign Frames, Graph Tooltips, Highlight Point, Region Selection, Lasso Selection, Show Selection and Invert Case Selection.

One or more of these buttons can be deleted and new ones can be added as described previously, but the toolbars themselves cannot be deleted. They can however be closed. The Format Bar, Data and Graph Editing toolbars can be closed by right-clicking on the corresponding tabs, and unchecking 'Show Toolbar'; repeat the same steps to display them again. The Data Edit Bar can be closed by right-clicking on the Data editor and unchecking 'Show Data Edit Bar'; repeat the same steps to display it again. Other toolbars can be displayed or closed using the Toolbars tab of the Customize dialog or the View -> Toolbars menu.

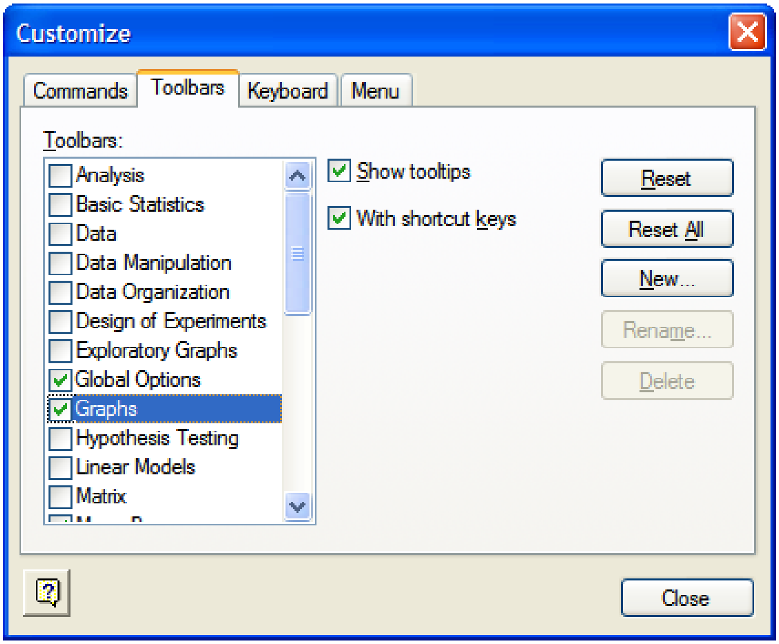
Positioning Toolbars

Toolbars can be docked to pane borders or left "floating" in front of the user interface. To move a toolbar, click the handlebar at the left or top and drag the toolbar to the new location.

- Dragging a toolbar to the left or right side of a pane that is in the docked state attaches or docks the toolbar vertically to that side.
- Dragging a toolbar to the top or bottom of a pane that is in the docked state attaches or docks the toolbar horizontally.
- Dragging a toolbar anywhere other than window borders creates a detached, floating toolbar. Alternatively, you can hold down the Ctrl key while dragging to prevent toolbar docking. Clicking the in the upper right corner closes floating toolbars.

Toolbar Customization

The Toolbars tab of the Customize dialog enables you to close or display SYSTAT toolbars, as well as create new toolbars.



The Toolbars list contains the names of the available toolbars prefixed by check boxes. Notice that the Menu Bar, Standard, Graph and Statistics are checked (by default) , and also that Menu Bar cannot be unchecked. To close a toolbar except the Menu Bar, simply click on the checkmark to uncheck its name. Likewise, to display a toolbar, check the corresponding name in the list.

Apart from making use of the 32 built-in toolbars, you can create your own toolbars. Press the New button, enter the desired name, and press OK. The toolbar appears in front of the dialog. Drag it to the desired location or leave it floating in front of the interface. Drag and drop the desired menu, menu items, or toolbar buttons, from other toolbars or the Commands list in the Commands tab, into the new toolbar.

- To reset any toolbar to its default state, select its name in the Toolbars list, and press the Reset button. To reset all toolbars, just press the Reset All button.
- To rename or delete a toolbar that you have created, press the Rename or Delete buttons respectively.

The Toolbars tab also offers optional button appearance features:

- **Show tooltips.** Displays the button name when the mouse pauses on a button.

- **With shortcut keys.** Displays the shortcut key sequence to be pressed to invoke the same feature, along with the button tooltip.

Keyboard Shortcuts

Although SYSTAT runs in a Windows environment, many users find manipulating the mouse to be an annoyance. Fortunately for these users, every menu item can be accessed using the keyboard.

The F10 key activates the File menu. Once activated, use the arrow keys to navigate through the menu system. The up and down arrows scan vertically through the active menu. The left and right arrows open submenus or move between menus. Use Enter to execute a selected item.

SYSTAT also offers shortcut and access keys for keyboard control of the SYSTAT interface.

Shortcut (Accelerator) Keys. In general, shortcut keys involve holding down the Ctrl key with a single letter to perform a specific task. Most shortcut key combinations appear on the menus after the equivalent entry. Shortcut key behavior may depend on the active window. For example, Ctrl + P prints the content of the Output Editor if it is active, but prints a graph if the Graph Editor is active. The following shortcut keys are available:

Pane/Tab	Shortcut Key	Function
(Any)	Ctrl + N	create a new file in the active tab
	Ctrl + O	open a file in the active tab
	Ctrl + I	open data file.
	Ctrl + Shift + I	import a data file from a database.
	Ctrl + S	save the content of the active tab
	Ctrl +Alt + S	save all open files.
	Ctrl + D, Ctrl + E	save current data
	Ctrl + Q	quit the SYSTAT application.
	Ctrl + X	cut selection, placing contents on the clipboard
	Ctrl + C, Ctrl + Insert	copy selection to the clipboard
	Ctrl + V, Shift + Insert	paste clipboard contents at the current location
	Del	delete the current selection
	F6	invoke the Global Options dialog.
	Ctrl + 0	launch a full screen view of the Viewspace.

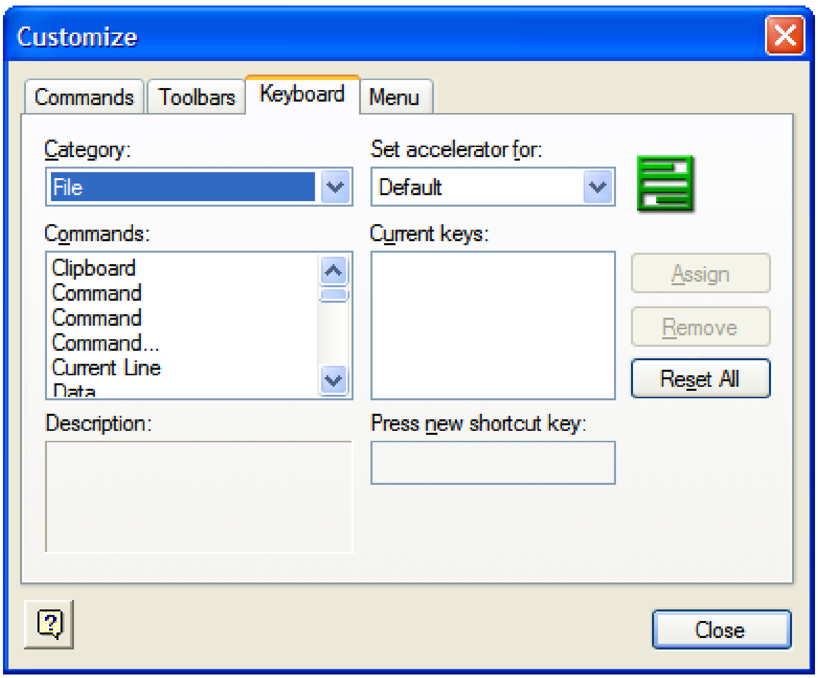
	Ctrl + Shift + O	activate the Output Editor
	Ctrl + Shift + D	activate the Data Editor
	Ctrl + Shift + G	activate the Graph Editor
	F4	invoke the Customize dialog.
	Ctrl + G	invoke the Graph Gallery
	Ctrl + Alt + F	invoke the Graph: Function Plot dialog.
	Ctrl + 1	activate the Workspace.
	Ctrl + 2	activate the Viewspace.
	Ctrl + 3	activate the Commandspace.
	Ctrl + Tab	move the focus between the three spaces of the user interface. This shortcut will not cycle between the three tabs of the Commandspace.
	Ctrl + Alt + Tab	cycle forward (to the right) through the tabs of the active space.
	Ctrl + Alt + Shift + Tab	cycle backward (to the left) through the tabs of the active space.
	Ctrl + Home	move the cursor to the top of the active tab.
	Ctrl + End	move the cursor to the end of the active tab.
	F10	activate the File menu
	Esc	closes an open dialog box
Output Editor	Ctrl + Shift + Alt + P	specify the printer, paper size, source and orientation to be considered while printing.
	Ctrl + Alt + P	preview the content of the Output Editor before printing.
	Ctrl + P	print the content of the Output Editor.
	Ctrl + Z, Alt + Backspace	undo step by step, a few steps of editing done
	Ctrl + Y	redo step-by-step, a few steps of editing done
	Ctrl + F	find text.
	F3	find the next instance of the text specified for the search.
	Ctrl + H, Ctrl + R	replace text.
	Ctrl + A	select entire contents of the active tab.
	Ctrl + Shift + F	set the font of subsequently typed (not generated) or selected text in the Output Editor.
Data/Variable Editor	Ctrl + Shift + Alt + P	specify the printer, paper size, source and orientation to be considered while printing.
	Ctrl + Alt + P	preview the data/variable information before printing.

	Ctrl + P	print data/variable information.
	Ctrl + Z, Alt + Backspace	undo step by step, upto 32 steps of editing done.
	Ctrl + Y	redo step-by-step, upto 32 steps of editing done.
	Ctrl + F	locate a variable in the Data Editor.
	Ctrl + H, Ctrl + R	replace instances of a string in a given column.
	Ctrl + A	select entire contents of the active tab.
	Alt + Insert	add empty rows in the Data Editor (appends at the end of a file if one is already open).
	Ctrl + Shift + Insert	insert variables in the Data Editor before or after a selected column.
	Ctrl + Shift + Del	delete the selected variables in the Data Editor.
	Ctrl + Shift + P	open Variable Properties for the current column
	Shift + Del	cut the selected variable or case
Graph Editor	Ctrl + P	print the graph that is in the Graph Editor.
	Del	delete any annotation that you may have created
Commandspace	F7	submit the contents of the active tab in the Commandspace.
	Ctrl + L	submit the command line on which the cursor is currently positioned.
	F8	submit the selection in the active tab of the Commandspace.
	Ctrl + F7	submit a command file.
	Ctrl + Shift + V	submit the contents of the clipboard.
	Ctrl + Shift + Alt + P	specify the printer, paper size, source and orientation to be considered while printing.
	Ctrl + Alt + P	preview the output before printing.
	Ctrl + P	print data.
	Ctrl + Z, Alt + Backspace	toggle between undoing and redoing the last step of editing.
	Ctrl + Y	redo the step that was last undone.
	Ctrl + F	find text
	F3	find the next instance of the text specified for the search.
	Ctrl + H, Ctrl + R	replace text
	Ctrl + A	select entire contents of the active tab.
	Ctrl + Shift + F	set the font to be used in the active tab.
	F9	recall commands from the command buffer one-by-one starting from the latest.
	Ctrl + W	toggle visibility of Commandspace

Access keys. Access keys provide an alternative to accelerator keys for accessing menu entries. Access keys open menus using the Alt key and allow navigation to selected entries using designated letters.

- The name of each menu contains one underlined letter. Pressing Alt and the underlined letter opens the corresponding menu. After opening a menu, you can execute any of the displayed entries.
- Like the menu titles, each menu entry contains one underlined letter. Pressing this letter runs the entry as if it had been selected using the mouse.

The list of access keys is too long to be displayed here. To view the key required for a particular menu entry, open the menu and scan through the underlined letters. You will quickly become familiar with the procedures and graphs you use frequently.



Keyboard Shortcut Customization

The default keyboard shortcuts may be changed and new keyboard shortcuts can be defined using the Keyboard tab of the Customize dialog.

Category. Lists all the menus in the Menu Bar, and one entry for all commands put together.

Commands. Lists all the menu items under the menu selected in Category. Select a command to see its description in the Description area.

Current keys. Displays the keyboard shortcut(s) already assigned (either by SYSTAT or by you) to the command selected in Commands. If you do not want to use an existing keyboard shortcut key, select it and press the Remove button to remove the assignment. To reset keyboard shortcuts for all commands to their default assignments, press Reset All.

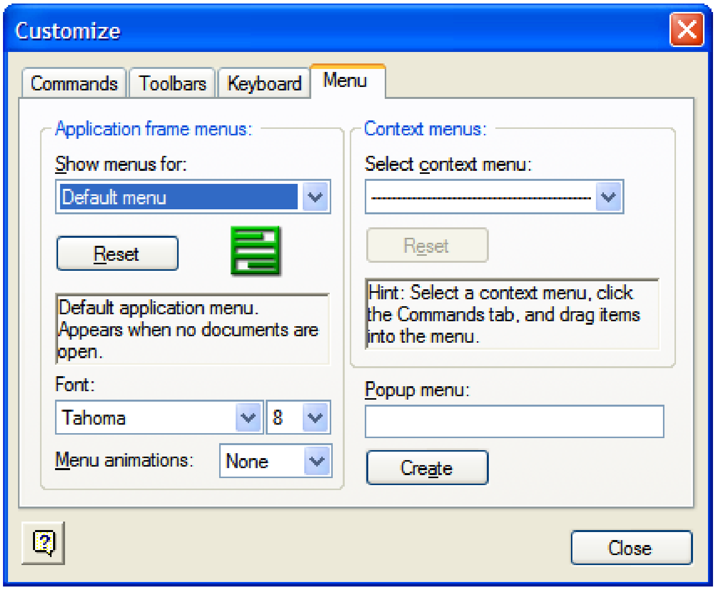
Press new shortcut key. Press the desired shortcut key or key combination for the selected command. The key name will be automatically displayed in this area as you press it. Key combinations will have to begin with Shift, Ctrl, Alt, or any combination of these, and end with one other key. When you are satisfied with the key combination you have typed, press Assign. You can define more than one keyboard shortcut for a command.

If a key combination you have typed in the new shortcut key area has already been assigned to some other command, then that command will be displayed in the Assigned to area, and the Assign button will be disabled. Also, the new shortcut key area will not register any external keyboard shortcuts, since such shortcuts may also be useful while working with SYSTAT. (In fact, pressing such shortcuts will perform the associated external task.) For instance, Alt + Tab is a Windows shortcut that lists all open windows, allowing you to select one by holding Alt down and repeatedly pressing Tab. This functionality offers quick navigation between the SYSTAT user interface and any other program you may be running concurrently.

Access Key Customization. The access key for a menu item is indicated by typing an ampersand before the underlined letter, in the Button text area of the Button Appearance dialog box. You can change the access key to use, by moving the ampersand to be just before the desired letter in the caption. Take care to see that you do not create duplicate access keys.

Menu Customization

SYSTAT has several context menus that pop up on right-click in various parts of its user interface. Use the Menu tab of the Customize dialog box to customize these menus, as well as set a few other options.



Reset. The default menu structure of SYSTAT may be modified according to the user's preferences and needs, as described earlier. Use the Reset button to reset the menu structure to its default state.

Context menus are available for the Startpage, Output Editor, Data Editor (columns, rows and cells), Graph Editor, Output Organizer (data, view data, graph, other, and main), Examples (folder and node), Interactive, Batch, and Log tabs of the Commandspace, and status bar. To customize a context menu, select it from the drop-down list (or right-click in the associated pane) so that it pops up. Customize it as you would customize any other menu or toolbar. If you drag and drop toolbar buttons, the associated text is automatically displayed (you cannot display only button images here). Any changes are immediately applied. Press the Reset button in the Context menus group to reset the selected context menu to its default state. Press the Close  button at the top right corner or close the Customize dialog to close the popped up menu.

Font. Select the desired font and font size to be used for all the menu items.

Menu animation. By default, all SYSTAT menus pop-up immediately on click. You may choose to leave it that way or use one of the two available animation effects: Unfold and Slide.

Select context menu. Select the context menu that you want to customize. Press Reset to reset any changes you may have made to the selected context menu to the installation default.

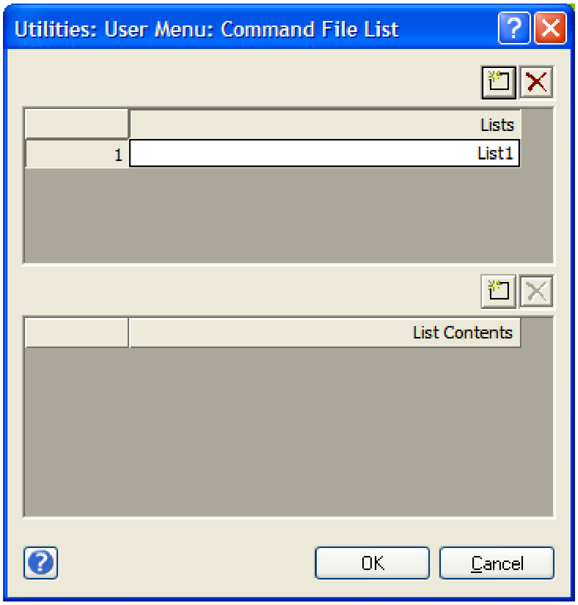
Popup menu. Use this to create new popup menus in the Menu Bar. Enter the name of a the popup menu and press Create. The new menu gets added as the first item in the Menu Bar. Drag and drop the menu to whatever location you want it to be in.

Command File Lists

Command files can be saved in any folder. If you elect to organize your files by projects, each folder will most likely contain data, output, and command files. This approach groups related command files together, but may result in similar files appearing in several project folders. On the other hand, you can store files by type, resulting in a single folder containing only command files. In either situation, finding a particular command file can be a difficult task. The Command File List dialog provides a command file classification scheme that is independent of your folder structure. Using this dialog box, you create lists of command files having some element in common, such as "Charts with Error Bars". A list can then be associated with the Submit From File List toolbar button or menu item for immediate processing of any file contained therein.

To open the Command File List dialog box, from the menus choose:

Utilities
  User Menu
    Command File List…



Lists. Displays all defined command file lists. Select a list to view the names of all command files assigned to the list, in the List Contents list. You can define lists or remove defined lists as described below. Once you do that, select a list to assign it to the Submit From File List button and menu item; SYSTAT automatically links the two. You can change the list assigned to the toolbar button by selecting a different list at any time.

List Contents. Displays the names of the command files assigned to the selected list. You can assign files to or remove assigned files from the list. For example, suppose you have a file in C:\Folder1 that produces a plot of residuals against predicted values and another file in D:\Folder2 that produces a probability plot of residuals. You can assign both files to a list called "Regression Diagnostics".The only condition is that the files should be text-based.

Modify the index of command file lists or the contents of any list using the two customization tools. For the index of command file lists, these buttons have the following functions:

- **Insert Row.** Creates a new command file list. Alternatively, right-click in the Lists header and select Insert Row. Once a row is created, you can even press the Enter key to create more rows. After inserting a row, type a name for the new list.The default name is set to List1.You can replace it by a suitable name.The name should be unique. Click on the row and press the Delete key if you want to clear a name. Press the Enter key or click outside the row to assign the name to the new list.

- **Delete Row.** Deletes the selected list. Alternatively, right-click on the list and select Delete Row.

For the set of command files in a list, the two buttons have the following functions:

- **New.** Adds a file to the selected list. When adding a file to a list, press the ellipsis button at the right of the new entry to browse for a particular file. Alternatively, type the path and filename into the list of command files. SYSTAT automatically appends the currently defined path for command files to any typed filenames without a path.
- **Delete.** Deletes the selected command file from the list. The command file is deleted from the list only; the file is not deleted from the user's system.

Submission From File Lists

In addition to offering a mechanism for organizing files, command file lists also allow submission of the files contained in the lists. As a result, you can create templates for custom graphs, assign them to a file list, and apply them to the current data via a mouse click.

Use the Submit from File List button on the Standard toolbar to submit files from previously defined command file lists.

Alternatively, from the menus choose:

File
  Submit
    From Command File List...

This presents the names of all files in the command file list that is currently selected in the Command File List dialog. The display contains only the filename, not the path. As a result, some lists may contain multiple entries with the same name, but which invoke different command files. Using unique names for command files avoids this potentially confusing situation.

Selecting a file from the displayed list submits the corresponding file for processing. The commands contained in the file do not appear on the middle tab of the Commandspace; file submission does not affect this tab. As a result, you can have a command file open and submit a second file at the same time.

Command file lists and the list of recent command files appearing on the File menu offer similar functionality, but differ in several notable ways. First, command file lists allow you to group your files into categories, whereas file lists based on recency of use do not. Second, you can create multiple command file lists, each having an unlimited number of entries. The recent command list allows only nine entries. Third, the structure of command file lists persists across sessions, but lists of recent files change each time you open a file. Finally, command file lists submit the selected file for processing. The recent file list merely opens the file on the middle tab of the Commandspace.

Recent Dialogs

SYSTAT provides quick, easy access to frequently used dialog boxes. Every time you use (invoke and execute) a dialog from the Data, Graph, Analyze, Advanced or Quick Access menus, or even from the corresponding DIALOG command, it is added to the list of recently used dialog boxes. This list persists across SYSTAT sessions, so if you consistently use the same set of dialog boxes, they are always just a click away. Simply click the Recent Dialogs button  on the Standard toolbar, or from the menus choose:

Utilities
  Recent Dialogs...

Selecting an item from the list presents the corresponding dialog box. All options and variable lists in the recalled dialog box reflect your specifications from the last use of that dialog. However, opening a different data file changes the variables available for an analysis and consequently resets all dialog boxes to their default settings.

SYSTAT automatically updates the list of dialog boxes during your sessions. The list contains up to fifteen dialog boxes, ordered according to recency of use. Each use of a dialog box results in a corresponding entry at the top of the Recent Dialogs list. Any other instance of that dialog in the list is removed. As a result, no dialog box appears in the list more than once. If your list contains fifteen entries and you use a dialog box not appearing in the list, SYSTAT adds the new dialog to the top of the list and removes the oldest entry.

Some main dialog boxes require preliminary results before they can be used. For instance, the Hypothesis Test dialog can only be used after estimating a model successfully. These contingent dialogs do appear in the Recent Dialogs list, but are removed each time a data file is opened.

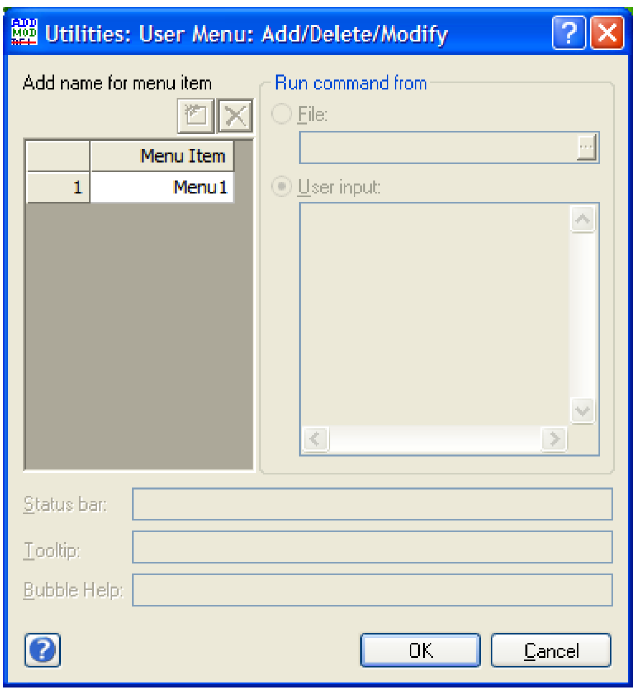
Although the goal of Recent Dialogs is to present the most recently used dialogs, some main dialogs do not appear in the list. The Variable Properties and Add Empty Rows dialog boxes, for example, do not receive list entries. Furthermore, wizards that result in a sequence of dialogs only receive an entry for the first dialog of the sequence.

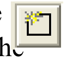
Note: Because most dialog boxes require variable specifications, Dialog Recall is disabled if there is no open data file.

User Menus

SYSTAT's menus offer a dialog interface to most of the underlying command language. You can also create an additional menu with entries designed to process sets of commands that you frequently run. To add a user menu item, from the menus choose:

Utilities
  User Menu
    Add/Delete/Modify…



Menu item. Displays all the menu item names that are currently defined. Use the  and  buttons to insert new items and delete unwanted items respectively. The names in this list will be displayed under the Menu List sub-menu of User Menu. You can define any number of menu items here, but the Menu List will display the first 30.

You have to associate each menu item you define to either of the following:

File. Displays the SYSTAT command filename, if any, associated with the currently selected menu item name. To specify a different filename or when you are defining the menu item for the first time, type the name of a command file including its path or press the  button and browse for it.

User input. Displays the set of commands, if any, associated with the currently selected menu item name. Edit existing commands or type a new set of commands just as you would in the Commandspace.You may want to type one or more DIALOG commands here that would pop up frequently used dialog boxes, or a command template that you could apply on various data files.

Status bar. Displays the status bar help content currently associated with the selected menu item.You can edit existing content or type new content.

Tooltip. Displays the tooltip that will appear on mouse hover if the selected menu item is placed on a toolbar. You can edit an existing tooltip or type a new one.

Bubble Help. Displays the Bubble help content currently associated with the selected menu item.You can edit existing content or type new content.

An alternative way of creating a user menu item is by using the Record Script feature. This feature automatically creates a menu entry if you request it to do so, and associates it with the command scripts it has just recorded. You can see the menu item list, and the recorded set of commands when you open the User Menu Profile dialog subsequently. For more information about this feature, see *Command Language*.

To access a menu item created using the Add/Delete/Modify dialog or Record Script feature, from the menus choose:

Utilities
  User Menu
    Menu List...

and, under this, the corresponding menu item name. Clicking the name will execute the underlying set of commands.

Keyboard shortcuts. Any user menu item can be accessed using the keyboard by pressing the underlined number preceding its name (the full sequence would be ALT + U, U, L, the underlined number).

Themes

The themes feature of SYSTAT allows you to create, store and apply any number of fully customized interface themes each with its own set of menu items and toolbars as well as the position and size of spaces, content of the status bar, and keyboard shortcuts. These will be very useful if you do not need some of the menu items at all. If you are comfortable with a different menu arrangement or terminology, work with just a subset of all the data processing, analyses and graphing techniques available in SYSTAT, or work with one of several sets of features that you will need at various times. For instance, if you conduct various courses in Statistics starting from a basic course to an advanced one, execute projects catering to various industries, or do research in various application areas like Psychology, Engineering or Chemistry, you may create one theme for each case and apply the appropriate theme as required.

You can save the changes you make to the default theme or any existing theme of SYSTAT in a theme file. To do this, from the menus choose:

Utilities
  Themes
    Save Current Theme…

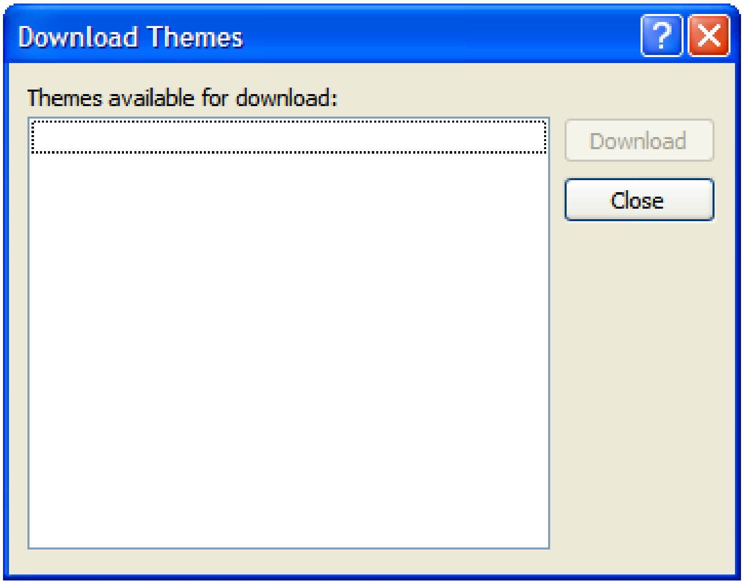
In the dialog that pops up, enter a suitable file name, and press Save. All menu items, status bar content, toolbar layout and location, as well as those of the Workspace, Viewspace and Commandspace will be saved in this file. By default, the file will be saved to the Themes folder of SYSTAT. You may specify a different folder to save to; the advantage of saving in the Themes folder is that the theme will be listed in the Themes section of the Startpage. The name of the theme will be the same as the filename; you simply have to double-click the desired theme name to apply it. In any case, to apply any stored menu theme, from the menus choose:

Utilities
  Themes
    Apply Theme…

Navigate to your themes folder, select the desired file and press Open.

New themes will be available on the SYSTAT server from time-to-time. To download these, from the menus choose:

Utilities
  Themes
    Download Themes...



In the dialog box that opens, check the themes that you want to install, uncheck the ones that you do not need, and press Download. If you do not want to install themes at this time, press Close.

To revert to the default menu theme, from the menus choose:

Utilities
  Themes
    Apply Default Theme...

Global Options

SYSTAT has a host of global settings that you can customize according to your preferences. These settings are automatically saved at the end of a session, and remain in effect for subsequent sessions. Most of them can also be accessed through the Global Options toolbar or the status bar. To open the Global Options dialog box, from the menus choose:

Edit
  Options…

The six tabs in the Options dialog box control different settings in SYSTAT.

General. Specify general appearance and behavior options.

Data. Specify Data Editor display options.

Output. Specify the general appearance of output.

Output Scheme. Specify font and color for individual components of the output, as well as the background image or color for all of the output.

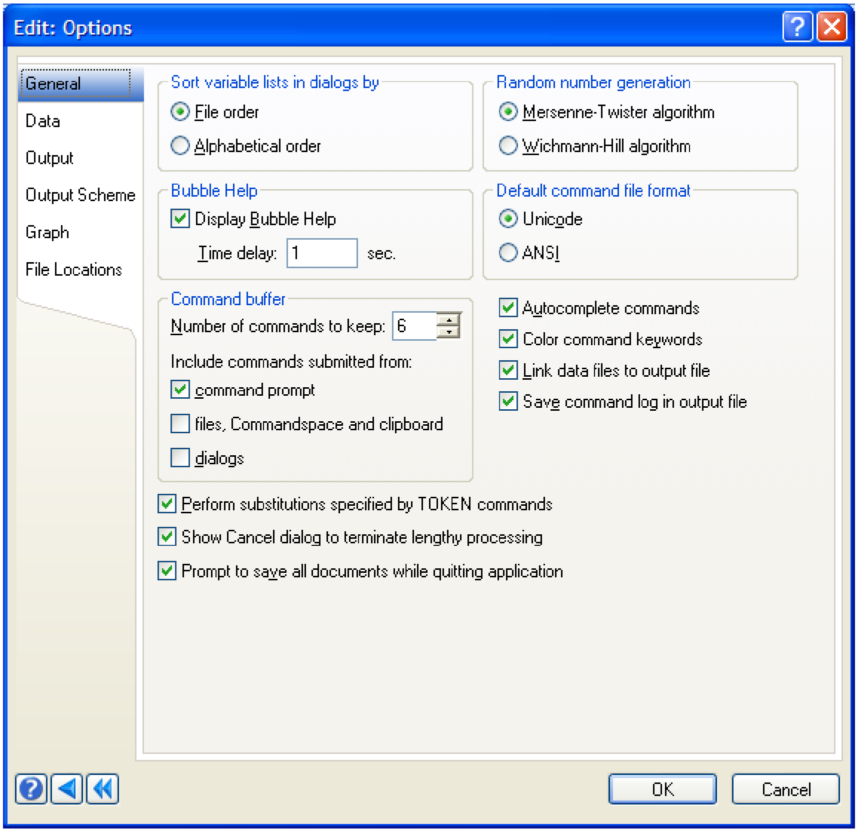
Graph. Specify graph scaling, line thickness, character size, and measurement units for all subsequent graphs.

File Locations. Set folders in which SYSTAT should look for files of different types.

The General, Output, Output Scheme, and File Locations tabs are described here. For information about Data options, see SYSTAT Data. For information about Graph options, see SYSTAT Graphics.

General Options

The General tab of the Global Options dialog controls the ordering of variables in dialog boxes, token processing, and command recall.



Sort variable lists in dialogs by. You can sort source variable lists in dialog boxes by file order or alphabetical order. For data files with a large number of variables, it is often easier to find variables in source lists if the variables are sorted alphabetically. If variables are grouped together in the file for a specific reason, it may be easier to select related groups of variables if the variables are sorted in file order.

Random number generation. SYSTAT provides two algorithms for generating random numbers:

- **Mersenne-Twister.** This is believed to have a far longer period and far higher order of equidistribution than other random number generators. It is the recommended option especially for Monte Carlo studies.
- **Wichmann-Hill.** This generates random numbers by a triple modulo method.

Mersenne-Twister (MT) is the default option. We recommend the MT option, especially if the number of uniform random numbers to be generated for your Monte Carlo exercise is large, say more than 10,000.

If you would like to reproduce results involving random number generation from earlier SYSTAT versions, with old command files or otherwise, make sure that your random number generation option is Wichmann-Hill (and, of course, that your seed is the same as before).

For more details, see Chapter 4 (Data Transformations) of the *Data* volume and user documentation on Monte Carlo if you have the Monte Carlo add-on module.

Bubble Help. Apart from the help provided on the status bar about each menu item, a more detailed description is provided in a "bubble" that appears when you pause the mouse on the menu item for a few seconds.You can specify the number of seconds to pause the mouse before the help appears, or even turn off the help completely.

Default command file format. SYSTAT provides two formats for saving command files. For a given file, you do have the option of saving in the ANSI format using the File type dropdown in the Save File dialog box. The default choice may be set to one of the following:

- **Unicode.** SYSTAT command files will be saved in the unicode format by default.
- **ANSI.** SYSTAT command files will be saved in the ANSI format by default.

Command buffer. The command buffer contains the most recently processed commands. Use this buffer for quick recall, modification, and resubmission of commands using the F9 key. The number of commands to keep defines the size of the buffer; use the up and down arrows to adjust the number of retrievable command lines. The software uses the buffer to store commands generated from any of the following sources:

- **Command prompt.** Commands submitted using the Interactive tab of the Commandspace.
- **Files, Commandspace and clipboard.** Commands submitted from the middle and Log tabs of the Commandspace. This option also includes commands submitted directly from the Windows Clipboard and command files submitted via the SUBMIT command.
- **Dialogs.** Commands generated after clicking the OK button in any dialog. Select this option to use the dialog interface to generate a command line that you expect to refine iteratively.

Autocomplete commands. As you type commands in any tab of the Commandspace, you will be prompted with the possible command keywords, arguments, options, option values, available data files, or available variables. For instance, the data files in the folder specified under Open data in the Global Options dialog will be listed if you type "USE ". This feature is enabled by default. You can turn it off if you do not want commands to be autocompleted.

Color command keywords. By default, in any tab of the Commandspace, SYSTAT displays command keywords in colored font with specific colors denoting specific kinds of keywords. You may uncheck this option if you do not want commands to be colored.

Link data files to output file. When a SYSTAT output file is saved, the data files are linked to the output file. That means you can open an output file saved in a previous session and continue working with it provided the underlying data files exist in the same path. Uncheck this option if you do not want to use output files across sessions.

Save command log in output file. When a SYSTAT output file is saved, the command log will also be saved with it. That means you can open an output file, saved in a previous session, and re-use the commands from that session. Uncheck this option if you do not use output files across sessions.

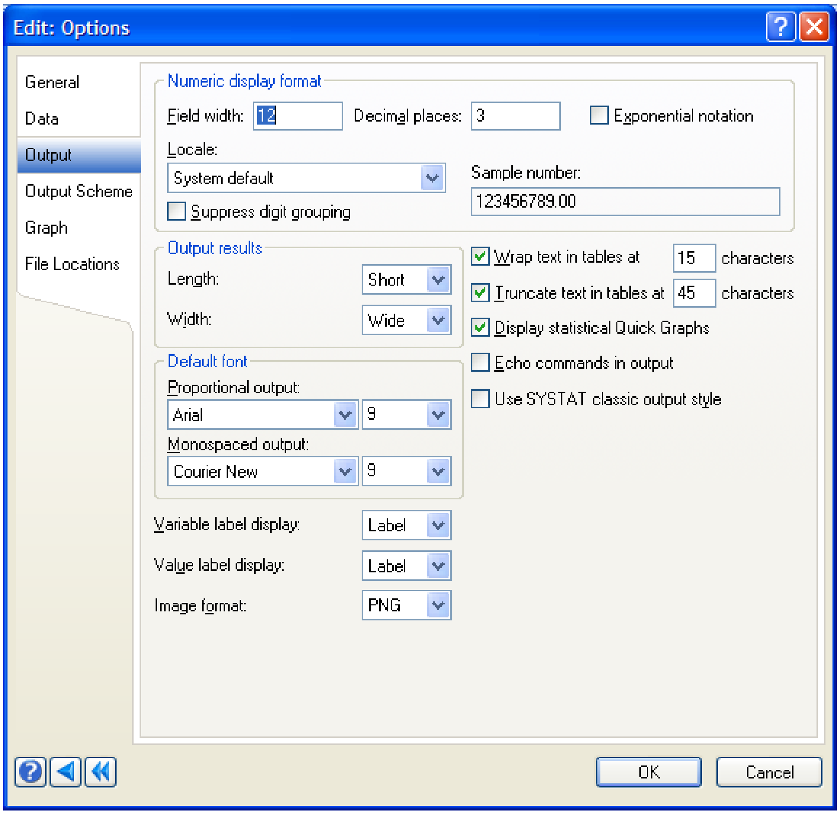
Perform substitutions specified by TOKEN commands. With this option selected, SYSTAT treats the ampersand (&) character as a token indicator. During processing, predefined or user-specified values replace every '&' and the text immediately following it. Deselect this option to prevent these substitutions.

Show Cancel dialog to terminate lengthy processing. Whenever processing by SYSTAT takes some time before results can be displayed, a Cancel dialog pops up so that you can cancel processing. You may want to uncheck this option to avoid accidental cancellation of a process.

Prompt to save all documents while quitting SYSTAT. By default, SYSTAT prompts you to save all open documents, including any new unsaved data and commands that you may have entered, when you quit the application.You may want to uncheck this option when you run the application unattended in the batch mode.

Output Options

The Output tab of the Global Options dialog determines the format and content of subsequently created output.



Numeric display format. These settings control the default display of numeric data in the output. Field width is the total number of digits in the data value, including decimal places. Exponential notation is used to display very small values. This is particularly useful for data values that might otherwise appear as 0 in the chosen data format. For example, a value of 0.00001 is displayed as 0.000 in the default 12.3 format but is displayed as 1.00000E-5 in exponential notation. A number that would otherwise violate the specified field width will also be converted to exponential notation while maintaining the number of decimal places. Individual variable formats in the Data Editor override the default setting.

Locale. SYSTAT determines the initial default decimal and digit grouping symbols for numbers from the current settings in the Regional and Language Options dialog of the Windows Control Panel. This is recognized as the System default. You may change the setting to any of the locales provided in the dropdown list. A sample number will be displayed alongside. You may suppress digit grouping if you do not want digits to be grouped.

With this option, you will be able to enter numbers in the Data Editor using the decimal and digit grouping symbols of your chosen locale. The output displayed in the Output editor will also adhere to these locale specific settings. You can thus create output suitable for any given locale.

Output results. These settings control the display of the results of your analyses.

- Length specifies the amount of statistical output that is generated. Short provides standard output (the default). Some statistical analyses provide additional results when you select Medium or Long. Note that some procedures have no additional output. (*Tip*: In command mode, DISCRIM, LOGLIN, and XTAB allow you to add or delete items selectively. Specify PLENGTH NONE and then individually specify the items you want to print.)
- To control Width, select Narrow (80[77 (82) characters wide in the HTML (Classic) format, for a font size of 10],) or Wide (132[106 (113) characters wide in the HTML (Classic) format, for a font size of 10]), or None. This applies to screen output (how output is saved and printed). The wide setting is useful for data listings and correlation matrices when there are more than five variables. Selecting None prevents tables from splitting no matter how wide they are.
- To control Width, select Narrow (80 characters wide) or Wide (132 characters wide). This applies to screen output (how output is saved and printed). The wide setting is useful for data listings and correlation matrices when there are more than five variables.

Default font. You can specify the font used in the output.

- Proportional output sets the font and font size for the HTML based output.
- Monospaced output sets the font and font size for output appearing in the classic style, and any output requiring fixed-width font (that facilitates automatic alignment of text) like stem-and-leaf diagrams.

Wrap text in tables. The text written in tables can be sometimes very long, especially when variable and/or value labels are defined. In such cases, by default, in each cell, the text will be wrapped into multiple lines if they extend beyond 15 characters. Row headers will be wrapped if they extend beyond thrice this number, i.e., 45 characters.

You can set a different number here as desired. You can even uncheck this option to prevent wrapping.

Truncate text in tables. Apart from wrapping, the text in tables can also be truncated. By default, in each cell, the truncation will happen at 45 characters. You can change this number or even turn off truncation.

Display statistical Quick Graphs. You can turn the display of the Quick Graphs on and off. By default, SYSTAT automatically displays Quick Graphs.

Echo commands in output. Includes commands in the Output Editor before the subsequent output.

Use SYSTAT classic output style. Displays all subsequent statistical output as ASCII text using the Courier font. With this option selected, no output appears in formatted tables.

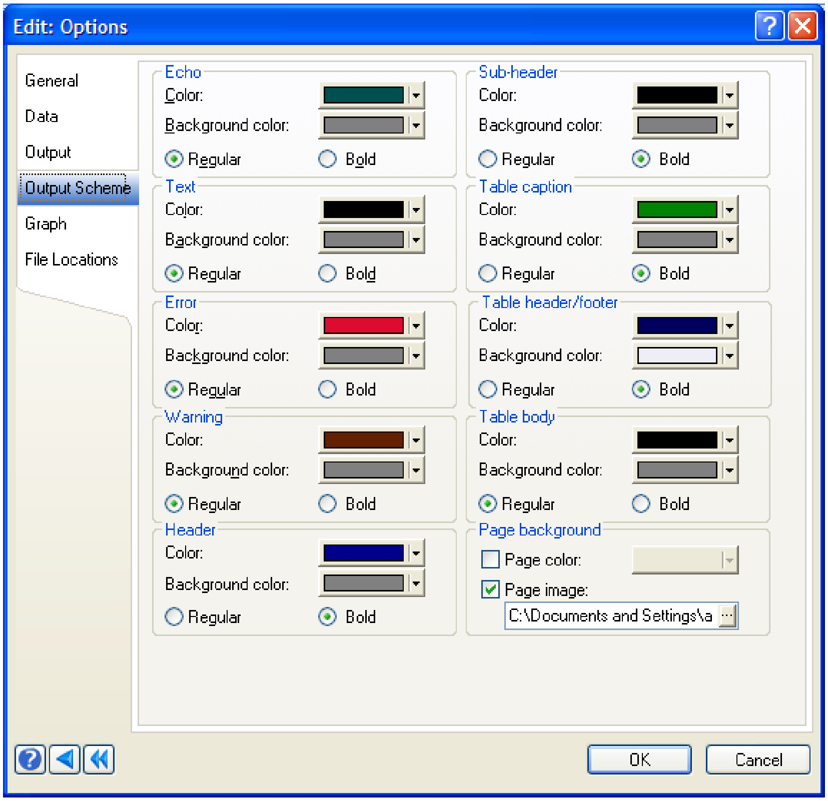
Variable label display. If a variable label is defined for a variable, it will be used to identify the corresponding variable in the output instead of the variable name itself. Select "Both" if you want both variable names and labels to be used, or "Name" if you want just the variable names to be used.

Value label display. If value labels are defined for a variable, they will be used to represent the underlying data values in the output.You can select "Both" to display both value labels and data values, and "Data" to display just the data values.

Image format. The graphs created by SYSTAT in the Output Editor are in the "portable network graphics (PNG)" format. You can choose this or any one of the formats: BMP, JPG, GIF and EMF.

Output Scheme

The Output Scheme tab of the Global Options dialog allows you to customize the output format in terms of the font color, style (regular or bold) and background color of various components of the output (excluding graphs), as well as the page background.



Echo. Specify the font color, style and background color of echoed commands. The default is a shade of teal, in the regular font style with a white background.

Text. Specify the font color, style and background color of all text. The default is black color, in the regular font style with a white background.

Error. Specify the font color, style and background color of error messages. The default is a crimson color, in the regular font style with a white background.

Warning. Specify the font color, style and background color of warning messages. The default is a shade of brown, in the regular font style with a white background.

Header. Specify the font color, style and background color of text headings. The default is a shade of blue, in the bold font style with a white background.

Sub-header. Specify the font color, style and background color of text sub-headings. The default is a shade of blue, in the bold font style with a white background.

Table caption. Specify the font color, style and background color of table captions. The default is a shade of blue, in the bold font style with a white background.

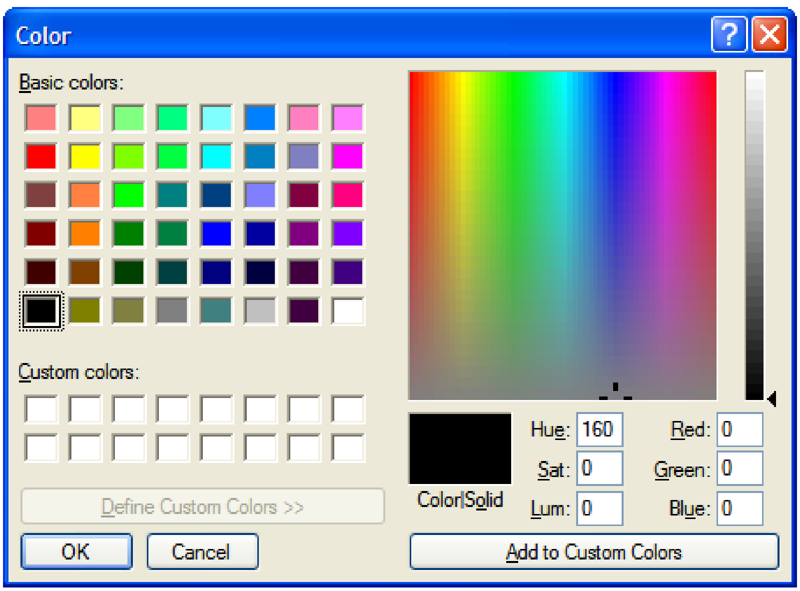
Table header/footer. Specify the font color, style and background color of the text in table headers and footers. The default is black color, in the bold font style with an off-white background.

Table body. Specify the font color, style and background color of the text in table body. The default is black color, in the bold font style with a white background.

Page background. Specify the background color and/or image for the entire page. The image should be stored in the PNG, BMP, JPG, GIF or EMF format, and can be in any location.

Color Palette

To change a color, click the corresponding color button, click on a pre-defined color in the color palette, or create your own color by clicking More colors. Clicking this opens the Color dialog.



Basic colors. Click one of the basic colors and press OK to use that color.

Custom colors. Click a basic color to begin with. It shows up in the Color|Solid area, with the cross-hair at the corresponding point in the full color spectrum above it, and an arrow at the corresponding point in the color bar beside the spectrum. You can move the cross-hair to any point in the full spectrum, and slide the arrow to any height in the color bar. You can also enter hue, saturation, luminosity, and RGB values. Press Add to Custom Colors to add the color to the Custom color palette. You can create any number of colors in this way. Finally, click on a color and press OK to use that color.

File Locations

Use the File Locations tab to specify the folder containing the files used in the Graph Gallery, to designate file paths to append to filenames used in SYSTAT commands, and define paths to store command, graph and output files.

Set project directory. Resets file paths for all file types to the appropriate sub-folders within the designated folder. Check Use common directory if you want all subsequent file opening and saving to occur directly within this folder.

Set custom directories. As an alternative to specifying a project directory, you can specify individual folders based on file type or file operation.

- **Graph Gallery.** Specify the folder containing the command files and graphics used to generate the Graph Gallery.
- **Open data.** Sets the folder used for opening all SYSTAT data files (.SYZ and .SYS). When opening data files using the menus, the Open dialog initially defaults to this folder. This is set to the SYSTAT Data folder at the time of installation.
- **Save data.** Defines the folder used for saving all SYSTAT data files (.SYZ). When saving data files using the menus, the Save As dialog initially defaults to this folder. If a USE command is issued without a path, SYSTAT also looks for the file in this folder. This is set to the SYSTAT Data folder at the time of installation.
- **Work data.** Sets the folder used for saving all temporary data files (.SYZ). If a USE command is issued without a path, SYSTAT also looks for the file in this folder. This is set to the Windows temporary folder at the time of installation.
- **Import data.** Identifies the folder used for all data file importing.
- **Export data.** Identifies the folder used for all data file exporting.
- **Command files.** Sets the folder used for opening and saving of SYSTAT command files. When opening or saving command files using the menus, the dialogs initially default to this folder. This is set to the SYSTAT Command folder at the time of installation.

- **Output files.** Associates the designated folder with all SYSTAT (.SYO) as well as HTML (.MHT) output files (.SYO). When opening or saving output files using the menus, the dialogs initially default to this folder.
- **ASCII output files.** Sets the folder used for saving ASCII output files (.DAT) created using the OUTPUT command.
- **Export graphs.** Identifies the folder used for saving all graphic formats.
- **Basic GET.** Defines the folder used for reading ASCII files (.DAT) using the GET command.
- **Basic PUT.** Defines the folder used for writing ASCII files (.DAT) using the PUT command.
- **Export HTML.** Identifies the folder used for saving all HTML files.
- **Export RTF.** Identifies the folder used for saving all RTF files.

Using Commands

Among the general options, use TOKEN/ON or OFF to switch token substitution on or off.

The following commands specify global output display options:

`FORMAT m,n / UNDERFLOW`	Indicates the format for numeric output.

`DISPLAY SHORT`
`MEDIUM`
`LONG`
	Defines the length of statistical output.

`PAGE NARROW` `WIDE`	Indicates the width of the output.

`VDISPLAY LABEL`
`NAME`
`BOTH`
	Defines the use of variable labels in the output.

`LDISPLAY LABEL`
`NAME`
`BOTH`
	Defines the use of value labels in the output.

`GRAPH`	Includes Quick Graphs generated by statistical procedures in the output. Use GRAPH NONE to suppress Quick Graphs.
`ECHO ON` `     OFF`	Indicates whether to echo commands in the output or not.
`CLASSIC ON` `        OFF`	Controls the appearance of statistical results.
`FPATH` *`path`* `/PROJECT or` *`filetype`*	Specifies a path prefix to append to filenames. If *path* is not specified, all file locations are set to the program folder. If no option is specified, all directories are set to the specified path. PROJECT will set *path* as the root directory under which sub-folders Gallery, Data, Command, and Output will be created.

For the *filetype* in the FPATH statement, specify one of the following: GALLERY, USE, SAVE, WORK, IMPORT, EXPORT, SUBMIT, OSAVE, OUTPUT, GSAVE, GET and PUT.

Chapter

8

Applications

SYSTAT offers applications in the following fields:

- Anthropology
- Astronomy
- Biology
- Chemistry
- Engineering
- Environmental Sciences
- Genetics
- Manufacturing
- Medical Research
- Psychology
- Sociology
- Statistics
- Toxicology

You can find these applications in the online Help. Use the Contents tab of the Help system to access the Application Gallery. In the gallery, you will find sample analyses with their associated commands and menu selections. All relevant data and command files are included.

Anthropology

Egyptian Skulls Data

EGYPTDM data consists of four measurements of male Egyptian skulls from five different time periods ranging from 4000 B.C. to 150 A.D.

Variable	Description
MB, BH, BL, NH	Skull measurements
YEAR	Year of measurement

The data can be analyzed to determine if there are any changes in the skull sizes between the time periods. The researchers theorize that a change in skull size over time is evidence of the interbreeding of the Egyptians with immigrant populations over the years. Because there are four different measurements that characterize skull size, multivariate techniques that allow multiple dependent variables can be used. Dependent variables are the measurements *MB*, *BH*, *BL*, and *NH*. The predictor variable is *YEAR*. Assuming that *YEAR* is a discrete predictor variable, then data can be analyzed using MANOVA. Assuming that there is a linear trend to the change in skull size, then *YEAR* can be treated as a continuous predictor variable.

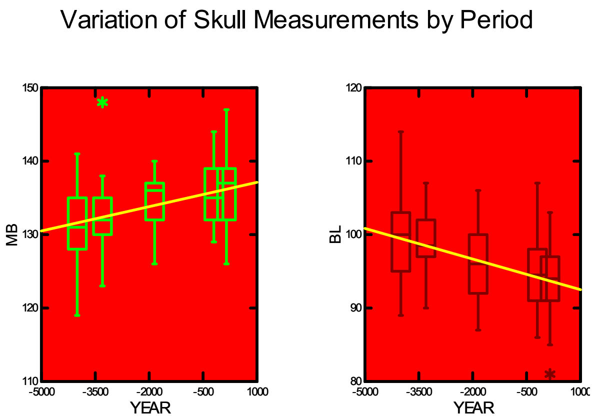
Potential analyses include MANOVA, regression, and principal components.

Box Plot and Regression

The input is:

```
USE EGYPTDM
THICK 2.5
BEGIN
   DENSITY MB BL*YEAR/BOX, FCOLOR=1, FILL=1, XMAX=1000,
             XMIN=-5000, COLOR= {3, 11}, HEIGHT=5.5, WIDTH=4,
             XTIC=4,
             TITLE='Variation of Skull Measurements by Period'
   PLOT MB BL * YEAR / SMOOTH=LINEAR, SIZE=0, XMAX=1000,
                    XMIN=-5000, XTIC=4, COLOR=4, HEIGHT=5.5,
                    WIDTH=4
END
```

The output is:



MANOVA

The input is:

```
PLENGTH SHORT
USE EGYPTDM
MANOVA
   MODEL MB BH BL NH = CONSTANT + YEAR
   ESTIMATE
```

The output is:

```
N of Cases Processed : 150
```

Dependent Variable Means

MB	BH	BL	NH
133.973	132.547	96.460	50.933

Regression Coefficients B = $(X'X)^{-1}X'Y$

Factor	MB	BH	BL	NH
CONSTANT	136.004	131.545	93.901	51.542
YEAR	0.001	-0.001	-0.001	0.000

Information Criteria

AIC	3468.115
AIC (Corrected)	3473.336
Schwarz's BIC	3522.306

Multiple Correlations

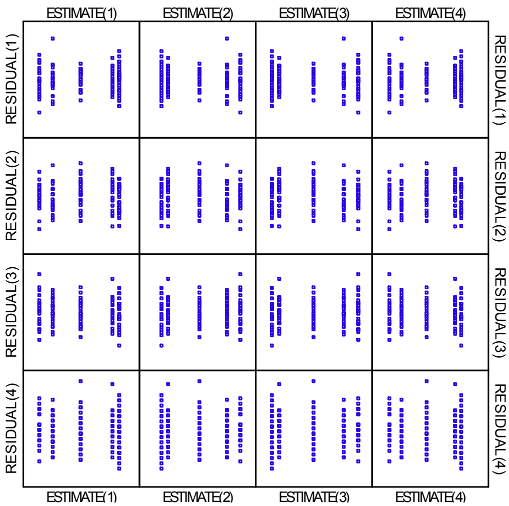
MB	BH	BL	NH
0.371	0.181	0.425	0.170

Adjusted R^2 = $1-(1-R^2)*(N-1)/df$, where N = 150, and df = 148

Adjusted R^2

MB	BH	BL	NH
0.132	0.026	0.175	0.022

Plot of Residuals vs Predicted Values



Astronomy

Sunspot Cycles

SUNSPTDM data consists of a calculated relative measure of the daily number of sunspots compiled from the observations of a number of different observatories.

Variables	**Description**
YEAR	The year the observations were made
JAN-DEC	The relative measure of sunspots for the indicated month
ANNUAL	The mean relative measure of sunspots for the entire year

Sunspots exhibit cyclical behavior on a 10 to 11 year cycle. These cycles have potentially important effects on the earth's ecosystem, including weather and the growth and development of living organisms. Understanding the natural causes and effects of sunspot behavior are all important areas of scientific exploration.

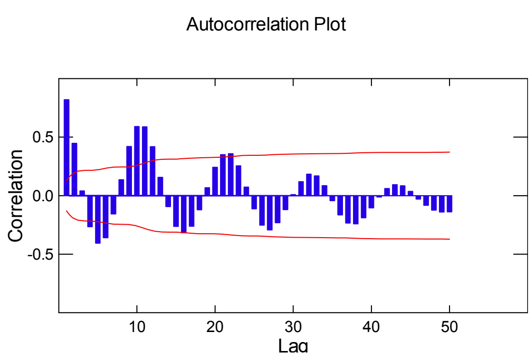
Potential analyses include Time Series (smoothing, autocorrelation, Fourier analysis, ARIMA, etc.) and Descriptive Statistics (variance and distribution).

Autocorrelation Plot

The input is:

```
USE SUNSPTDM
SERIES
   ACF ANNUAL
```

The output is:



Biology

Mortality Rates of Mediterranean Fruit Flies

The *FRTFLYDM* data contains information on mortality rates for Mediterranean fruit flies over 172 days, after which all flies died. Experimenters recorded the number of flies dying each day and divided this by the number alive at the beginning of the day to measure the mortality rate for each day.

Variable	Description
DAY	Day number
LIVING	Number of fruit flies alive at the beginning of the day
MORTRATE	Mortality rate of the fruit flies for each day

The Mediterranean fruit fly data can be used to determine the functional form of mortality rate as a function of time. A scatterplot of these two variables suggests that mortality rate might be a cubic function of time. Since the number of fruit flies alive is directly determined by these two variables, the mortality rate function can be substituted into an equation for the number of fruit flies living as a function of time (which appears to be exponentially decreasing) to estimate parameters for the nonlinear model.

Potential analyses include nonlinear modeling, linear regression, and transformations.

Nonlinear Modeling Showing an Exponential Decline in Fruit Flies Over Time

The input is:

```
USE FRTFLYDM
NONLIN
   MODEL LIVING = 1203646*exp (-(A+B*DAY+C*DAY^2)*DAY)
   ESTIMATE / ITER=50
```

The output is:

Iteration History

No.	Loss	A	B	C
0	1.541E+013	0.010	-0.010	0.010
1	1.508E+013	-0.016	0.011	0.006
2	1.468E+013	-0.041	0.029	0.003
3	1.416E+013	-0.064	0.046	0.000
4	1.411E+013	-0.066	0.047	0.000
5	1.411E+013	-0.066	0.047	0.000
6	1.411E+013	-0.066	0.047	0.000
7	1.410E+013	-0.066	0.047	0.000
8	1.410E+013	-0.066	0.047	0.000
9	1.410E+013	-0.066	0.047	0.000
10	1.410E+013	-0.066	0.047	0.000
11	1.410E+013	-0.066	0.047	0.000
12	1.410E+013	-0.066	0.047	0.000
13	1.410E+013	-0.066	0.047	0.000
14	1.127E+013	0.006	0.019	0.000
15	7.117E+012	0.049	0.005	0.000
16	4.213E+012	0.053	0.002	0.000
17	5.111E+011	0.015	0.002	0.000
18	1.621E+011	-0.004	0.002	0.000
19	2.562E+010	-0.021	0.003	0.000
20	2.282E+010	-0.021	0.003	0.000
21	2.228E+010	-0.021	0.003	0.000
22	2.164E+010	-0.021	0.003	0.000
23	1.384E+010	-0.015	0.002	0.000
24	1.309E+010	-0.013	0.002	0.000
25	1.305E+010	-0.013	0.002	0.000
26	1.305E+010	-0.013	0.002	0.000
27	1.305E+010	-0.013	0.002	0.000
28	1.305E+010	-0.013	0.002	0.000
29	1.305E+010	-0.013	0.002	0.000
30	1.305E+010	-0.013	0.002	0.000
31	1.305E+010	-0.013	0.002	0.000

Dependent Variable :LIVING

Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	2.363E+013	3	7.877E+012
Residual	1.305E+010	170	76738341.153
Total	2.364E+013	173	
Mean corrected	1.983E+013	172	

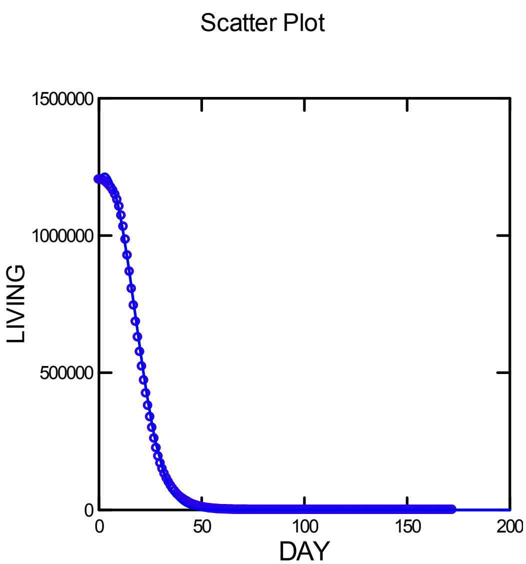
R-squares

Raw R-square (1-Residual/Total) : 0.999
Mean Corrected R-square (1-Residual/Corrected) : 0.999
R-square(Observed vs Predicted) : 0.999

Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Lower	Interval Upper
A	-0.013	0.001	-14.165	-0.014	-0.011
B	0.002	0.000	21.259	0.002	0.002
C	0.000	0.000	4.773	0.000	0.000

```
Asymptotic Correlation Matrix of Parameters

   ¦      A         B        C
---+------------------------
A  ¦  1.000
B  ¦ -0.952     1.000
C  ¦  0.866    -0.971    1.000
```

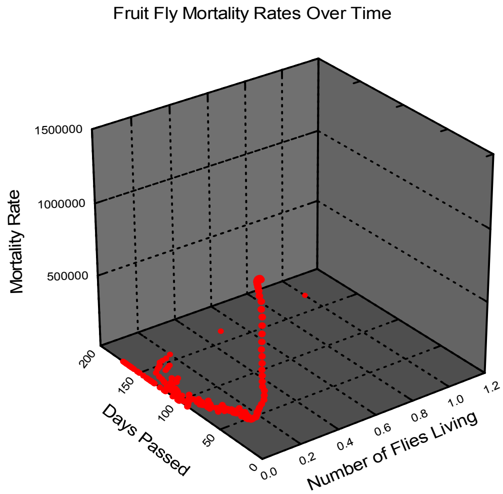


Scatterplot

The input is:

```
USE FRTFLYDM
PLOT LIVING*DAY*MORTRATE/AX=CORNER, FILL, FCOLOR=GRAY,
                    COLOR=RED, XLAB='Number of Flies Living',
                    YLAB='Days Passed', ZLAB='Mortality Rate',
                    XGRID, YGRID, ZGRID,
                    TITLE='Fruit Fly Mortality Rates Over Time'
```

The output is:



Animal Predatory Danger

SLEEPDM data contains information from a study on the effects of physical and biological characteristics and sleep patterns influencing the danger of a mammal being eaten by predators. The study includes data on the hours of dreaming and nondreaming sleep, gestation age, and body and brain weight for 62 mammals.

Variable	**Description**
SPECIES$	Type of species
BODY	Body weight of the mammal in kg
BRAIN	Brain weight of the mammal in g
SLO_SLP	Number of hours of non-dreaming sleep
DREAM_SLP	Number of hours of dreaming sleep
TOTAL_SLEEP	Number of hours of total sleep
LIFE	The life span in years
GESTATE	The gestation age
PREDATION	Index of predation as a quantitative variable
EXPOSURE	Index of exposure as a quantitative variable

The danger faced by mammals may be due to the environment they are in or their biological and physical characteristics. These studies are used to assess whether physical and biological attributes in mammals play a significant role in determining the predatory danger faced by mammals.

Potential analyses include regression trees, multiple regression, and discriminant analysis.

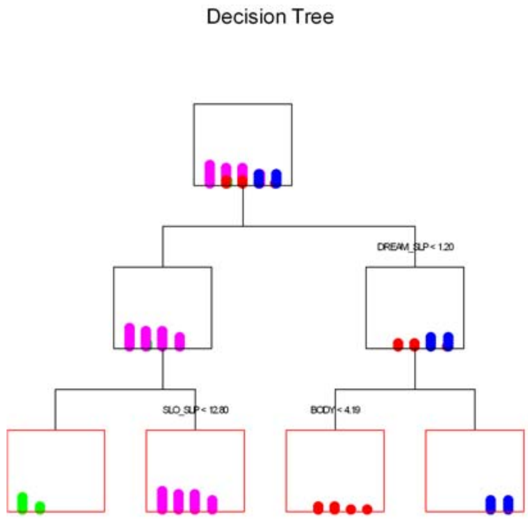
Regression Tree with DIT Plots

The input is:

```
USE SLEEPDM
TREES
    MODEL DANGER=BODY, BRAIN, SLO_SLP, DREAM_SLP, GESTATE
    ESTIMATE / DENSITY=DIT
```

The output is:

```
18 Cases Deleted due to Missing Data.

 Split   Variable      PRE   Improvement
----------------------------------------
     1   DREAM_SLP   0.404         0.404
     2   BODY        0.479         0.074
     3   SLO_SLP     0.547         0.068

 Fitting Method                        : Least Squares
 Predicted Variable                    : DANGER
 Minimum Split Index Value             : 0.050
 Minimum Improvement in PRE            : 0.050
 Maximum Number of Nodes Allowed       : 21
 Minimum Count Allowed in Each Node    : 5
 Number of Terminal Nodes in Final Tree : 4
 Proportional Reduction in Error (PRE)  : 0.547

 Node   From   Count    Mean      SD   Split Variable   Cut Value     Fit
--------------------------------------------------------------------------
    1      0      44   2.659   1.380   DREAM_SLP            1.200   0.404
    2      1      14   3.929   1.072   BODY                 4.190   0.408
    3      1      30   2.067   1.081   SLO_SLP             12.800   0.164
    4      2       6   3.167   1.169
    5      2       8   4.500   0.535
    6      3      23   2.304   1.105
    7      3       7   1.286   0.488
```



Chemistry

Enzyme Reaction Velocity

ENZYMDM data consists of measurements of an enzymatic reaction measuring the effects of an inhibitor on the reaction velocity of an enzyme and substrate.

Variable	Description
VELOCITY	Reaction velocity
SUB_CONC	Substrate concentration
INH_CONC	Inhibitor concentration

Understanding how reaction rates depend on the various reaction conditions is critical to optimizing the yield of a reaction. Also, the functional form of the rate on reaction parameters serves as a test of the theoretical models used to interpret a chemical reaction.

Potential analyses include nonlinear modeling, bootstrapping, and smoothing.

Estimation using Bootstrap Method

The input is:

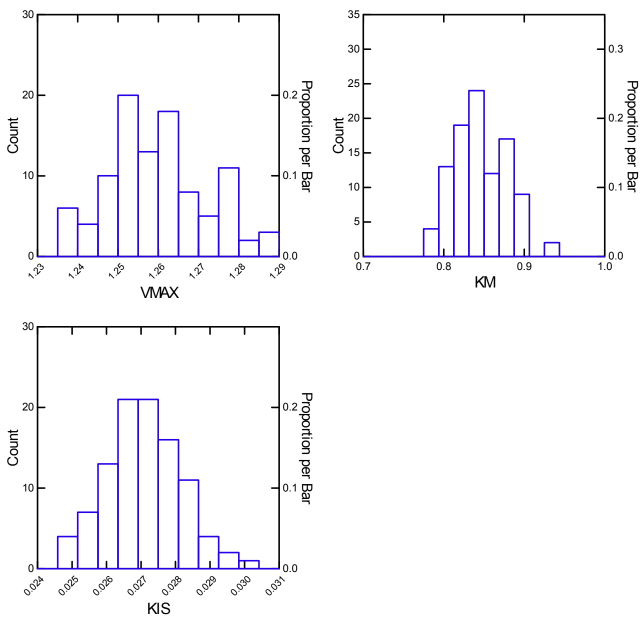
```
USE ENZYMDM
NONLIN
   MODEL VELOCITY =VMAX*SUB_CONC/(KM*(1+INH_CONC/KIS)+SUB_CONC)
   ESTIMATE / SAMPLE=BOOT(100)
```

Next, the *ESTIM* file is used to draw the density plots. *ESTIM* contains the estimated parameters for each sample.

```
USE ESTIM
   CBSTAT / MEAN, SD, SEM
   DENSITY VMAX, KM, KIS
```

The output is:

	VMAX	KM	KIS
Arithmetic Mean	1.260	0.846	0.027
Standard Error of Arithmetic Mean	0.001	0.003	0.000
Standard Deviation	0.012	0.033	0.001



Nonlinear Analysis

The input is:

```
USE ENZYMDM
NONLIN
   MODEL VELOCITY=VMAX*SUB_CONC/(KM*(1+INH_CONC/KIS)+SUB_CONC)
   ESTIMATE
```

The output is:

Iteration History

No.	Loss	VMAX	KM	KIS
0	3.568	1.010	1.020	1.030
1	3.192	1.009	0.988	0.651
2	2.897	1.011	0.961	0.481
3	0.772	1.021	0.873	0.075
4	0.154	1.134	0.845	0.029
5	0.014	1.260	0.847	0.027
6	0.014	1.259	0.847	0.027
7	0.014	1.260	0.847	0.027
8	0.014	1.260	0.847	0.027

Dependent Variable :VELOCITY

Sum of Squares and Mean Squares

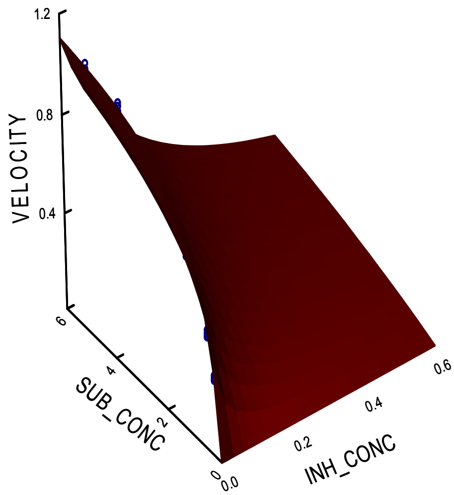
Source	SS	df	Mean Squares
Regression	15.404	3	5.135
Residual	0.014	43	0.000
Total	15.418	46	
Mean corrected	5.763	45	

R-squares

```
Raw R-square (1-Residual/Total)                  : 0.999
Mean Corrected R-square (1-Residual/Corrected)   : 0.998
R-square(Observed vs Predicted)                  : 0.998
```

Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval Lower	Upper
VMAX	1.260	0.012	104.191	1.235	1.284
KM	0.847	0.027	31.876	0.793	0.900
KIS	0.027	0.001	31.033	0.025	0.029



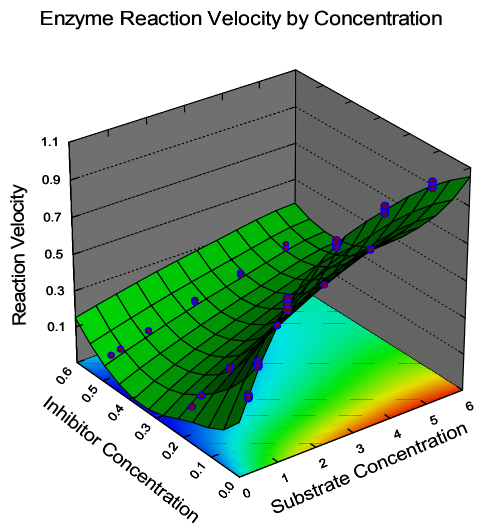
DWLS Smoother

The input is:

```
USE ENZYMDM
THICK 1.7
BEGIN
  PLOT VELOCITY*INH_CONC*SUB_CONC /SIZE=0, SMOOTH=DWLS,
                 TENSION=0.500,TITLE='', XLABEL='', YLABEL='',
                 ZLABEL='', AXES=CORNER, ACOLOR=BLACK, YGRID,
                 ZGRID,FCOLOR =gray, ZMAX =1.1,
                 HEIGHT=3.75,WIDTH=3.75, ALTITUDE = 3.75
  FACET XY
  PLOT VELOCITY*INH_CONC*SUB_CONC /SIZE=0, SMOOTH=DWLS,
                 TENSION =0.500,TITLE='', XLABEL='', YLABEL='',
                 ZLABEL='', AXES=no,SC=no,legend=no, FCOLOR= white,
                 ZMAX = 1.1, tile,HEIGHT=3.75,WIDTH=3.75,
                 ALTITUDE = 3.75
  FACET
  PLOT VELOCITY*INH_CONC*SUB_CONC / SIZE=0,SMOOTH=DWLS,
                 TENSION =0.500, TITLE='', XLABEL='', YLABEL='',
                 ZLABEL='',
                 ZMAX = 1.1,HEIGHT=3.75,WIDTH=3.75,
                 ALTITUDE = 3.75
  PLOT VELOCITY*INH_CONC*SUB_CONC /SIZE=0,SMOOTH=DWLS,
                 SURF=XYCUT, TENSION =0.500, TITLE='', XLABEL='',
                 YLABEL='', ZLABEL='',ZMAX =1.1,
                 HEIGHT=3.75,WIDTH=3.75,
                 ALTITUDE = 3.75
```

```
    PLOT VELOCITY*INH_CONC*SUB_CONC/ COLOR=11,FILL=1,SIZE=1.3,
                    TITLE='Enzyme Reaction Velocity by Concentration',
                    XLABEL= 'Substrate Concentration',
                    YLABEL= 'Inhibitor Concentration',
                    ZLABEL= 'Reaction Velocity',
                    ZMAX = 1.1,HEIGHT=3.75,WIDTH=3.75,
                    ALTITUDE = 3.75
    PLOT VELOCITY*INH_CONC*SUB_CONC / COLOR=2,FILL=0,SIZE=1.3,
                    TITLE= 'Enzyme Reaction Velocity by
                    Concentration', XLABEL= 'Substrate Concentration',
                    YLABEL= 'Inhibitor Concentration',
                    ZLABEL= 'Reaction Velocity',
                    ZMAX = 1.1,HEIGHT=3.75,WIDTH=3.75,
                    ALTITUDE = 3.75
END
THICK 1
```

The output is:



Engineering

Robust Design - Design of Experiments

DESIGNDM data consists of the results of a designed experiment to improve the performance of a fuel gauge.

Variable	Description
RUN	The case ID
SPRING	Dummy variable for the type of spring used
POINTER	Dummy variable for the type of pointer used
VENDOR	Dummy variable for the vendor used
ANGLE	Dummy variable for the type of angle bracket used
READING	The reading of the fuel gauge under the designed conditions

This example is a demonstration of the use of Design of Experiments (DOE) in the product development process. A four-factor, two-level fractional design is used to minimize the data collection needed to analyze the factors affecting the performance of a fuel gauge: *SPRING*, *POINTER*, *VENDOR*, and *ANGLE*.

ANOVA

The input is:

```
USE DESIGNDM
ANOVA
   CATEGORY SPRING / REPLACE
    DEPEND READING
    ESTIMATE
ANOVA
   CATEGORY POINTER / REPLACE
    DEPEND READING
    ESTIMATE
ANOVA
    CATEGORY VENDOR / REPLACE
    DEPEND READING
    ESTIMATE
ANOVA
   CATEGORY ANGLE / REPLACE
    DEPEND READING
    ESTIMATE
```

The output is:

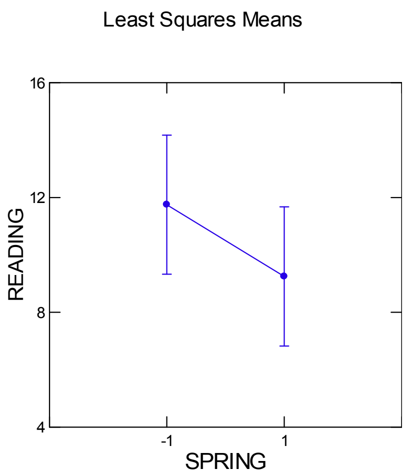
```
Effects coding used for categorical variables in model.
The categorical values encountered during processing are

  Variables          ¦        Levels
 --------------------+-------------------
  SPRING (2 levels)  ¦  -1.000     1.000

  Dependent Variable ¦ READING
  N                  ¦      16
  Multiple R         ¦   0.386
  Squared Multiple R ¦   0.149
```

Estimates of Effects B = $(X'X)^{-1}X'Y$

```
Factor   ¦ Level   READING
---------+----------------
CONSTANT ¦           10.500
SPRING   ¦ -1         1.250
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
SPRING	25.000	1	25.000	2.448	0.140
Error	143.000	14	10.214		



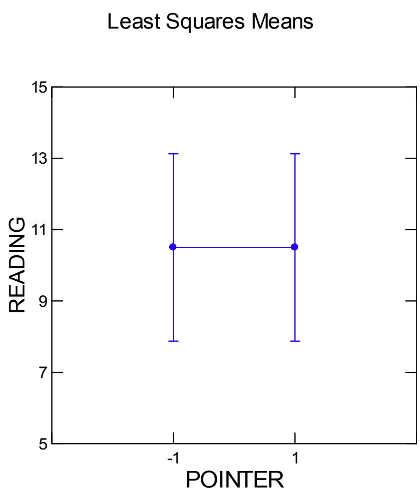

```
Durbin-Watson D Statistic    ¦ 1.103
First Order Autocorrelation  ¦ 0.404

Effects coding used for categorical variables in model.
Categorical values encountered during processing are

  Variables           ¦        Levels
 ---------------------+-------------------
  POINTER (2 levels)  ¦  -1.000     1.000

  Dependent Variable ¦ READING
  N                  ¦      16
  Multiple R         ¦   0.000
  Squared Multiple R ¦   0.000
```

Estimates of Effects $B = (X'X)^{-1}X'Y$

Factor	Level	READING
CONSTANT		10.500
POINTER	-1	0.000

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
POINTER	0.000	1	0.000	0.000	1.000
Error	168.000	14	12.000		



***** WARNING ***** :
Case 11 is an Outlier (Studentized Residual : : 2.839)

Durbin-Watson D Statistic	1.512
First Order Autocorrelation	0.201

Effects coding used for categorical variables in model.
The categorical values encountered during processing are

Variables	Levels	
VENDOR (2 levels)	-1.000	1.000

Dependent Variable	READING
N	16
Multiple R	0.270
Squared Multiple R	0.073

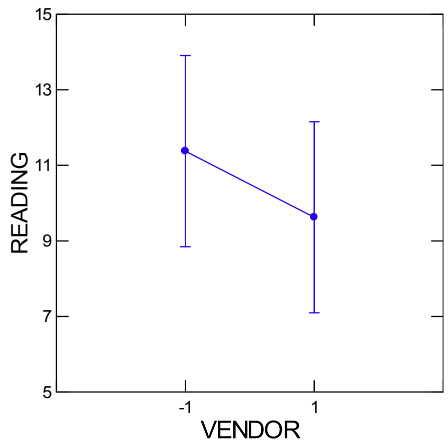
Estimates of Effects $B = (X'X)^{-1}X'Y$

Factor	Level	READING
CONSTANT		10.500
VENDOR	-1	0.875

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
VENDOR	12.250	1	12.250	1.101	0.312
Error	155.750	14	11.125		

Least Squares Means



Durbin-Watson D Statistic	1.645
First Order Autocorrelation	0.137

Effects coding used for categorical variables in model.
The categorical values encountered during processing are

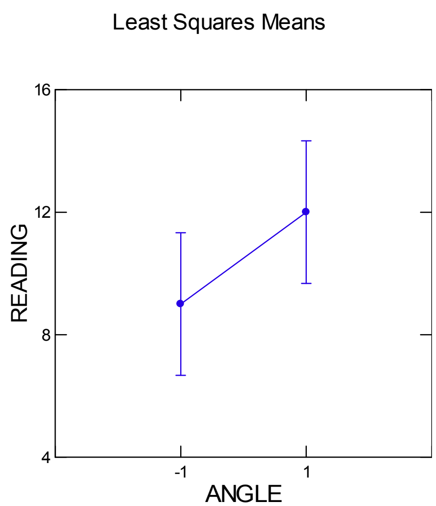
Variables	Levels	
ANGLE (2 levels)	-1.000	1.000

Dependent Variable	READING
N	16
Multiple R	0.463
Squared Multiple R	0.214

```
Estimates of Effects B = (X'X)-1X'Y

Factor   ¦ Level   READING
---------+-----------------
CONSTANT ¦            10.500
   ANGLE ¦ -1         -1.500
```

Analysis of Variance

```
Source ¦ Type III SS   df   Mean Squares   F-ratio   p-value
-------+------------------------------------------------------
ANGLE  ¦      36.000    1         36.000     3.818     0.071
Error  ¦     132.000   14          9.429
```




```
Durbin-Watson D Statistic    ¦ 1.765
First Order Autocorrelation ¦ 0.023s
```

Creating the Four Factor, Two Level Design Matrix

The input is:

```
DESIGN
   SAVE XDESIGN
   FACTORIAL / LEVELS=2 FACTORS=4 REPS=1
```

Once the design matrix is created, the following steps complete the DOE process:

- Assigning variable names
- Assigning factor level labels

- Collecting and entering data
- Performing analyses

The output is:

C:\XDESIGN.syz

	RUN	X1	X2	X3	X4
1	1.000	-1.000	-1.000	-1.000	-1.000
2	2.000	1.000	-1.000	-1.000	-1.000
3	3.000	-1.000	1.000	-1.000	-1.000
4	4.000	1.000	1.000	-1.000	-1.000
5	5.000	-1.000	-1.000	1.000	-1.000
6	6.000	1.000	-1.000	1.000	-1.000
7	7.000	-1.000	1.000	1.000	-1.000
8	8.000	1.000	1.000	1.000	-1.000
9	9.000	-1.000	-1.000	-1.000	1.000
10	10.000	1.000	-1.000	-1.000	1.000
11	11.000	-1.000	1.000	-1.000	1.000
12	12.000	1.000	1.000	-1.000	1.000
13	13.000	-1.000	-1.000	1.000	1.000
14	14.000	1.000	-1.000	1.000	1.000
15	15.000	-1.000	1.000	1.000	1.000
16	16.000	1.000	1.000	1.000	1.000

Data | Variable

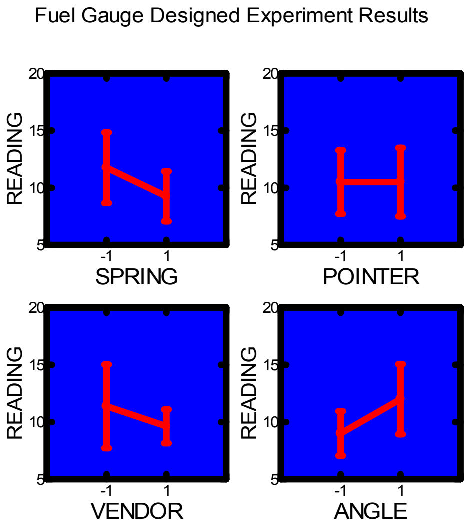
Dot Plots

The input is:

```
USE DESIGNDM
CATEGORY SPRING POINTER VENDOR ANGLE
THICK 6
CSIZE 2
DOT READING*SPRING POINTER VENDOR ANGLE/LINE, SERROR=.95,
            COLOR = 1, FCOLOR = 2,
            TITLE = 'Fuel Gauge Designed Experiment Results'
THICK 1
```

The following plots assume that we have collected data in accordance with a generated experimental design.

The output is:



Environmental Science

Mercury Levels in Freshwater Fish

The *MRCURYDM* data consists of measurements of largemouth bass in 53 different Florida lakes to examine the factors that influence the level of mercury contamination. The pH level, amount of chlorophyll, calcium, and alkalinity were measured from water samples that were collected. The age of each fish and the mercury concentration in the muscle tissue were measured (older fish tend to have higher concentrations) from a sample of fish taken from each lake. To make a fair comparison of the fish in different lakes, the investigators used a regression estimate of the expected mercury concentration in a three-year-old fish as the standardized value for each lake. Finally, in 10 of the 53 lakes, the age of the individual fish could not be determined and the average mercury concentration of the sampled fish was used.

Variable	Description
ID	Lake ID
LAKE$	Lake name
ALKLNTY	Measured alkalinity of the lake (mg/L as Calcium Carbonate)
PH	Measured PH of the lake
CALCIUM	Measured Calcium of the lake (mg/l)
CHLORO	Measured Chlorophyll of the lake (mg/l)
AVGMERC	Average mercury concentration (parts per million) in the tissue of the fish sampled from the lake
SAMPLES	Number of fish sampled in the lake
MIN	Minimum mercury concentration in sampled fish from lake
MAX	Maximum mercury concentration in sampled fish from lake
STDMERC	Regression estimate of the mercury concentration in a 3-year-old fish from the lake
AGEDATA	Indicator of the availability of age data on fish sampled
LNCHLORO	Log of *CHLORO*

Mercury is a toxic element. Its presence in the environment arises from pollution, and it subsequently becomes part of the food chain, creating potentially harmful effects for both animals and humans. Understanding the level and causes of contamination of the environment by such pollutants is an important problem in environmental science.

Potential analyses include descriptive statistics (variance and distribution), transformations, correlation and regression.

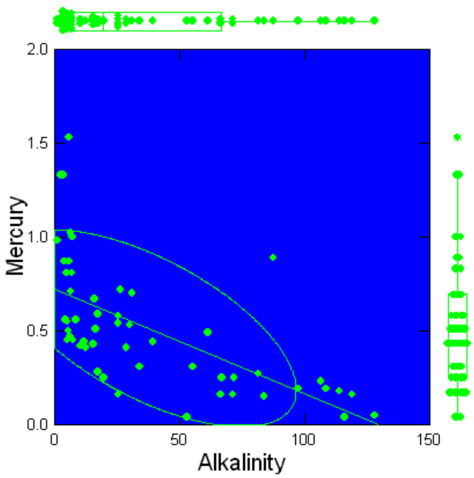
Regression of Standard Mercury Level on Lake Alkalinity

The input is:

```
USE MRCURYDM
PLOT STDMERC*ALKLNTY/ELL, SMOOTH=LINEAR, BORDER=DOX,
FILL=1,XLAB='Alkalinity', YLAB='Mercury',
TITLE='Measured Mercury Levels in Freshwater Fish vs Alkalinity',
COLOR=3, FCOLOR=2
```

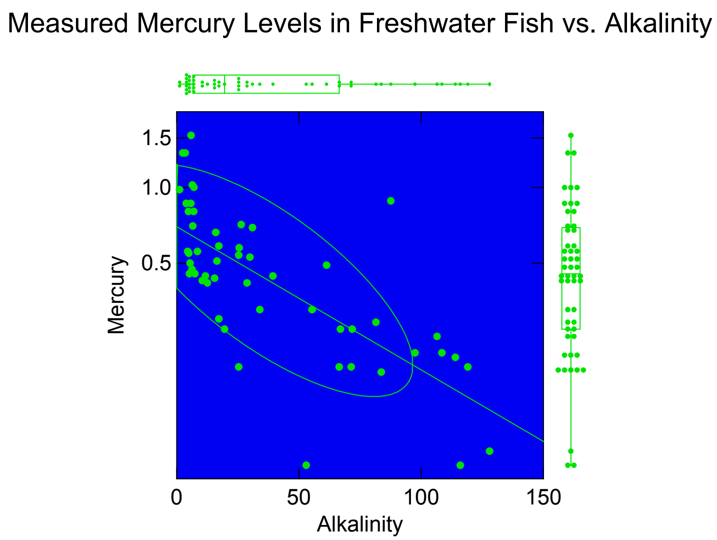
The output is:

Measured Mercury Levels in Freshwater Fish vs Alkalinity



The Graph Window can be used to transform both the Alkalinity and Standard Mercury variables so that they meet the assumptions of linear regression.

The graph below has X-Power=0.7; Y-Power=0.4



Genetics

Bayesian Estimation of Gene Frequency

Note: This example will work with the Monte-Carlo add-on module version 1.

Rao (1973) illustrated maximum likelihood estimation of gene frequencies of O, A and B blood groups through the method of scoring. McLachlan and Krishnan (1997) used the EM algorithm for the same problem. This application illustrates Bayesian estimation of these gene frequencies by the Gibbs Sampling method.

Consider the following multinomial model with four cell frequencies and their probabilities with parameters p, q, and r with $p + q + r = 1$.
Let $n = n_O + n_A + n_B + n_{AB}$.

Data	**Model**
n_O	176
n_A	182
n_B	60
n_{AB}	17

Let us consider a hypothetical augmented data for this problem to be n_O, n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB} with a multinomial model $\{n; (1-p-q)^2, p^2, 2p(1-p-q), q^2, 2q(1-p-q), 2pq\}$. With respect to the latter full model, n_{AA}, n_{BB} could be considered as missing data.

MODEL:

$$X \sim Multinomial_6\,(435; (1-p-q)^2, p^2, 2p(1-p-q), q^2, 2q(1-p-q), 2pq)$$

Prior information:

$$(p, q, r) \sim Dirichlet\,(\alpha, \beta, \gamma)$$

The full conditional densities take the form:

$$n_{AA} \sim Binomial\left(n_A, \frac{p^2}{p^2 + 2p(1-p-q)}\right)$$

$$n_{BB} \sim Binomial\left(n_B, \frac{q^2}{q^2 + 2q(1-p-q)}\right)$$

$$p \sim (1-q)Beta(2n_{AA} + n_{AO} + n_{AB} + \alpha, \quad 2n_{OO} + n_{AO} + n_{BO} + \gamma)$$

$$q \sim (1-p)Beta(2n_{BB} + n_{BO} + n_{AB} + \beta, \quad 2n_{OO} + n_{AO} + n_{BO} + \gamma)$$

For generating random samples from p and q, the generated value from the beta distribution is to be multiplied with $(1-q)$ and $(1-p)$ respectively. Since it is not possible in our system to implement this, let us consider:

$$p \sim Beta(2n_{AA} + n_{AO} + n_{AB} + \alpha, \quad 2n_{OO} + n_{AO} + n_{BO} + \gamma)$$

$$q \sim Beta(2n_{BB} + n_{BO} + n_{AB} + \beta, \quad 2n_{OO} + n_{AO} + n_{BO} + \gamma)$$

and, whenever p and q appear in other full conditionals, p is replaced by $(1-q)\,p$ and q is replaced by $(1-p)\,q$. Take $\alpha=2$, $\beta=2$ and $\gamma=2$.

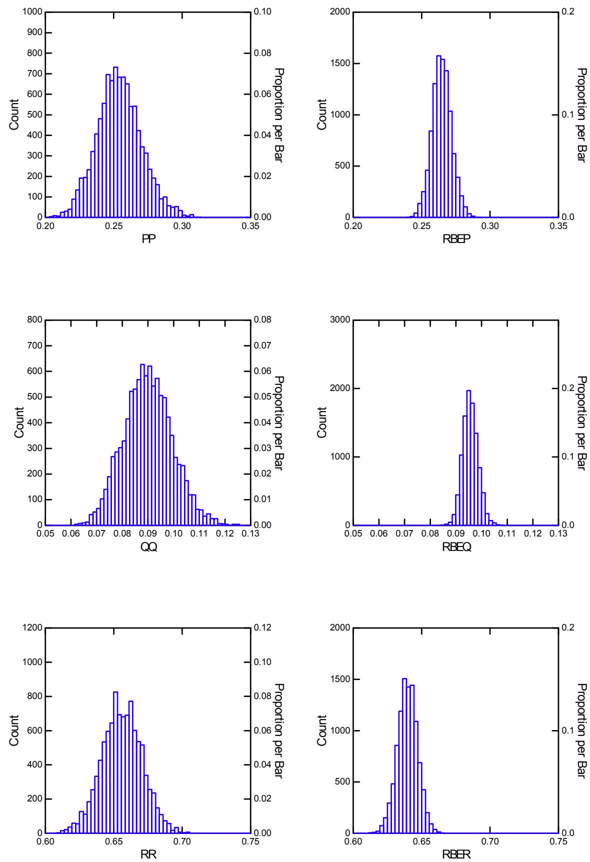
Gene Frequency Estimation using Gibbs Sampling

The input is:

```
FORMAT 10 5
MCMC
TMP N1~=182
TMP N2~=60
TMP P1~ = 0.04762
TMP P2~= 0.31034
TMP B1~=240
TMP B2~=550
TMP D1~=83
TMP D2~=550
GVAR NAA~=40,NBB~=5,P~=0.1,Q~=0.5
FUNCTION TMP FC1()
     TMP NAA~=NRN(N1~,P1~)
ENDFUNC
FUNCTION TMP FC2()
     TMP NBB~= NRN(N2~,P2~)
ENDFUNC
FUNCTION TMP FC3()
     TMP P~=BRN(B1~,B2~)
ENDFUNC
FUNCTION TMP FC4()
     TMP Q~= BRN(D1~,D2~)
ENDFUNC
SAVE GIBBSGENETIC
GIBBS FCOND(FC1(),FC2(),FC3(),FC4()) / SIZE=10000 NSAMP=1
BURNIN=1000 GAP=1, RSEED=1783
USE GIBBSGENETIC
LET PP=(1-Q1)*P1
LET QQ=(1-P1)*Q1
LET RR=1-PP-QQ
LET RBEP= (1-
QQ)*((NAA1+182+17+2)/((NAA1+182+17+2)+((2*176)+182+60-NAA1-
NBB1+2)))
LET RBEQ=(1-
PP)*((NBB1+60+17+2)/((NBB1+60+17+2)+((2*176)+182+60-NAA1-
NBB1+2)))
LET RBER=1-RBEP-RBEQ
STATISTICS PP QQ RR RBEP RBEQ RBER/ MAXIMUM MEAN MEDIAN,
MINIMUM SD VARIANCE, N PTILE={2.5 50 97.5}
BEGIN
DENSITY PP RBEP/HIST XMIN=0.20 XMAX=0.35 LOC={0,0}
DENSITY QQ RBEQ/HIST  XMIN=0.05 XMAX=0.13 LOC={0,-3}
DENSITY RR RBER/HIST  XMIN=0.60 XMAX=0.75 LOC={0,-6}
END
FORMAT
CLEAR function = fc1, fc2, fc3,fc4
```

The output is:

```
                   ¦      PP        QQ        RR      RBEP      RBEQ
-------------------+---------------------------------------------
N of Cases         ¦   10000     10000     10000     10000     10000
Minimum            ¦ 0.21275   0.06147   0.58743   0.22834   0.09148
Maximum            ¦ 0.32731   0.12441   0.68789   0.26139   0.12460
Median             ¦ 0.26412   0.09108   0.64442   0.24480   0.10736
Arithmetic Mean    ¦ 0.26461   0.09119   0.64420   0.24486   0.10753
Standard Deviation ¦ 0.01516   0.00922   0.01334   0.00436   0.00448
Variance           ¦ 0.00023   0.00009   0.00018   0.00002   0.00002
Method = CLEVELAND ¦
 2.500%            ¦ 0.23545   0.07382   0.61744   0.23642   0.09916
 50.000%           ¦ 0.26412   0.09108   0.64442   0.24480   0.10736
 97.500%           ¦ 0.29642   0.11032   0.67021   0.25359   0.11642

                   ¦    RBER
-------------------+--------
N of Cases         ¦   10000
Minimum            ¦ 0.62002
Maximum            ¦ 0.67243
Median             ¦ 0.64772
Arithmetic Mean    ¦ 0.64761
Standard Deviation ¦ 0.00671
Variance           ¦ 0.00005
Method = CLEVELAND ¦
            2.500% ¦ 0.63412
           50.000% ¦ 0.64772
           97.500% ¦ 0.66050
```



Maximum likelihood estimates of *p, q* and *r* evaluated by the scoring method or the EM algorithm are 0.26444, 0.09317 and 0.64239. With the available prior information, the estimates of *p, q* and *r* are approximated by the Gibbs Sampling method. The empirical estimates of *p, q,* and *r* are 0.25407, 0.09003 and 0.65589 respectively. Rao-Blackwellized estimates are 0.26470, 0.09564, and 0.63966 respectively.

Manufacturing

Quality Control

The *BOXES* data consists of daily measurements of five randomly selected computer components.

Variable	Description
DAY	The day the sample was taken
SAMPLE	The sample number for the day (1-5)
OHMS	The resistance of the component in ohms

Quality control charts are used regularly in manufacturing environments to keep track of manufacturing processes, diagnose problems, and improve operations.

Potential analyses include descriptive statistics, quality control charts, ANOVA, and time series.

R Chart of Ohms vs Days

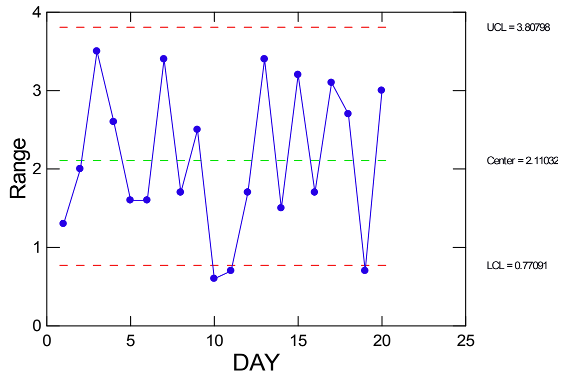
The input is:

```
USE BOXES
QC
   SHEWHART OHMS*DAY / TYPE=R PLIMITS = {.025,.975}
```

The output is:

```
Number of Lines of Input Data Read           : 100.00000
Number with Missing Data or Zero Weight      :   0.00000
Number of Samples to be Plotted              :  20.00000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                    :  19.93100
Estimated Population Standard Deviation      :   0.90730
Total N (Excluding Missing Data)             :     100
```

R Chart for OHMS with Alpha = 0.05



X-bar Chart of Ohms vs Days

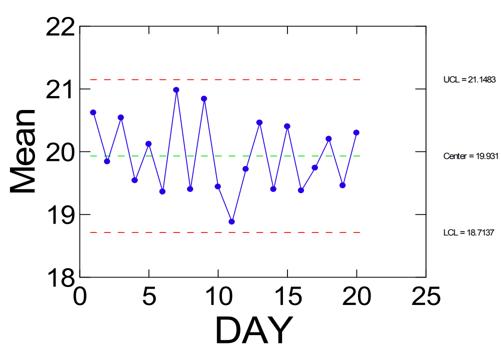
The input is:

```
USE BOXES
QC
  SHEWHART OHMS*DAY / TYPE=XBAR
```

The output is:

```
Number of Lines of Input Data Read           : 100.00000
Number with Missing Data or Zero Weight      :   0.00000
Number of Samples to be Plotted              :  20.00000
(Only Subgroups Containing Data are Plotted)
Estimated Population Mean                    :  19.93100
Estimated Population Standard Deviation      :   0.90730
Total N (Excluding Missing Data)             :     100
```

X-BAR Chart for OHMS with Alpha = 0.0027



Medical Research

Clinical Trials

The *CANCERDM* data set contains information from a study of the effects of supplemental Vitamin C as part of routine cancer treatment for 100 patients and 1000 controls (that is, 10 controls for each patient).

Variable	Description
CASE	Case ID
ORGAN$	Organ affected by cancer
SEX$	Sex of patient
AGE	Age of the patient
SURVATD	Survival of patient measured from first hospital attendance
CNTLATD	Survival of control group from first hospital attendance
SURVUNTR	Survival of patient from time cancer deemed un-treatable
CNTLUNTR	Survival of control from time cancer deemed untreatable
LOGSURVA	Logarithm of *SURVATD*
LOGCNTLA	Logarithm of *CNTLAD*
LOGSURVU	Logarithm of *SURVUNTR*
LOGCNTLU	Logarithm of *CNTLUNTR*

Clinical trials of this sort are the basis for evaluating the effectiveness of any new drug or medical treatment. They are a critical part of the FDA approval process in the U.S. and similar evaluations in virtually all developed countries.

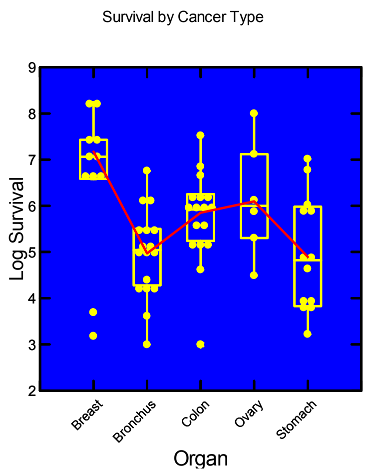
Potential analyses include descriptive statistics, transformations, ANOVA and survival analysis.

Box Plot of Selected Cancer Types

The input is:

```
USE CANCERDM
SELECT (ORGAN$= 'Breast') OR (ORGAN$= 'Bronchus') OR,
       (ORGAN$= 'Colon') OR (ORGAN$= 'Ovary') OR,
       (ORGAN$= 'Stomach')
THICK 3
CATEGORY ORGAN$
BEGIN
   DEN LOGSURVA*ORGAN$ / DOX,SIZE=1.2,FILL=1, FCOLOR=BLUE,
                         COLOR=YELLOW,YLAB='Log Survival',
                         XLAB='Organ',HEI=5IN,WID=5IN,
                         TITLE='Survival by Cancer Type'
   PLOT LOGSURVA*ORGAN$ / SMOOTH=LOWESS,TENSION=0,SIZE=0,
                          COLOR=1,YLAB='',XLAB='',HEI=5IN,
                          WID=5IN,TITLE=''
END
THICK 1
```

The output is:

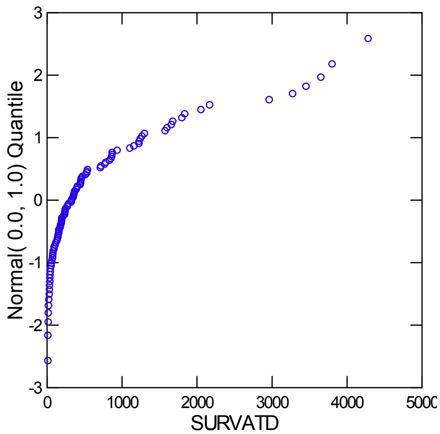


Transformation of Survival Variable

The input is:

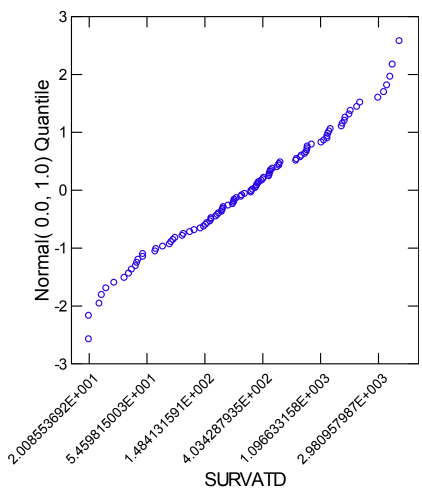
```
USE CANCERDM
    PPLOT SURVATD
```

The output is:



To perform an ANOVA, the variable used must produce a straight line in a probability plot. Clearly the distribution of *SURVATD* is skewed and must be transformed.

You can use the Graph Window to reduce the X-axis power from 1 through successive exponential power transformation 0.9 to 0.1 and finally to 0, which is same as the log transformation.



The second plot should appear. Since the probability plot is much closer to a straight line we see that a log transformation is appropriate.

Survival Rates of Melanoma Patients

MELANMDM data contains reports on melanoma patients.

Variable	**Description**
TIME	The survival time for melanoma patients in days
CENSOR	The censoring variable
WEIGHT	The weight variable
ULCER	Presence or absence of ulcers
DEPTH	Depth of ulceration
NODES	Number of lymph nodes that are affected
SEX$	The sex of the patient
SEX	The stratification variable coded for the analysis

Survival studies are used in the area of drug development. Survival rates of the patients on an experimental drug are studied to determine the effectiveness of the drug in treating melanoma. Sex may be used as a stratification variable to examine the difference in the survival patterns of male and female patients.

Potential analyses include survival analysis and logistic regression.

Stratified Cox Regression

The input is:

```
USE MELNMADM
SURVIVAL
   MODEL TIME =ULCER, DEPTH, NODES / CENSOR=CENSOR STRATA=SEX
   ESTIMATE / COX
   LTAB / CHAZ
```

The output is:

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records            : 69
Records Kept for Analysis : 69

Censoring      ¦ Observations
---------------+-------------
Exact Failures ¦           36
Right Censored ¦           33
```

Covariate Means

```
ULCER ¦ 1.507
DEPTH ¦ 2.562
NODES ¦ 3.246
```

Type 1: Exact Failures and Right Censoring

```
Overall Time Range:      [72.000, 7307.000]
Failure Time Range:      [72.000, 1606.000]
Stratification on SEX specified, 2 levels
```

Cox Proportional Hazards Estimation

With stratification on SEX

Iteration	Step	Log-Likelih-ood
0	0	-112.564
1	0	-108.343
2	0	-103.570
3	0	-103.533
4	0	-103.533

Results after 4 Iterations

```
Final Convergence Criterion      : 0.000
Maximum Gradient Element         : 0.000
Initial Score Test of Regression : 32.533 with 3 df
Significance Level (p-Value)     : 0.000
Final Log-Likelihood             : -103.533
AIC                              : 213.066
Schwarz's BIC                    : 217.816
```

Parameter	Estimate	Standard Error	Z	p-Value
ULCER	-0.817	0.385	-2.123	0.034
DEPTH	0.083	0.053	1.587	0.112
NODES	0.131	0.057	2.289	0.022

Life Table for Last Cox Model

With stratification on SEX
The following results are for SEX = 0.
Evaluated at Mean Values of Covariates:

```
ULCER : 1.507
DEPTH : 2.562
NODES : 3.246
```

No Tied Failure Times

Number at Risk	Number Failing	Time	Model Survival Probability	Model Hazard Rate	Cumulative Hazard
31.000	1.000	133.000	0.967	0.032	0.033
30.000	1.000	184.000	0.934	0.034	0.069
29.000	1.000	251.000	0.900	0.036	0.106
28.000	1.000	320.000	0.865	0.038	0.146
27.000	1.000	391.000	0.829	0.041	0.188
26.000	1.000	414.000	0.793	0.042	0.232
25.000	1.000	434.000	0.758	0.043	0.277
23.000	1.000	471.000	0.721	0.048	0.327
22.000	1.000	544.000	0.682	0.053	0.383
20.000	1.000	788.000	0.638	0.062	0.449
19.000	1.000	812.000	0.596	0.065	0.518

15.000	1.000	1151.000	0.547	0.079	0.603
13.000	1.000	1239.000	0.491	0.098	0.711
5.000	1.000	1579.000	0.361	0.236	1.018
4.000	1.000	1606.000	0.230	0.308	1.468

Group size : 31.000
Number Failing : 15.000

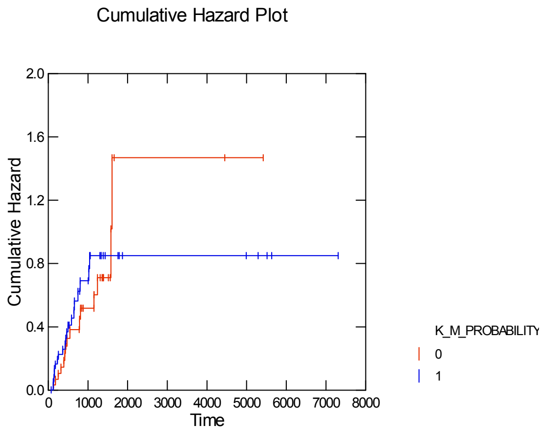
The following results are for SEX = 1.
Evaluated at Mean Values of Covariates:

ULCER : 1.507
DEPTH : 2.562
NODES : 3.246

No Tied Failure Times

Number at Risk	Number Failing	Time	Model Survival Probability	Model Hazard Rate	Cumulative Hazard
38.000	1.000	72.000	0.998	0.002	0.002
37.000	1.000	125.000	0.973	0.024	0.027
36.000	1.000	127.000	0.949	0.025	0.053
35.000	1.000	142.000	0.923	0.026	0.080
34.000	1.000	151.000	0.898	0.027	0.108
33.000	1.000	154.000	0.873	0.028	0.136
32.000	1.000	176.000	0.848	0.028	0.165
31.000	1.000	229.000	0.823	0.029	0.195
30.000	1.000	256.000	0.798	0.030	0.226
29.000	1.000	362.000	0.772	0.031	0.258
28.000	1.000	422.000	0.747	0.033	0.292
27.000	1.000	441.000	0.720	0.035	0.329
26.000	1.000	465.000	0.692	0.038	0.368
25.000	1.000	495.000	0.663	0.041	0.411
23.000	1.000	584.000	0.634	0.043	0.455
22.000	1.000	645.000	0.603	0.048	0.505
21.000	1.000	659.000	0.569	0.055	0.563
20.000	1.000	749.000	0.536	0.058	0.624
18.000	1.000	803.000	0.501	0.063	0.691
16.000	1.000	1020.000	0.464	0.071	0.767
15.000	1.000	1042.000	0.427	0.077	0.850

Group size : 38.000
Number Failing : 21.000




```
Log-Rank Test, Stratification on SEX, Strata Range 1 to 2

                ¦ Chi-Square
                ¦  Statistic
Method          ¦  with 1 df   p-Value
----------------+---------------------
Mantel          ¦      0.568     0.451
Breslow-Gehan   ¦      1.589     0.207
Tarone-Ware     ¦      1.167     0.280
```

Stratified Kaplan-Meier Estimation

The input is:

```
USE MELNMADM
SURVIVAL
   MODEL TIME / CENSOR=CENSOR, STRATA=SEX
   ESTIMATE
   LTAB
```

The output is:

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records            : 69
Records Kept for Analysis : 69

Censoring       ¦ Observations
----------------+-------------
Exact Failures  ¦           36
Right Censored  ¦           33
```

Type 1: Exact Failures and Right Censoring

Overall Time Range: [72.000, 7307.000]
Failure Time Range: [72.000, 1606.000]
Stratification on SEX specified, 2 levels

Nonparametric Estimation
Table of Kaplan-Meier Probabilities

With stratification on SEX
The following results are for SEX = 0.

Number at Risk	Number Failing	Time	K-M Probability	Standard Error	95.0% Confidence Interval Lower	Upper
31.000	1.000	133.000	0.968	0.032	0.792	0.995
30.000	1.000	184.000	0.935	0.044	0.766	0.983
29.000	1.000	251.000	0.903	0.053	0.729	0.968
28.000	1.000	320.000	0.871	0.060	0.692	0.950
27.000	1.000	391.000	0.839	0.066	0.655	0.929
26.000	1.000	414.000	0.806	0.071	0.619	0.908
25.000	1.000	434.000	0.774	0.075	0.584	0.885
23.000	1.000	471.000	0.741	0.079	0.547	0.861
22.000	1.000	544.000	0.707	0.082	0.512	0.836
20.000	1.000	788.000	0.672	0.085	0.475	0.808
19.000	1.000	812.000	0.636	0.088	0.439	0.780
15.000	1.000	1151.000	0.594	0.092	0.394	0.747
13.000	1.000	1239.000	0.548	0.095	0.346	0.711
5.000	1.000	1579.000	0.438	0.124	0.199	0.657
4.000	1.000	1606.000	0.329	0.133	0.103	0.580

Group size : 31.000
Number Failing : 15.000
Product Limit Likelihood : -58.200

Mean Survival Time

Mean Survival Time	95.0% Confidence Interval Lower	Upper
2395.302	1278.588	3512.017

Survival Quantiles

Probability	Survival Time	95.0% Confidence Interval Lower	Upper
0.250	.	1579.000	.
0.500	1579.000	788.000	.
0.750	471.000	251.000	1151.000

The following results are for SEX = 1.

Number at Risk	Number Failing	Time	K-M Probability	Standard Error	95.0% Confidence Interval Lower	Upper
38.000	1.000	72.000	0.974	0.026	0.828	0.996
37.000	1.000	125.000	0.947	0.036	0.806	0.987
36.000	1.000	127.000	0.921	0.044	0.775	0.974
35.000	1.000	142.000	0.895	0.050	0.743	0.959
34.000	1.000	151.000	0.868	0.055	0.712	0.943
33.000	1.000	154.000	0.842	0.059	0.682	0.926
32.000	1.000	176.000	0.816	0.063	0.652	0.908
31.000	1.000	229.000	0.789	0.066	0.623	0.889
30.000	1.000	256.000	0.763	0.069	0.594	0.869
29.000	1.000	362.000	0.737	0.071	0.566	0.849
28.000	1.000	422.000	0.711	0.074	0.539	0.828

27.000	1.000	441.000	0.684	0.075	0.511	0.807
26.000	1.000	465.000	0.658	0.077	0.485	0.785
25.000	1.000	495.000	0.632	0.078	0.459	0.763
23.000	1.000	584.000	0.604	0.080	0.431	0.739
22.000	1.000	645.000	0.577	0.081	0.405	0.715
21.000	1.000	659.000	0.549	0.081	0.378	0.691
20.000	1.000	749.000	0.522	0.082	0.353	0.666
18.000	1.000	803.000	0.493	0.082	0.325	0.640
16.000	1.000	1020.000	0.462	0.083	0.297	0.612
15.000	1.000	1042.000	0.431	0.083	0.269	0.584

```
Group size                 :  38.000
Number Failing             :  21.000
Product Limit Likelihood : -89.404
```

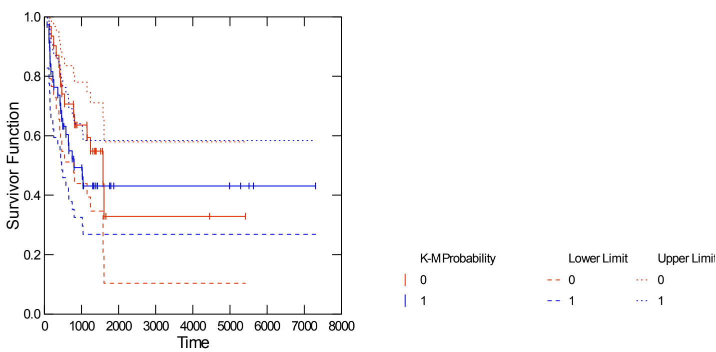
Mean Survival Time

Mean Survival Time	95.0% Confidence Interval Lower	Upper
3404.857	2282.604	4527.110

Survival Quantiles

Probability	Survival Time	95.0% Confidence Interval Lower	Upper
0.250	.	.	.
0.500	803.000	465.000	.
0.750	362.000	142.000	584.000

Survival Plot



Log-Rank Test, Stratification on SEX, Strata Range 1 to 2

Method	Chi-Square Statistic with 1 df	p-Value
Mantel	0.568	0.451
Breslow-Gehan	1.589	0.207
Tarone-Ware	1.167	0.280

Weibull Estimation

The input is:

```
USE MELNMADM
SURVIVAL
  MODEL TIME = ULCER, DEPTH, NODES / CENSOR=CENSOR
  ESTIMATE / EWB
  QNTL
```

The output is:

```
Time Variable   : TIME
Censor Variable : CENSOR

Input Records            : 69
Records Kept for Analysis : 69

Censoring      ¦ Observations
---------------+-------------
Exact Failures ¦           36
Right Censored ¦           33


Covariate Means

ULCER ¦ 1.50725
DEPTH ¦ 2.56203
NODES ¦ 3.24638
```

Type 1: Exact Failures and Right Censoring

```
Overall Time Range:      [72.00000 , 7307.00000]
Failure Time Range:      [72.00000 , 1606.00000]
```

Weibull Model B(1)--shape, B(2)--scale

```
Extreme value parameterization

Convergence : 0.00000
Tolerance   : 0.00000


Iteration   Step   Log-Likelihood   Method
------------------------------------------
0           0      -346.02864       BHHH
1           0      -333.96139       BHHH
2           0      -325.72128       BHHH
3           0      -318.69616       BHHH
4           0      -316.15793       BHHH
5           0      -312.05797       N-R
6           0      -307.55232       BHHH
7           0      -306.81388       BHHH
8           1      -306.61528       N-R
9           0      -306.50985       N-R
10          0      -306.50812       N-R
11          0      -306.50812       N-R
```

Results after 11 Iterations

```
Final Convergence Criterion      : 0.00000
Maximum Gradient Element         : 0.00001
Initial Score Test of Regression : 14.73796 with 5 df
Significance Level (p-value)     : 0.01154
Final Log-Likelihood             : -306.50812
```

```
AIC                                : 623.01624
Schwarz's BIC                      : 634.18677

Parameter ¦ Estimate   Standard Error          Z   p-value
----------+-----------------------------------------------
B(1)      ¦  1.20162          0.16086    7.47021   0.00000
B(2)      ¦  7.27717          0.72848    9.98955   0.00000
ULCER     ¦  0.77647          0.43142    1.79978   0.07190
DEPTH     ¦ -0.15354          0.05740   -2.67495   0.00747
NODES     ¦ -0.06307          0.01995   -3.16235   0.00156

1.0/B(1): 0.83221, EXP(B(2)): 1446.88707

       ¦ Mean Failure
Vector ¦         Time       Variance
-------+----------------------------
ZERO   ¦   1595.59198   3.71688E+006
MEAN   ¦    900.37653   1.18354E+006

Coefficient of Variation:  1.20828

          ¦             95.0% Confidence Interval
Parameter ¦ Estimate          Lower            Upper
----------+--------------------------------------
B(1)      ¦  1.20162        0.88635          1.51689
B(2)      ¦  7.27717        5.84938          8.70496
ULCER     ¦  0.77647       -0.06911          1.62204
DEPTH     ¦ -0.15354       -0.26604         -0.04104
NODES     ¦ -0.06307       -0.10217         -0.02398
```

Covariance Matrix

```
      ¦ B(1)       B(2)       ULCER     DEPTH     NODES
------+--------------------------------------------------
B(1)  ¦ 0.02587
B(2)  ¦ 0.00284    0.53068
ULCER ¦ 0.00750    -0.28760   0.18613
DEPTH ¦ -0.00122   -0.02138   0.00720   0.00329
NODES ¦ -0.00025   -0.00290   0.00068   0.00002   0.00040
```

Correlation Matrix

```
      ¦ B(1)       B(2)       ULCER     DEPTH     NODES
------+--------------------------------------------------
B(1)  ¦ 1.00000
B(2)  ¦ 0.02421    1.00000
ULCER ¦ 0.10803    -0.91511   1.00000
DEPTH ¦ -0.13193   -0.51120   0.29073   1.00000
NODES ¦ -0.07699   -0.19929   0.07878   0.02046   1.00000
```

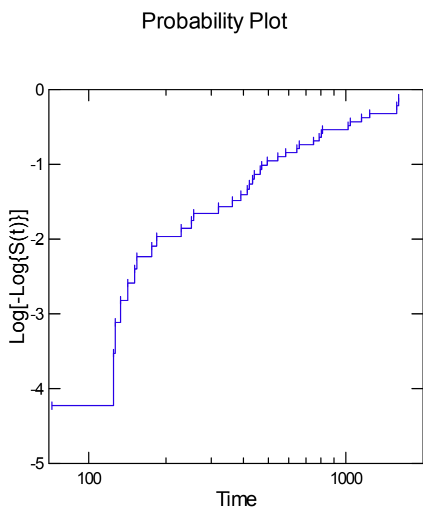
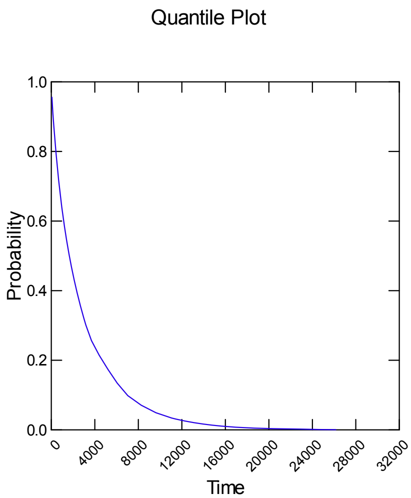


Table of Estimated Quantiles for Last Accelerated Weibull Model

Covariate Vector

ULCER : 1.507
DEPTH : 2.562
NODES : 3.246

Survival Probability	Estimated Time	95.0% Confidence Interval Lower	Upper	Log of Estimated Time	Standard Error of Log Time
0.999	0.637	0.079	5.166	-0.451	1.068
0.995	4.418	0.895	21.825	1.486	0.815
0.990	10.193	2.549	40.769	2.322	0.707
0.975	30.935	10.186	93.952	3.432	0.567
0.950	72.263	29.169	179.023	4.280	0.463
0.900	171.618	84.262	349.534	5.145	0.363
0.750	573.787	353.087	932.437	6.352	0.248
0.667	866.645	560.840	1339.193	6.765	0.222
0.500	1650.688	1101.241	2474.271	7.409	0.207
0.333	2870.859	1861.913	4426.540	7.962	0.221
0.250	3796.547	2386.677	6039.263	8.242	0.237
0.100	6985.190	3989.200	12231.245	8.852	0.286
0.050	9583.149	5152.747	17822.869	9.168	0.317
0.025	12306.215	6287.225	24087.403	9.418	0.343
0.010	16065.792	7752.889	33292.060	9.684	0.372
0.005	19013.916	8840.918	40892.701	9.853	0.391
0.001	26151.527	11313.122	60452.137	10.172	0.428


Quantile Plot
1.0
0.8
0.6
0.4
0.2
0.0
Probability
0
4000
8000
12000
16000
20000
24000
28000
32000
Time

Psychology

Day Care Effects on Child Development

The *DAYCREDM* data consists of three measures of a child's social competence: a measure for behavior at dinner, a measure for behavior in dealing with strangers, and a measure involving social problem solving in a cognitive test. In addition, there is a categorical variable for the setting in which a child was raised, either by parents, by a babysitter, or in a day-care center.

Variable	Description
SETTING$	Daycare setting in which child is raised
SETTING	Coded setting
DINNER	Behavioral measure of skill during dinner
STRANGER	Measure of skill in dealing with a stranger
PROBLEM	Social problem solving skill in a cognitive test

An important issue in child development is whether the daycare setting in which a child is raised has a differential effect on social behavior. This data set offers three measures of social competence for children in three different daycare settings--some cared for during the day by parents, others by a babysitter, and the rest in a daycare center. The data set is a good candidate for MANOVA because it offers three ways of measuring for a single latent variable—social competence. One critical issue is whether the data satisfy the assumptions of MANOVA, especially regarding homogeneity of variance and covariance across settings.

Potential analyses include ANOVA, MANOVA, regression, and factor analysis.

MANOVA

The input is:

```
USE DAYCREDM
MANOVA
  PLENGTH LONG
  CATEGORY SETTING
  DEPEND DINNER, STRANGER, PROBLEM
  ESTIMATE
```

The output is:

```
Effects coding used for categorical variables in model.
The categorical values encountered during processing are

Variables          ¦        Levels
-------------------+----------------------
SETTING (3 levels) ¦ 1.000   2.000   3.000

N of Cases Processed : 48

Dependent Variable Means

  DINNER   STRANGER   PROBLEM
-----------------------------
1288.188    714.250    54.083

Estimates of Effects B = (X'X)-1X'Y

Factor   ¦ Level     DINNER   STRANGER   PROBLEM
---------+-------------------------------------
CONSTANT ¦         1308.795    690.589    51.733
SETTING  ¦ 1       -166.479    -62.116    -2.207
SETTING  ¦ 2        109.905   -126.189   -12.533

Standardized Estimates of Effects

Factor   ¦ Level   DINNER   STRANGER   PROBLEM
---------+-----------------------------------
CONSTANT ¦          0.000      0.000     0.000
SETTING  ¦ 1       -0.278     -0.176    -0.069
SETTING  ¦ 2        0.156     -0.304    -0.331

Total Sum of Product Matrix

         ¦       DINNER      STRANGER     PROBLEM
---------+--------------------------------------
DINNER   ¦ 13624387.313
STRANGER ¦  2382747.750   4713117.000
PROBLEM  ¦   241634.250    218044.000   39267.667

Residual Sum of Product Matrix E'E = Y'Y-Y'XB

         ¦       DINNER      STRANGER     PROBLEM
---------+--------------------------------------
DINNER   ¦ 12936578.626
STRANGER ¦  2099145.095   3833722.926
PROBLEM  ¦   230259.126    149554.411   33741.074

Residual Covariance Matrix SY.X

         ¦     DINNER    STRANGER   PROBLEM
---------+--------------------------------
DINNER   ¦ 287479.525
STRANGER ¦  46647.669   85193.843
PROBLEM  ¦   5116.869    3323.431   749.802

Residual Correlation Matrix RY.X

         ¦ DINNER   STRANGER   PROBLEM
---------+---------------------------
DINNER   ¦  1.000
STRANGER ¦  0.298      1.000
PROBLEM  ¦  0.349      0.416     1.000

Information Criteria

AIC             ¦ 1878.445
AIC (Corrected) ¦ 1893.445
Schwarz's BIC   ¦ 1906.513
```

SETTING : 1
N of Cases : 19

Least Squares Means

	DINNER	STRANGER	PROBLEM
LS Mean	1142.316	628.474	49.526
Standard Error	123.006	66.962	6.282

SETTING : 2
N of Cases : 10

Least Squares Means

	DINNER	STRANGER	PROBLEM
LS Mean	1418.700	564.400	39.200
Standard Error	169.552	92.301	8.659

SETTING : 3
N of Cases : 19

Least Squares Means

	DINNER	STRANGER	PROBLEM
LS Mean	1365.368	878.895	66.474
Standard Error	123.006	66.962	6.282

Test for effect called: CONSTANT

Null Hypothesis Contrast AB

DINNER	STRANGER	PROBLEM
1308.795	690.589	51.733

Inverse Contrast A(X'X)-1A'

0.023

Hypothesis Sum of Product Matrix H = B'A'(A(X'X)-1A')-1AB

	DINNER	STRANGER	PROBLEM
DINNER	75105991.386		
STRANGER	39629901.926	20910836.774	
PROBLEM	2968749.169	1566469.415	117347.118

Error Sum of Product Matrix G = E'E

	DINNER	STRANGER	PROBLEM
DINNER	12936578.626		
STRANGER	2099145.095	3833722.926	
PROBLEM	230259.126	149554.411	33741.074

Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
DINNER	75105991.386	1	75105991.386	261.257	0.000
Error	12936578.626	45	287479.525		
STRANGER	20910836.774	1	20910836.774	245.450	0.000
Error	3833722.926	45	85193.843		
PROBLEM	117347.118	1	117347.118	156.504	0.000
Error	33741.074	45	749.802		

Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.100	128.489	3, 43	0.000
Pillai Trace	0.900	128.489	3, 43	0.000
Hotelling-Lawley Trace	8.964	128.489	3, 43	0.000

Test of Residual Roots

Roots	Chi-square	df
1 through 1	102.306	3

Canonical Correlations

0.948

Dependent Variable Canonical Coefficients Standardized
by Conditional (within Groups) Standard Deviations

DINNER	0.578
STRANGER	0.523
PROBLEM	0.204

Canonical Loadings (Correlations between Conditional
Dependent Variables and Dependent Canonical Factors)

DINNER	0.805
STRANGER	0.780
PROBLEM	0.623

Information Criteria

AIC	1878.445
AIC (Corrected)	1893.445
Schwarz's BIC	1906.513

Test for effect called: SETTING

Null Hypothesis Contrast AB

	DINNER	STRANGER	PROBLEM
1	-166.479	-62.116	-2.207
2	109.905	-126.189	-12.533

Inverse Contrast A(X'X)-1A'

	1	2
1	0.040	
2	-0.028	0.056

Hypothesis Sum of Product Matrix H = B'A'(A(X'X)-1A')-1AB

	DINNER	STRANGER	PROBLEM
DINNER	687808.686		
STRANGER	283602.655	879394.074	
PROBLEM	11375.124	68489.589	5526.593

Error Sum of Product Matrix G = E'E

	DINNER	STRANGER	PROBLEM
DINNER	12936578.626		
STRANGER	2099145.095	3833722.926	
PROBLEM	230259.126	149554.411	33741.074

Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
DINNER	687808.686	2	343904.343	1.196	0.312
Error	12936578.626	45	287479.525		
STRANGER	879394.074	2	439697.037	5.161	0.010
Error	3833722.926	45	85193.843		
PROBLEM	5526.593	2	2763.296	3.685	0.033
Error	33741.074	45	749.802		

Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.723	2.519	6, 86	0.027
Pillai Trace	0.290	2.488	6, 88	0.029
Hotelling-Lawley Trace	0.364	2.547	6, 84	0.026

THETA	S	M	N	p-value
0.232	2	0.000	20.500	0.035

Test of Residual Roots

Roots	Chi-square	df
1 through 2	14.250	6
2 through 2	2.624	2

Canonical Correlations

1	2
0.482	0.241

Dependent Variable Canonical Coefficients Standardized
by Conditional (within Groups) Standard Deviations

	1	2
DINNER	-0.341	0.980
STRANGER	0.723	0.288
PROBLEM	0.554	-0.424

Canonical Loadings (Correlations between Conditional
Dependent Variables and Dependent Canonical Factors)

	1	2
DINNER	0.068	0.918
STRANGER	0.852	0.404
PROBLEM	0.736	0.037

Information Criteria

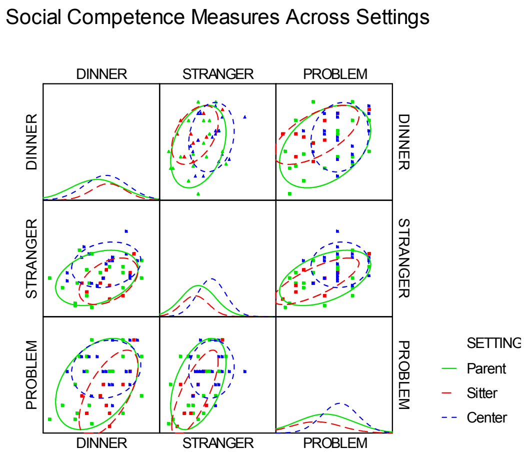
AIC	1878.445
AIC (Corrected)	1893.445
Schwarz's BIC	1906.513

Scatterplot Matrix (SPLOM)

The input is:

```
USE DAYCREDM
LABEL SETTING / 1='Parent', 2 ='Sitter', 3='Center'
SPLOM DINNER STRANGER PROBLEM /GROUP=SETTING, DEN=NORM,
ELL, DASH={1,7,10}, COLOR={3,1,2}, FILL, SYMBOL={1,4,8},
OVERLAY,
TITLE='Social Competence Measures Across Settings'
```

The output is:



A scatterplot matrix can be used to check the assumptions of MANOVA, i.e., that the variance and covariances are homogeneous across settings. From the SPLOM, there does not seem to be any systematic violations of the assumptions, which might require a variable transformation.

Analysis of Fear Symptoms of U.S. Soldiers using Item-Response Theory

COMBATDM data contains reports of fear symptoms by selected U.S. soldiers after being withdrawn from World War II combat. There are nine symptoms that are included for analysis and the number of soldiers in each profile of symptom is reported.

Variable	**Description**
COUNT	Number of soldiers in each profile of symptom
POUNDING	Violent pounding of the heart
SINKING	Sinking feeling of the stomach
SHAKING	Shaking or trembling all over
NAUSEOUS	Feeling sick at the stomach
STIFF	Cold sweat
FAINT	Feeling of weakness or feeling faint
VOMIT	Vomiting
BOWELS	Losing control of the bowels
URINE	Urinating in the pants

Determining which withdrawal fear symptoms are common to the soldiers after a combat and the probability of each taking place is useful in preparing the soldiers for future encounters.

Potential analyses include Test item analysis, factor analysis, multidimensional scaling, and cluster analysis.

Classical Test Item Analysis

The input is:

```
USE COMBATDM
TESTAT
  MODEL POUNDING.. URINE
  FREQUENCY COUNT
  IDVAR COUNT
  ESTIMATE/CLASSICAL
```

The output is:

Case frequencies determined by value of variable COUNT

Data Below are Based on 93 Complete Cases for 9 Data Items

Test Score Statistics

	Total	Average	Odd	Even
Mean	4.538	0.504	2.473	2.065
Standard Deviation	2.399	0.267	1.333	1.277
Standard Error	0.250	0.028	0.139	0.133
Maximum	9.000	1.000	5.000	4.000
Minimum	1.000	0.111	0.000	0.000
N of Cases	93	93	93	93

Internal Consistency Data

Split-half Correlation	: 0.690
Spearman-Brown Coefficient	: 0.816
Guttman (Rulon) Coefficient	: 0.816
Coefficient Alpha - All Items	: 0.787
Coefficient Alpha - Odd Items	: 0.613
Coefficient Alpha - Even Items	: 0.661

Approximate Standard Error of Measurement of Total Score for 15 z score Intervals

z-score	Total Score	N	Standard Error
-3.750	-4.458	0	.
-3.250	-3.258	0	.
-2.750	-2.059	0	.
-2.250	-0.860	0	.
-1.750	0.340	10	1.000
-1.250	1.539	16	1.000
-0.750	2.739	6	1.000
-0.250	3.938	29	1.390
0.250	5.137	10	1.095
0.750	6.337	8	1.000
1.250	7.536	8	0.000
1.750	8.735	6	1.000
2.250	9.935	0	.
2.750	11.134	0	.
3.250	12.334	0	.

Item Reliability Statistics

Item	Label	Mean	Standard Deviation	Item Total R	Item Reliability Index	Excl Item R	Excl Item Alpha
1	POUNDING	0.903	0.296	0.331	0.098	0.215	0.794
2	SINKING	0.785	0.411	0.499	0.205	0.354	0.782
3	SHAKING	0.559	0.496	0.678	0.336	0.539	0.757
4	NAUSEOUS	0.613	0.487	0.721	0.351	0.599	0.747
5	STIFF	0.538	0.499	0.693	0.346	0.559	0.754
6	FAINT	0.452	0.498	0.715	0.356	0.588	0.749
7	VOMIT	0.376	0.484	0.622	0.301	0.472	0.767
8	BOWELS	0.215	0.411	0.625	0.257	0.502	0.763
9	URINE	0.097	0.296	0.503	0.149	0.402	0.777

Logistic Test Item Analysis

The input is:

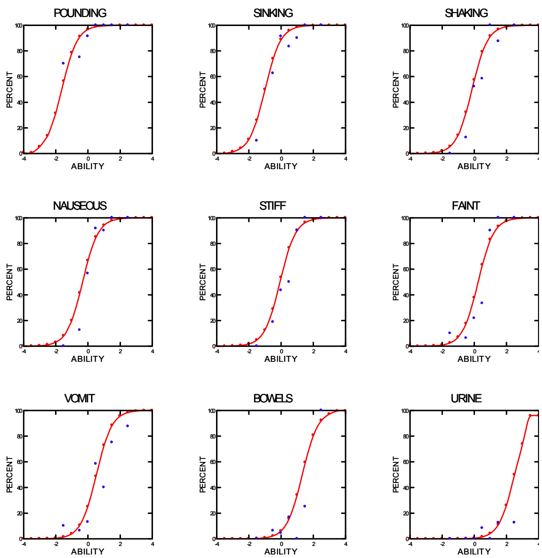
```
USE COMBATDM
TESTAT
  MODEL POUNDING.. URINE
  FREQUENCY COUNT
  IDVAR COUNT
  ESTIMATE/LOG1
```

The output is:

```
Case frequencies determined by value of variable COUNT
93 Cases were processed, each containing 9 items
6 Cases were deleted by editing for missing data or for zero or
perfect total scores after item editing.
0 Items were deleted by editing for missing data or for zero or
perfect total scores after item editing.
Data below are based on 87 Cases and 9 Items

Total Score Mean                                     :   4.230
Standard Deviation                                   :   2.164
-Log(Likelihood) Using Initial Parameter Estimates : 270.982

STEP 1 Convergence Criterion : 0.050

Stage 1: Estimate Ability with Item Parameter(s) Constant

        -Log
(Likelihood)   Change      LR
-----------------------------
     270.071   -0.911   2.486

Greatest Change in Ability Estimate was for Case 80

Change from Old Estimate : 0.134
Current Estimate         : 2.005

Stage 2: Estimate Item Parameter(s) with Ability Constant

        -Log
(Likelihood)   Change      LR
-----------------------------
     269.662   -0.409   1.505

Greatest Change in Difficulty Estimate was for Item BOWELS

Change from Old Estimate : 0.084
Current Estimate         : 1.301

Current Value of Discrimination Index : 1.206

STEP 2 Convergence Criterion : 0.050

Stage 1: Estimate Ability with Item Parameter(s) Constant

        -Log
(Likelihood)   Change      LR
-----------------------------
     269.590   -0.072   1.075

Greatest Change in Ability Estimate was for Case 87

Change from Old Estimate : 0.006
```

```
Current Estimate          : 2.011

Stage 2: Estimate Item Parameter(s) with Ability Constant

        -Log
(Likelihood)   Change      LR
-----------------------------
     269.549   -0.041   1.042

Greatest Change in Difficulty Estimate was for Item BOWELS

Change from Old Estimate : 0.032
Current Estimate          : 1.315

Current Value of Discrimination Index : 1.226
```

Latent Trait Model Item Plots



Sociology

World Population Characteristics

The *WORLDDM* data contains 1990 information on 30 countries and includes birth and death rates, life expectancies (male and female), types of government, whether mostly urban or rural, and latitude and longitude.

Variable	**Description**
COUNTRY$	Country name
BIRTH_RT	Number of births per 1000 people in 1990
DEATH_RT	Number of deaths per 1000 people in 1990
MALE	Years of life expectancy for males
FEMALE	Years of life expectancy for females
GOV$	Type of government
URBAN$	Rural or city
LAT	Latitude of the country's centroid
LON	Longitude of the country's centroid

Countries are often classified into categories (for example, developed or third world) based on certain socioeconomic criteria (one key group of criteria being population statistics). This data set contains such criteria for 30 countries of various regions and per capita income levels, allowing countries to be clustered according to population characteristics. In addition, variables such as the type of government and whether the country is mostly rural or urban may have an impact on these population characteristics.

Potential analyses include ANOVA, regression, cluster analysis, multidimensional scaling, and mapping.

Cluster Analysis

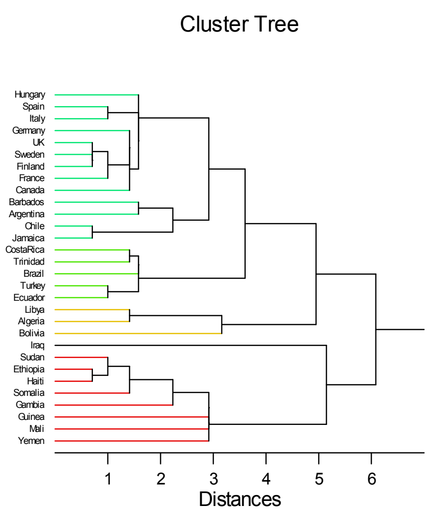
The input is:

```
USE WORLDDM
CLUSTER
  IDVAR COUNTRY$
  JOIN BIRTH_RT DEATH_RT
```

The output is:

```
Distance Metric is Euclidean Distance
Single Linkage Method (Nearest Neighbor)

  Clusters Joining     at Distance   No. of Members
---------------------------------------------------
Sweden     Finland           0.707                2
UK         Sweden            0.707                3
Haiti      Ethiopia          0.707                2
Jamaica    Chile             0.707                2
France     UK                1.000                4
Italy      Spain             1.000                2
Haiti      Sudan             1.000                3
Ecuador    Turkey            1.000                2
France     Germany           1.414                5
Canada     France            1.414                6
Algeria    Libya             1.414                2
Somalia    Haiti             1.414                4
Trinidad   CostaRica         1.414                2
Italy      Canada            1.581                8
Hungary    Italy             1.581                9
Barbados   Argentina         1.581                2
Brazil     Trinidad          1.581                3
Ecuador    Brazil            1.581                5
Somalia    Gambia            2.236                5
Jamaica    Barbados          2.236                4
Jamaica    Hungary           2.915               13
Mali       Guinea            2.915                2
Somalia    Mali              2.915                7
Yemen      Somalia           2.915                8
Algeria    Bolivia           3.162                3
Jamaica    Ecuador           3.606               18
Jamaica    Algeria           4.950               21
Yemen      Iraq              5.148                9
Jamaica    Yemen             6.083               30
```

Clustering Countries by Birth and Death Rates.

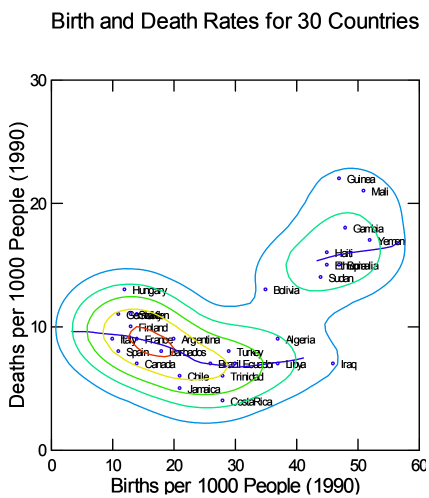


Kernel Densities Ellipses and Modal Smoothers

The input is:

```
USE WORLDDM
BEGIN
PLOT DEATH_RT*BIRTH_RT / XMIN=0, XMAX=60, YMIN=0, YMAX=30,
                         XTICK=6, SYMBOL=1, SIZE=.5,
                         LABEL=COUNTRY$, SMOO=MODE,
                         XLAB="Births per 1000 People (1990)",
                         YLAB="Deaths per 1000 People (1990)"
DEN .*DEATH_RT*BIRTH_RT / XMIN=0, XMAX=60, YMIN=0, YMAX=30,
                 XTICK=6, KERNEL, CONTOUR, ZTICK=10, ZPIP=0,
                 AX=0, SC=0,
                 TITLE="Birth and Death Rates for 30 Countries"
END
```

The output is:



Statistics

Instructional Methods

The INSTRDM data consists of measures of achievement on a biology exam for two groups of students—one group simply told to study everything from a biology text in general and the other given terms and concepts that they were expected to master. An additional covariate, the student's aptitude, is also included in the data set.

Variable	Description
STUDENT	Student ID
INSTRUCT$	Type of instruction given
INSTRUCT	Coded variable for INSTRUCT$
APTITUDE	Student's underlying ability to learn
ACHEIVE	Student's score on the exam

From an education-theory standpoint, this data set is interesting because it demonstrates the effect on "achievement" due to different study instructions. A student is likely to show a higher level of achievement when given specific instructions on what to know for an exam than a student who gets only general instructions. From a statistical standpoint, it demonstrates the importance of considering covariates when using ANOVA models. A straight ANOVA of ACHIEVE on INSTRUCT shows no significance at the 95% confidence level, but after separating out some of the variance using the covariate APTITUDE in an ANCOVA model, there is a significant difference between instruction groups.

Potential analyses include ANOVA, ANCOVA, and regression.

Analysis of Covariance

The input is:

```
USE INSTRDM
GLM
  CATEGORY INSTRUCT / EFFECT
  MODEL ACHIEVE = CONSTANT + INSTRUCT + APTITUDE
  ESTIMATE
```

The output is:

Effects coding used for categorical variables in model.
The categorical values encountered during processing are

Variables	Levels	
INSTRUCT (2 levels)	1.000	2.000

Dependent Variable	ACHIEVE
N	20
Multiple R	0.760
Squared Multiple R	0.578

Estimates of Effects B = (X'X)-1X'Y

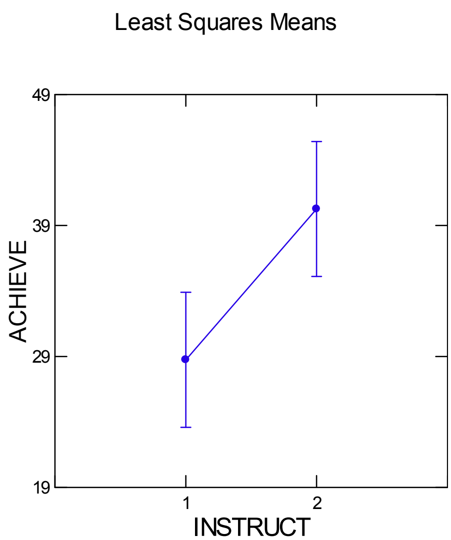
Factor	Level	ACHIEVE
CONSTANT		9.646
INSTRUCT	1	-5.755
APTITUDE		0.502

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
INSTRUCT	641.424	1	641.424	10.915	0.004
APTITUDE	961.017	1	961.017	16.354	0.001
Error	998.983	17	58.764		

Least Squares Means

Factor	Level	LS Mean	Standard Error	N
INSTRUCT	1	28.745	2.444	10.000
INSTRUCT	2	40.255	2.444	10.000



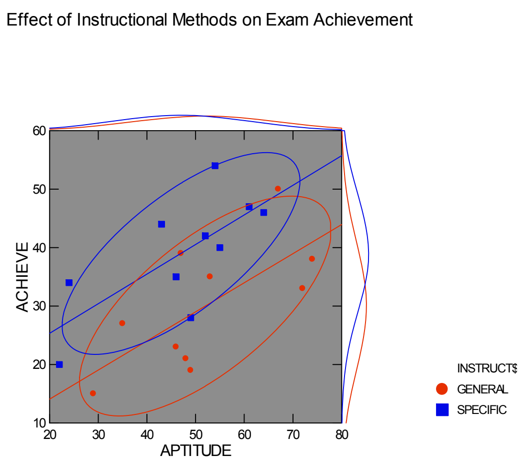
```
Durbin-Watson D Statistic    ¦  2.197
First Order Autocorrelation ¦ -0.171
```

Scatterplot

The input is:

```
USE INSTRDM
PLOT ACHIEVE * APTITUDE / GROUP=INSTRUCT$, OVERLAY,
BORDER=NORMAL, ELL, SMOOTH=LINEAR, FCOLOR=GRAY, SYMBOL={1, 8},
FILL,
TITLE="Effect of Instructional Methods on Exam Achievement"
```

The output is:



Toxicology

Concentration of nicotine sulfate required to kill 50% of a group of common fruit flies

The *WILLMSDM* data contains the results of a bioassay conducted to determine the concentration of nicotine sulfate required to kill 50% of a group of common fruit flies. The experimenters recorded the number of fruit flies that are killed at different dosage levels.

Variable	Description
RESPONSE	The dependent variable, which is the response of the fruit fly to the dose of nicotine sulfate (stimulus).
LDOSE	The logarithm of the dose.
COUNT	The number of fruit flies with that response.

In bioassay, it is common to estimate the dose required to kill 50% of a target population. For example, a toxicity experiment may be conducted to establish the concentration of nicotine sulfate required to kill 50% of a group of common fruit flies. The goal is to identify the level of stimulus required to induce a 50% response rate, where response may be any binary outcome variable and the stimulus is a continuous variate. In bioassay, stimuli include drugs, toxins, hormones, and insecticides; responses include death, weight gain, bacterial growth, and color change.

Potential analyses include logistic regression and survival analysis.

Logistic regression

The input is:

```
USE WILLMSDM
FREQ=COUNT
LOGIT
  MODEL RESPONSE=CONSTANT+LDOSE / REF = 0
  ESTIMATE
  QNTL
  LET LDOSEB=LDOSE-.4895
  MODEL RESPONSE=LDOSEB / REF = 0
  ESTIMATE
  LET LDOSEB=LDOSE+2.634
  MODEL RESPONSE=LDOSEB / REF = 0
  ESTIMATE
```

The output is:

```
Case frequencies determined by value of variable COUNT
The categorical values encountered during processing are

Variables           ¦    Levels
--------------------+--------------
RESPONSE (2 levels) ¦ 0.000   1.000

Dependent Variable       : RESPONSE
Analysis is Weighted by  : COUNT
Sum of Weights           : 25.000
Input Records            : 9
Records for Analysis     : 9
```

Sample Split

```
Category ¦ Count       Weighted Count
---------+---------------------------
0        ¦ RESPONSE            15.000
1        ¦ REFERENCE           10.000
```

Log-Likelihood Iteration History

```
Log-Likelihood at Iteration1 ¦ -17.329
Log-Likelihood at Iteration2 ¦ -13.277
Log-Likelihood at Iteration3 ¦ -13.114
Log-Likelihood at Iteration4 ¦ -13.112
Log-Likelihood at Iteration5 ¦ -13.112
Log-Likelihood               ¦ -13.112
```

Information Criteria

```
AIC            ¦ 30.224
Schwarz's BIC  ¦ 30.618
```

Parameter Estimates

```
          ¦                                                   95% Confidence Interval
Parameter ¦ Estimate   Standard Error        Z   p-value          Lower         Upper
----------+--------------------------------------------------------------------------
CONSTANT  ¦   -0.564            0.496   -1.138     0.255         -1.536         0.408
LDOSE     ¦   -0.919            0.394   -2.334     0.020         -1.691        -0.147
```

Odds Ratio Estimates

```
          ¦                                 95% Confidence Interval
Parameter ¦ Odds Ratio   Standard Error          Lower         Upper
----------+---------------------------------------------------------
LDOSE     ¦      0.399            0.157          0.184         0.863
```

```
Overall Model Fit

Log-likelihood of Constant only Model ¦ -16.825
Log-likelihood of Full Model          ¦ -13.112
Chi-square value                      ¦   7.427
df                                    ¦       1
p-value                               ¦   0.006

R square Measures

McFadden's Rho-squared ¦ 0.221
Cox and Snell R-square ¦ 0.562
Naglekerke's R-square  ¦ 0.576
```

Evaluation Vector

```
  CONSTANT ¦ 1.000
  LDOSE    ¦ VALUE
```

Quantile Table

Probability	LOGIT	LDOSE	95% Bounds Upper	Lower
0.999	6.907	-8.127	-4.486	-49.055
0.995	5.293	-6.372	-3.508	-38.136
0.990	4.595	-5.612	-3.081	-33.416
0.975	3.664	-4.599	-2.503	-27.126
0.950	2.944	-3.817	-2.046	-22.281
0.900	2.197	-3.004	-1.552	-17.266
0.750	1.099	-1.809	-0.731	-9.987
0.667	0.695	-1.369	-0.347	-7.392
0.500	0.000	-0.613	0.746	-3.364
0.333	-0.695	0.142	3.551	-1.047
0.250	-1.099	0.582	5.928	-0.445
0.100	-2.197	1.777	13.053	0.530
0.050	-2.944	2.590	18.042	1.050
0.025	-3.664	3.372	22.875	1.519
0.010	-4.595	4.385	29.157	2.105
0.005	-5.293	5.145	33.873	2.536
0.001	-6.907	6.900	44.788	3.518

Case frequencies determined by value of variable COUNT
The categorical values encountered during processing are

Variables	Levels	
RESPONSE (2 levels)	0.000	1.000

Dependent Variable : RESPONSE
Analysis is Weighted by : COUNT
Sum of Weights : 25.000
Input Records : 9
Records for Analysis : 9

Sample Split

Category	Count	Weighted Count
0	RESPONSE	15.000
1	REFERENCE	10.000

Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-17.329
Log-Likelihood at Iteration2	-15.060
Log-Likelihood at Iteration3	-15.032
Log-Likelihood at Iteration4	-15.032
Log-Likelihood at Iteration5	-15.032
Log-Likelihood	-15.032

Information Criteria

AIC	32.064
Schwarz's BIC	32.261

Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95% Confidence Interval Lower	Upper
LDOSEB	-0.631	0.323	-1.950	0.051	-1.265	0.003

Odds Ratio Estimates

Parameter	Odds Ratio	Standard Error	95% Confidence Interval Lower	Upper
LDOSEB	0.532	0.172	0.282	1.003

Case frequencies determined by value of variable COUNT
The categorical values encountered during processing are

Variables	Levels	
RESPONSE (2 levels)	0.000	1.000

```
Dependent Variable       : RESPONSE
Analysis is Weighted by  : COUNT
Sum of Weights           : 25.000
Input Records            : 9
Records for Analysis     : 9
```

Sample Split

Category	Count	Weighted Count
0	RESPONSE	15.000
1	REFERENCE	10.000

Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-17.329
Log-Likelihood at Iteration2	-15.055
Log-Likelihood at Iteration3	-15.032
Log-Likelihood at Iteration4	-15.032
Log-Likelihood at Iteration5	-15.032
Log-Likelihood	-15.032

Information Criteria

AIC	32.064
Schwarz's BIC	32.262

Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95% Confidence Interval Lower	Upper
LDOSEB	-0.312	0.159	-1.968	0.049	-0.624	-0.001

Odds Ratio Estimates

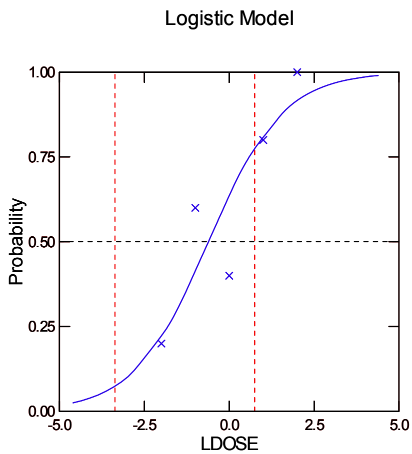
Parameter	Odds Ratio	Standard Error	95% Confidence Interval Lower	Upper
LDOSEB	0.732	0.116	0.536	0.999

Plot of Logistic Model

The input is:

```
USE WILLMSDM
FREQ=COUNT
LOGIT
MODEL RESPONSE = CONSTANT + LDOSE / REF = 0
ESTIMATE
SAVE QUANT
QNTL

REM CREATES PLOT OF LOGISTIC MODEL WITH LIMIT LINES ADDED AT THE
REM UPPER
REM AND LOWER LIMITS FOR THE LDOSE VALUE CORRESPONDING TO A
REM PROBABILITY HAS .50
USE QUANT
BEGIN
PLOT PROB*LDOSE / SIZE=0 XLAB=" " YLAB=" " XLIMIT={-3.364,0.746},
                  XMIN=-5 XMAX=5 XTICK=4,
                  ACOLOR=RED YTICK=4, YMAX=1 YMIN=0
PLOT PROB*LDOSE / SIZE=0 SMOOTH=SPLINE TENSION =0.500,
                  XMIN=-5 XMAX=5 XTICK=4 XLAB="LDOSE",
                  YLAB="Probability" YLIMIT=0.5,
                  YTICK=4 YMAX=1 YMIN=0
USE WILLMSDM
LET PDEAD=COUNT/5
SELECT (RESPONSE=1)
PLOT PDEAD*LDOSE / SYM=2 YTICK=4 YMAX=1 YMIN=0 XMIN=-5,
                   XMAX=5 XTICK=4 XLAB=" " YLAB=" " ,
                   SCALES=NONE TITLE="Logistic Model"
END
```

The output is:



Data References

Anthropology Data Sources

Original Source. Thomson, A. and Randall-McIver, R. (1905). *Ancient races of the Thebaid*. Oxford: Oxford University Press.

Data Reference. Hand, D. J., Daly, F., Lunn, A.D., McConway, K.J., and Ostrowski, E. (1994). *A handbook of small data sets*. New York: Chapman & Hall. pp. 299-301.

Manly, B.F.J. (1986). *Multivariate statistical methods*. New York: Chapman & Hall.

STATLIB. *http://lib.stat.cmu.edu/DASL/Datafiles/EgyptianSkulls.html*

Astronomy Data Source

Original Source. Waldmeir, M. (1961). The sunspot activity in the years 1610-1960. *Zurich: Schulthess and International Astronomical Union Quarterly Bulletin on Solar Activity*. Tokyo.

Data Reference. Andrews, D.F. and Herzberg, A.M. (1985). Data, pp. 67-76. Springer-Verlag.

Biology Data Source

Data Source. Carey, J.R., Liedo, P., Orozco, D., and Vaupel, J.W. (1992). Slowing of mortality rates at older ages in large med fly cohorts. *Science*, pp. 258, 457-461.
Data Reference. STATLIB http://lib.Stat.cmu.edu/DASL/Datafiles/Medflies.html
Data Source. Allison, T. and Cicchetti, D. V. (1976). Sleep in mammals: Ecological and constitutional correlates. *Science*, pp. 194, 732-734.

Chemistry Data Sources

Original Source. Adapted from a conference session on statistical computing (Greco et al., 1982).
Data Reference. Wilkinson L. and Engelman, L. (1996). SYSTAT 6.0 for Windows: *Statistics*, pp. 487-488, SPSS Inc.

Engineering Reference

Devor, R.E., Chang, T. and Sutherland, J.W. (1992). *Statistical quality design and control*, pp. 756-761. New York: MacMillan.

Environmental Science Sources

Original Source. Lange, Royals, and Connor. (1993). Transactions of the American fisheries society.
Data Reference. STATLIB http://lib.Stat.cmu.edu/DASL/Datafiles/MercuryinBass.html

Genetics Data Sources

Data Source. Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. 2nd edition, New York: John Wiley & Sons.
McLachlan, G.J. and Krishnan. T. (1997). *The EM algorithm and extensions*. New York: John Wiley & Sons.

Manufacturing Data Sources

Original Source. Messina, W.S. (1987). *Statistical quality control for manufacturing managers*. New York: Wiley.

Data Reference. Stenson, H. and Wilkinson, L. (1996). SYSTAT 6.0 for Windows: *Graphics, SPSS*, pp.291-369.

Medicine Data Sources

Original Source. Cameron, E. and Pauling, L. (1978). Supplemental ascorbate in the supportive treatment of cancer: Reevaluation of prolongation of survival times in terminal human cancer. *Proc. Natl. Acad. Sci. U.S.A*, 75, 4538-4542.

Data Reference. Andrews, D.F. and Herzberg, A.M. (1985). *Data*, pp. 203-207. Springer-Verlag.

Medical Research Data Reference

Wilkinson L. and Engelman, L. (1996), SYSTAT 7.0: *New Statistics*, pp.235, SPSS Inc.

Psychology Data Reference

Wilkinson, L., Blank, G. and Gruber, C. (1996). *Desktop data analysis with SYSTAT*. Upper Saddle River, NJ: Prentice Hall, p.454.

Stroufer, S.A., Guttmann, L., Suchman, E.A., Lazarsfeld, P.F., Staf, S.A., and Clausen, J. A. (1950). *Measurement and prediction*. Princeton, N. J.: Princeton University Press.

Sociology Data Reference

Wilkinson, L., Blank, G. and Gruber, C. (1996). *Desktop data analysis with SYSTAT*. Upper Saddle River, NJ: Prentice Hall, p.738.

Statistics Data Sources

Original Source. Huitema, B.E. (1980). *The Analysis of covariance and alternatives*. New York: John Wiley & Sons.

Data Reference. Wilkinson, L., Blank, G., and Gruber, C. (1996). *Desktop data analysis with SYSTAT.* Upper Saddle River, NJ: Prentice Hall, p. 442.

Toxicology Data Source

Hubert J. J. (1991). *Bioassay*. 3rd ed. Dubuque, Iowa: Kendall Hunt.

Appendix
9

Data Files

SYSTAT software comes with a folder of data files, which can be accessed through the File => Open => Data dialog. The folder contains over 350 files of data used in the nearly 600 examples provided in the user manual and online help. This Appendix gives details of these files, with sources of data, a brief description of the study which generated the data, and a description of the variables in the file.

These data files not only contain the data, but also a great deal of information on the data file. The information given in this Appendix is available in the data file itself. When you have clicked on the data file name in the dialog and opened it in the Data editor, by hovering the mouse over the corner rectangle (the top left cell) you will see the general information on the file. Then in the Variable Properties dialog of a variable (which can be opened by Data => Variable Properties with the variable name selected by clicking on it or by simply right-clicking on the variable name in the data file), in the Comments box at the bottom, you will see information on the variable. This information on the variable is also seen as a tooltip by simply moving the mouse over the variable name.

For a data file you create, you may construct this general file information by filling it in the File Comments dialog, which can be opened by right-clicking on the file name in the Data editor, or on the top left cell. Information on individual variables may be entered in the Comments box of the Variable Properties dialog.

The data file contains even more information, which can be seen by clicking the Variable tab in the Data editor, which opens the Variable editor. This contains information on each variable as to its name, label, value labels, type (string or numeric), categorical or not, the number of characters, number of decimals, display type and comments. It also contains information on which variables are involved in case selection, has been chosen to be a frequency or a weight variable, for BY groups analysis, a category variable or an order variable.

The following data files are 'Read only':

ACCIDENT• Jobson (1992). The data set relates to automobile accidents in Alberta, Canada. The variables are – *SEATBELT$, IMPACT$, INJURY$, DRIVER$, FREQ.*

ADAPTOR• The 'adaptor body' is one of the components of a machine. Its outer diameter is denoted by *DIA*. The data set contains the *DIA* of 16 adaptor bodies produced over a period of 16 hours one in each hour. The total time period is divided into two periods of eight hours each and the variable '*EIGHT* ' takes value 1 or 2 depending upon the period of its production. Similarly variables '*FOUR*' and '*TWO*' are constructed. Thus the 'design' is a nested one with 'four' nested inside '*EIGHT*' and '*TWO*' nested inside '*FOUR*'. The variables are - *DIA, EIGHT, FOUR, TWO.*

ADJADAPTOR• The data set consists of the outer diameter of a component named adaptor body, before and after correction. The two variables are – *BEFORE, AFTER.*

ADMIRE• Cohen and Brook (1987). In a large-scale longitudinal study of childhood and adolescent mental health, data were obtained on personal qualities that the subjects admired and what they thought other children admired, as well as the sex and age of the subjects. The admired qualities were organized into scales for antisocial, materialistic, and conventional values for the self and as ascribed to others. In one phase of the investigation, the researchers wanted to study the relationship between the sets of self versus others. However, several of these scales exhibited sex differences, were nonlinearly (specifically quadratically) related to age, and/or were differently related to age for the sexes. For the self-other association to be assessed free of the confounding influence of age, sex, and their interactions, it was desirable to partial those effects from the association. Using SYSTAT, the variables *SEX* times *AGE* and their squares were created. The variables are – *ID$*, *ANTISO_S*, *MATER_S*, *CONVEN_S*, *ANTISO_O*, *MATER_O*, *CONVEN_O*, *AGE*, *SEX*, *AGESQ*, *SEXAGE*, *SEXAGESQ*.

ADMIT• Graduate Record Examination Verbal (*GREV*) and Quantitative (*GREQ*) scores with a binary indicator of whether or not a student was awarded a Ph.D. (*PHD$*) in a graduate psychology department. The variables are – *YEAR, GPA, GREV, GREQ, GRE, PHD, GROUP, N, PHD$.*

AEROSOL• Beckman, Nachtsheim and Cook (1987). This is a study of high efficiency particulate air HEPA cartridges. For this two aerosol types (*AEROSOL*) were used to test the three HEPA respirator filters (*FILTER*) from each of two different manufacturers (*MANUFACTURER*).

AFIFI• Afifi and Azen (1974). The dependent variable, *SYSINCR*, is the increase in systolic blood pressure after administering one of four different drugs *(DRUG)* to patients with one of three different diseases *(DISEASE)*. Patients were assigned randomly to one of the four possible drugs.

AGE1• The data set consists of two variables *AGE$* and *SEX$.*

AGESEX• U.S. Census (1980). These data show the distribution of *MALES* and *FEMALES* within age groups. The variable *AGE* labels each age group by the upper age limit of its members.

AGESTAT• The data set is randomly generated data consisting of two variables *AGE* and *SEX$*.

AGR1 and **AGR2•** The data sets consist of a hypothetical agricultural data, where the yields of crops are related to the soil type and the type of fertilizer used. The variables are - *YIELD, FERTILIZER* and *SOIL*.

AIAG• Breyfogle (2003). This data set originated from Automotive Industry Action Group (AIAG)(1995). The data set deals with measures of a critical quality characteristic (*MEASURE*) of 80 samples. 5 samples collected in each of 16 subgroups *(SUBGROUP)*.

AIRCRAFT• Bennett and Desmarais (1975). These data show amplitude of vibration *(FLUTTER)* versus time *(TIME)* in an aircraft wing component.

AIRLINE• Box et al. (1994). The variable *PASS* contains monthly totals of international airline passengers for 12 years beginning in January, 1949.

AKIMA• Akima (1978). These data are topological measurements of a three-dimensional surface using the variables *X*, *Y*, and *Z*.

AM• Borg and Lingoes (1987), adapted from Green and Carmone (1970). This unfolding data set contains similarities only between the points delineating ‘A’ and ‘M,’ and these similarities are treated only as rank orders. Variables include *A1* through *A16 and ROW$*.

ANNEAL• Brownlee (1960). The experiment seeks to compare two different annealing methods for making cans. Three coils (*COIL*) of material were selected from the populations of coils made by each of the two methods (*METHOD*). Pair of samples was drawn from each of two locations (*LOCATION*) on the coil. The response is the life (*LIFE*) of the can.

ANSFIELD• Ansfield et al. (1977). This study examines the effects *(RESPONSE$)* of treatments *(TREAT$)* on two patient groups *(CANCER$)*, those with cancer of the colon or rectum and those with breast cancer. *NUMBER* gives the number of patients in each cancer/treatment/response group.

ANXIETY• Data are from a National Longitudinal Survey of Young Men conducted in 1979. The data set has been extracted from data set *NLS*.

BANK• The data set consists of the description of bank employees. The variables are –

WEIGHT	
ID	Employee code
SALBEG	Beginning salary
SEX	Sex of employee 0 Male 1 Female
TIME	Job seniority(in months)
AGE	Age of employee(in years)
SALNOW	Current salary
EDLEVEL	Educational level
WORK	Work experience
JOBCAT	Employment category 1 Clerical 2 Office trainee 3 Security officer 4 College trainee 5 Exempt employee 6 MBA trainee 7 Technical
MINORITY	Minority classification 0 White 1 Nonwhite
SEXRACE	Sex & race classification 1 Black Females 2 White Females 3 Black Males 4 White Males

BANKRUPTCY• Simonoff (2003). The data were collected on 25 telecommunication firms that were declared bankrupt during the period May 2000-January 2002 and 25 telecommunication firms that were not declared bankrupt from December 2000 in their issued financial statements. The potential predictors are based on five banking financial ratios:

WCTA	Working Capital as percentage of total assets.
RETA	Retained earnings as percentage of total assets.
EBITTA	Earning before interest and taxes as a percentage of total assets.
STA	Sales of total assets (in percentages).
BVEVL	Book value equity divided by book value of total liabilities.

BARLEY• Fisher (1935). The data are the yields of 10 varieties of barley in two years (1931 and 1932) at 6 sites in the Midwestern US. The variables are – *Y1931, Y1932, VARIETY$, SITE$*.

BBD• Myers & Montgomery (2002). This data set contains observations on viscosity (*VISCOSITY*) at different level combinations of the three factors:temperature (*TEMP*), agitation (*AGITATION*) and rate of addition (*RATE*). Each factor has 3 levels.

BIRTHS• Walser (1969). The data set consists of information on the *FREQUENCY* of births in each *MONTH* (labeled as 1,2,…,12) of a year in the University Hospital of Basel, Switzerland.

BIRTHS2• Conover (1999). These data were collected in a survey conducted in 7 hospitals of a certain city over a 12-month period divided into 4 seasons (*SEASON$*), and the numbers of newborn babies (*BIRTHS*) in each season were obtained. The variables are – *BIRTHS, SEASON$, HOSPITAL$*.

BIT5• The file contains five-item binary profiles fitting a two-dimensional structure perfectly. Variables in the SYSTAT data file are: *X(1)……X(5).*

BLOCK• Neter et al. (2004). These data comprise a randomized block design. Five blocks of judges *(BLOCK)* analyzed three treatments *(TREAT)*. Subjects (judges) are stratified within blocks, so the interaction of blocks and treatments cannot be analyzed, and the outcome of the analysis is *JUDGMENT*.

BLOCKCCD• Myers & Montgomery (2002). This data set contains observations on the yield of a chemical process (*YIELD*) at different level combinations of two factors, viz. time (*TIME*) and temperatute (*TEMP*) on 14 experimental units. However two different batches of raw materials were used. The variable *BLOCK* defines the different batches.

BOARDS• Montgomery (2005). It is an aggregated data set on the number of nonconformities found in 26 successive samples of 100 circuit boards. For convenience, the sample unit (or inspection unit) is defined as 100 boards. That is, although each sample contains 100 boards, each sample is considered a sample of size 1 from a Poisson distribution. The variables are-

SAMPLE	Identifier
DEFECTS	A total count of the number of defects in each group of 100 Boards

BOD• Bates and Watts (1988). Marske created these data from stream samples in 1967. Each sample bottle is inoculated with a mixed culture of microorganisms, sealed, incubated, and opened periodically for analysis of dissolved oxygen concentration. The variables are *DAYS* and *BOD*.

BOOKPREF• Conover (1999). The data set consists of the number of books sold in a week in 12 bookstores of four booksellers .The variables are – *BOOKS, STORE, BOOKSELLER*.

BOSTON• Belsley, Kuh, and Welsch (1980). The data set is Boston housing prices, used in Breiman et al. (1984). The variables are - *CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV*.

BOXES• Messina (1987). The ohms of electrical resistance in computer boxes are measured for five randomly selected boxes from each of 20 days of production. Thus, each *SAMPLE* contains five observations of resistance in *OHMS* for each of 20 days *(DAY)*.

BP• Hand et al. (1996). The data set gives the supine systolic and diastolic blood pressures (mm Hg) for 15 patients with moderate essential hypertension, immediately before and two hours after administering the drug, captopril. The variables are-

SYSBP_BEFORE	Systolic blood pressure (mm Hg) with moderate essential hypertension before administering the drug, captopril
SYSBP_AFTER	Systolic blood pressure (mm Hg) with moderate essential hypertension 2 hours after administering the drug, captopril
DIABP_BEFORE	Diastolic blood pressure (mm Hg) with moderate essential hypertension before administering the drug, captopril
DIABP_AFTER	Diastolic blood pressure (mm Hg) with moderate essential hypertension 2 hours after administering the drug, captopril

BRODLIE• Brodlie (1980). These data are *X* and *Y* coordinates taken from a figure in Brodlie's discussion of cubic spline interpolation.

BULB• Mendenhall et al. (2002). A manufacturer of industrial light bulbs tries to control the variability in length of life of the light bulbs so that standard deviation is less than 150 hours. The data consists of *LIFETIME* of 20 bulbs.

BUSES• Davis (1977). These data count the number of buses failing *(COUNT)* after driving 1 of 10 distances *(DISTANCE)*.

CANCER• Morrison (1990); Bishop et al. (1975). These studies examined breast cancer patients in three diagnostic centers *(CENTER$)*, three age groups *(AGE)*, whether they survived after three years post-diagnosis *(SURVIVE$)*, and the inflammation type (minimum/maximum) and appearance of the tumor *(TUMOR$)* (malignant/benign). The variable *NUMBER* contains the number of women in each cell.

CANCERDM• Cameron and Pauling (1978). The data set contains information from a study of the effects of supplemental vitamin C as part of routine cancer treatment for 100 patients and 1000 controls (10 controls for each patient).

CASE	Case ID
ORGAN$	Organ affected by cancer
SEX$	Sex of patient

AGE	Age of patient
SURVATD	Survival of patient measured from first hospital attendance
CNTLATD	Survival of control group from first hospital attendance
SURVUNTR	Survival of patient from time cancer deemed untreatable
CNTLUNTR	Survival of control from time cancer deemed untreatable
LOGSURVA	Logarithm of *SURVATD*
LOGCNTLA	Logarithm of *CNTLATD*
LOGSURVU	Logarithm of *SURVUNTR*
LOGCNTLU	Logarithm of *CNTLUNTR*

CARDOG• Wilkinson (1975). This data set contains the INDSCAL configurations of the scalings of cars and dogs. The variables are – *CAR$, DOG$, C1 ,C2, D1, D2.*

CARS• The data set reflects the attributes of the selected performance cars. The variables are – *ACCEL, BRAKE, SLALOM, MPG, SPEED, NAME$.*

CEMENT• Birkes and Dodge (1993). The data set consists of four kinds of ingredients *INGREDIENT1, INGREDIENT2, INGREDIENT3, INGREDIENT4* corresponding to the temperature (*HEAT)*.

CHOICE• McFadden (1979). The data set consists of hypothetical data .The *CHOICE* variable represents the three transportation alternatives (*AUTO, POOL, TRAIN*) each subject prefers. The first subscripted variable in each *CHOICE* category represents *TIME* and the second, *COST*. Finally, *SEX$* represents the gender of the chooser. *AGE* represents the age of the chooser.

CHOLESTEROL• The data set records the age and blood cholesterol levels for two groups of women. Women in the first group use contraceptive pills; women in the second group do not. A *PILL* value of 1 indicates that the woman takes the pill; a value of 2 indicates that she does not. Each case has the cholesterol value *CHOL* for a pill user and for her age-matched control *AGE*.

CITIES• Hartigan (1975). The data set is a dissimilarity matrix consisting of airline distances in hundreds of miles between ten global cities: *BERLIN, BOMBAY, CAPETOWN, CHICAGO, LONDON, MONTREAL, NEW YORK, PARIS, SANFRAN,* and *SEATTLE.*

CITYTEMP• These data consist of low and high July temperatures for eight U.S. cities in 1992.

CLINCOV• Hocking (2003). This example is based on a clinical data set where a pharmaceutical firm wants to test a new drug for a particular disease. The response is a measure of the improvement in the patients' status. A sample consisting of three clinics (*CLINIC*) is selected at random from a large population of clinics. From each clinic a sample of ten patients with the particular disease are selected. The drug is applied to each patient and we record the response (*Y*) of the drug as well as a relevant physical characteristic (*Z*) for each patient.

CLOTH• Montgomery (2005). Here, the occurrences of nonconformities (DEFECTS) in each of 10 rolls of dyed cloth were counted (ROLL). The rolls were not all the same size in square meters. Thus, the sample unit was defined as 50 square meters of cloth, and roll sizes were expressed in these units (UNITS).

COBDOUG• Judge et al. (1988). The data set is related to the Cobb-Douglas production function in Econometrics. The Cobb-Douglas Production function considers the effect of Labor (L) and Capital invested (K) over the output (Q). The data set consists of 20 observations containing the variables *Y, X1 and X2*, where we have *Y=lnQ* and *X1=lnL* and *X2=lnK*.

CODDER• These data contain the percentage of reader attention *(PERCENT)* in a certain geographical area *(LOCUS$)* for the local newspaper.

COFFEE• Hand et al.(1996). The data set contains the prices (in pence) of a 100gm pack of a particular brand of instant coffee, on sale in 15 different shops and amount (in gm) per pence in Milton Keynes on the same day in 1981. The variables are – *PRICE*, *GM_PER_PENCE*.

COLAS• Schiffman, Reynolds, and Young (1981). These data consist of judgments by 10 subjects of the dissimilarity (0–100) between pairs of colas, including *DIETPEPS, RC, YUKON, PEPPER, SHASTA, COKE, DIETPEPR, TAB, PEPSI, and DIETRITE.*

COLOR• These data provide the proportions of *RED, GREEN,* and *BLUE* that will produce the color specified in *COLOR$*.

COLRPREF• The data set contains color preferences *(RED, ORANGE, YELLOW, GREEN, BLUE)* among 15 people *(NAME$)* for five primary colors.

COMBAT• Stouffer et al. (1950). This data set contains reports of fear symptoms by selected U.S. soldiers after being withdrawn from World War II combat. Nine symptoms are included for analysis, and the number of soldiers in each profile of symptom is reported. The variables are-

COUNT	Number of soldiers in each profile of symptom
POUNDING	Violent pounding of the heart
SINKING	Sinking feeling in the stomach
SHAKING	Shaking or trembling all over
NAUSEOUS	Feeling sick to the stomach
STIFF	Cold sweat
FAINT	Feeling of weakness or feeling faint
VOMIT	Vomiting
BOWELS	Loss of bowel control
URINE	Loss of urinary control

COMFORT• Milliken and Johnson (1992). In an experiment the effects of temperature on the comfort level of 18 men and 18 women was carried out using nine environmental chambers. Three different temperatures (65F, 70F and 75F) were assigned to three randomly selected chambers. Two randomly selected men and two randomly selected women were assigned to each chamber. The comfort of each person was measured after three hours in a scale of 1 to 15, where 1= cold, 8= comfortable and 15= hot. The variables are - *TEMP, GENDER, PERSON, CHAMBER, COMFORT.*

COMPUTER• Montgomery (2005). The following data represent the results of inspecting all units of a personal computer produced for 10 consecutive days (*DAY*). *UNITS* are the number of computers inspected each day, and *NONCON* is the number of nonconforming units found.

CONDENSE• Messina (1987). The data file contains nonconformance data (defects) for 15 lots of condensers. *LOT$* is lot number, *TYPE$* is type of defect, and *TALLY* is the frequency of a particular defect in a particular lot. One thousand condensers were inspected in each lot.

CORK • Rao(2002). Observations are obtained on 28 trees for thickness of cork borings in the NORTH(N), EAST(E),SOUTH(S), and WEST(W) directions. The problem is to examine whether the bark deposit is same in all the directions. We may consider the three characters (contrast)

U1 = (N +S) - (E + W)
U2 = N - S
U3 = E - W

CORN• The data set gives the amount of inorganic phosphorous (X1), organic phosphorous (X2) present in the soil, and the plant-available phosphorous (Y) of corn grown in the soil.

COVAR• Winer (1971). Winer uses this artificial data set in an analysis of covariance in which *Y* is the dependent variable, *X* is the covariate, and *TREAT* is the treatment.

COVSTRUCT• It is a hypothetical data. The variables are-*P*, *Q*, *Y*.

COX• Cox (1970). These data record tests for failures among objects after certain times *(TIME)*. *FAILURE* is the number of failures, and *COUNT* is the total number of tests.

CRABS• Wilkinson (2005). These data record the location of 23 fiddler-crab holes in an 80 x 80 centimeter area of the Pamet River marsh in Truro, Massachusetts.The variables are-*CRAB*,*X*,*Y*

CRIMERW• Clausen (1998). These data show the information case-by-case about crimes in three different areas in Norway. The following is a list of the three different areas and three crimes. The SYSTAT names are within parentheses.

PLACE$	CRIME$
Mid Norrway *(Mid N)*	*Burglary*
North Norway (*NorthN*)	*Fraud*
Oslo Area (*Oslo*)	*Vandalism*

CRIMESTAT• FBI Uniform Crime Reports (1985). The data set consists of arrests by sex for selected crimes in United States in 1985. The variables are – *CRIME$, MALES, FEMALES.*

CROPS• Milliken and Johnson (1984). It is an agricultural data consists of yields in pounds (*YIELD*) of two varieties of wheat(*VARIETY*) grown in four different fertility regimes (*FERT*). To compare four fertilizers and two varieties of crops, four whole plots were grouped into two blocks (*BLOCK*). The two varieties were assigned randomly to the two whole plots in each group. Each whole plot is split into four subplots, and the four fertilizers are applied randomly to these.

DAYCREDM• Wilkinson, Blank, and Gruber (1996). This data set consists of three measures of a child's social competence, including a measure for behavior at dinner, a measure for behavior in dealing with strangers, and one involving social problem solving in a cognitive test. In addition, there is a categorical variable for the setting in which a child was raised, either by parents, by a babysitter, or by a daycare center. The variables are-

SETTING$	Daycare setting in which child is raised
SETTING	Coded setting
DINNER	Behavioral measure of skill during dinner
STRANGER	Measure of skill in dealing with a stranger
PROBLEM	Social problem-solving skills in a cognitive test

DELTIME• Montgomery, Peck, and Vining (2001). The data set deals with 25 delivery times of vending machines. The delivery time (*DELTIME*) of these machines is affected by the number of cases of product stocked (*CASES*) and the distance walked by the route driver (*DISTANCE*).

DESIGNDM• Devor, Chang, and Sutherland (1992). The data set consists of the results of an experiment designed to improve the performance of a fuel gauge.The variables are-

RUN	The case ID
SPRING	Dummy variable for the type of spring used
POINTER	Dummy variable for the type of pointer used

VENDOR	Dummy variable for the vendor used
ANGLE	Dummy variable for the type of angle bracket used
READING	The reading of the fuel gauge under the designed conditions

DEVMER• DEVEMER data file is derived from OURWORLD data file.

DIVORCE• Wilkinson, Blank, and Gruber (1996) and originally from Long (1971). This data set includes grounds for divorce in the United States in 1971.

DJONES• Brockwell and Davis(1991). The data set contains Dow-Jones Index of stocks on the New York Stock Exchange at closing on 251 trading days ending 26 August 1994. The data set contains the following variables:

DJSTOCK: Values of daily stocks of New York Stock Exchange

DJPRC: Percent relative price changes of the DJSTOCK series.

DOPTIMAL• Myers and Montgomery (2002). The data set is from an experiment based on a D-optimal design on adhesive bonding where the factors are amount of adhesive (*X1*) and cure temperature (*X2*). Here the response is the pull-off force (*Y*).

DOSE• These data are from a toxicity study for a drug designed to combat tumors. The data show the proportion of laboratory rats dying *(RESPONSE)* at each dose level *(DOSE)* of the drug. *LOGDOS*, dose in natural logarithm units.

ECLIPSE• These data are from the National Aeronautics and Space Administration web site and represent the longitude and latitude for the paths of eight future solar eclipses. Measurements occur at two minute intervals. The data are used courtesy of Fred Espenak, NASA/GSFC. The variables are-

MAPNUM	ID number
TIME$	Time in universal time at which eclipse will begin at the Latitude/Longitude for that case
MAXLAT	Northernmost latitude of total obstruction
MAXLON	Northernmost longitude of total obstruction
MINLAT	Southernmost latitude of total obstruction
MINLON	Southernmost longitude of total obstruction
LABLAT	Center latitude of total obstruction
LABLON	Center longitude of total obstruction
RATIO	Ratio of diameters of the Moon and the Sun
ALT	Altitude above horizon at the given Latitude/Longitude
AZIMUTH	Azimuth at which eclipse will occur
WIDTH	Width of the path of total obstruction

TOTALITY$	Time period of total obstruction at centerline
AUG_11_1999	Indicator for ellipse beginning on this date.
JUN_21_2001	Indicator for ellipse beginning on this date.
DEC_14_2001	Indicator for ellipse beginning on this date.
JUN_10_2002	Indicator for ellipse beginning on this date.
DEC_4_2002	Indicator for ellipse beginning on this date.
MAY_31_2003	Indicator for ellipse beginning on this date.
APR_8_2005	Indicator for ellipse beginning on this date.
OCT_3_2005	Indicator for ellipse beginning on this date.
LABEL$	Variable used for labeling eclipses on graphs

EDUCATN• This data set is a subset of the data set *SURVEY2*.

EGGS• Bliss (1967). An experiment was conducted to test the performance of laboratories and technicians to determine the fat content of dried eggs. A single can of dried eggs was stirred well. Samples were drawn and a pair of samples (claimed to be of two "types"), was sent to each of six commercial laboratories to be analyzed for fat content. Each laboratory assigned two technicians, who each analyzed both "types". The variables are-

FAT	Fat content as a percentage
LAB	Lab which ran the experiment
TECHNICIAN	Technician code
SAMPLE	Sample type used

EGYPTDM• Thomson and Randall-Maciver (1905). This data set consists of four measurements of male Egyptian skulls from five different time periods ranging from 4000 B.C. to 150 A.D. The four measurements of male Egyptian skulls are –

MB	Maximal breadth of skull
BH	Basibregmatic height of skull
BL	Basialveolar length of skull
NH	Nasal height of skull
YEAR	Time of measurement

EKMAN• Ekman (1954). These data are judged for similarities among 14 different spectral colors. The variable names are the colors' wavelengths *W584, W600, W610, W628, W651, W434,W445,W465,W472,W490,W504, W537,W55* and *W674*. The judgments are averaged across 31 subjects.

ELECSORT• This data set is obtained by sorting the data file *ELECTION,* by variable *NAME$*.

EMF• The data set consists of counts emfs of patients in urban and suburban areas affected by cancer or not. The variables are - *CANCER$, EMF$, RESIDENCE$, COUNT*.

ENERGY• SYSTAT created this file to demonstrate error bars. The variable *SE* determines the length of the error bar. *ENERGY$* is determined as low, medium, and high.

ENZYMDM• Greco et al. (1982). The data set consists of measurements of an enzymatic reaction measuring the effects on an inhibitor on the reaction velocity of an enzyme and substrate.

ENZYME• Greco et al. (1982). These data measure competitive inhibition for an enzyme inhibitor. *V* is the initial enzyme velocity, *S* is the concentration of the substrate, and *I* is the concentration of the inhibitor.

ESTIM• The data set consists of the estimated parameters for each sample of the data set *ENZYMDM*.

EURONEW• A subset of the *WORLD* data. These data include 27 European countries. The variable *LABLAT* is the latitude measurement of the capital, and *LABLON* is the longitude.

EX1• Wheaton, Muthén, Alwin, and Summers (1977). The data file is a covariance matrix of 6 manifest variables. The original data are attitude scales administered to 932 individuals in 1967 and 1971. The attitude scales measure anomia *(ANOMIA)*, powerlessness *(POWRLS)*, and alienation *(ALNTN)*. They also include a variable for socioeconomic index *(SEI)*, socioeconomic status *(SES)*, and years of schooling completed *(EDUCTN)*.

EX2• Duncan, Haller, and Portes (1971). The data is a correlation matrix of manifest variables. The original data measure peer influences on ambition. These data include the respondent's parental aspiration *(REPARASP)*, socioeconomic status *(RESOCIEC)*, intelligence *(REINTGCE)*, occupational aspiration *(REOCCASP)*, and educational aspiration *(REEDASP)*. These data also include the respondent's best friend's intelligence *(BFINTGCE)*, socioeconomic status *(BFSOCIEC)*, parental aspiration *(BFPARASP)*, occupational aspiration *(BFOCCASP)*, and ambition *(BFAMBITN)*.

EX3• Mels and Koorts (1989). These data are taken from a job satisfaction survey of 213 nurses. There are 10 manifest variables that serve as indicators of four latent variables: job security *(JOBSEC)*, attitude toward training *(TRAING)*, opportunities for promotion *(PROMOT)*, and relations with superiors *(RELSUP)*.

EX4A and **EX4B•** Lawley and Maxwell (1971). These data comprise a correlation matrix of nine ability tests administered to 72 children.

EXER• The data consist of people who were randomly assigned to two different diets (*DIET*) low-fat and not low-fat and three different types of exercise (*EXERTYPE*) at rest, walking leisurely and running. A baseline pulse measurement (*PULSE*) was obtained at time = 0 for every individual in the study. However, subsequent pulse measurements were taken at less regular time intervals. The second pulse measurements were taken at approximately 2 minutes (time = 120 seconds); the third pulse measurement was obtained at approximately 5 minutes (time = 300 seconds); and the fourth and final pulse measurement was obtained at approximately 10 minutes (time = 600 seconds).

EXPORTS • Hand, Daly, Lunn, McConway, and Ostrowski (1996). This data set consists of the value (in millions of £) of British exports (*EXPORTS*) during the years 1820 to 1850 (*YEAR*) .

FLEA• Lubischew (1962). The data set consists of measurements on the following four variables on two species (*SPECIES*) of flea beetles:

X1 Distance of the transverse groove to the posterior border of the paradox (in microns)

X2 Length of the elytra (in mm)

X3 Length of the second antennal point (in microns)

X4 Length of the third antennal joint. (in microns)

FLEABEETLE• Hand et al. (1996). Data were collected on the genus of flea beetle Chaetocnema, which contains three species (*SPECIES$*): concinna (Con), heikertingeri (Hei), and heptapotamica (Hep). Measurements were made on the width and angle of the aedeagus of 74 beetles. The goal of the original study was to form a classification rule to distinguish the three species. The data set consists of only measurements of angle of aedeagus of beetles. The variables are - *ANGLE, SPECIES$*.

FOOD• These data were gathered from food labels at a grocery store. The variables are-

BRAND$	Shortened name for brand
FOOD$	Type of dinner: chicken, pasta, or beef
CALORIES	Calories per serving
FAT	Grams of fat
PROTEIN	Grams of protein
VITAMINA, CALCIUM, IRON	Percentage of daily value of vitamin A, calcium, and iron
COST	Price per dinner
DIET$	*Yes* if low in calories; *no* if standard

FORBES• Bringham(1980). The data are various characteristics of financial performance in chemical companies reported by 30 largest companies. The variables are:

PE_RATIO: Price-to-earning ratio, which is the price of one share of common stock divided by the earnings per share for the past year. This ratio shows the dollar amount investors are willing to pay for the stock per dollar of current earnings of the company.

ROR5: Percent rate of return on total capital (invested plus debt) averaged over the past 5 years.

DE_RATIO: Bept-to-equity (invested capital) ratio for the past year. This ratio indicates the extents to which management is using borrowed funds to operate the company.

SALESGR5: percent annual compound growth rate of sales, computed from the most recent five years compared with the previous five years.

EPS5: percent annual compound growth in earning per share, computed from the most recent five years compared with the previous five years.

NPM1: Percent net profit margin, which is the net profits divided by the sales for the past year, expressed as a percentage.

PAYOUTR1: Annual dividend divided by the latest 12-month earnings per share. This value represents the proportion of earnings paid out to shareholders rather than retained to operate and expand the company.

FOREARM1• Pearson and Lee (1903). The data set consists of *ARMLENGH,* that is length of forearm (in inches) of 140 men.

FOSSILS• The data give the incidence of fossil specimens of various flora found at various elevations of a site in British Columbia. The variables are – *HEIGHT, CHARA, NITALLA, JUNCUS, RUMEX.*

FRACTION• These data are from a half of a 2^4 factorial design. Each cell contains two observations on a *Y* variable

FRTFLYDM• Carey, Liedo, Orozco, and Vaupel (1992). This data set contains information on mortality rates for Mediterranean fruit flies over 172 days, after which all flies were dead. Experimenters recorded the number of flies dying each day (*DAY*) and divided this by the number alive (*LIVING*) at the beginning of the day to measure mortality rate (*MORTRATE*) for each day.

GAUGE1• Smith (2001). The data set consists of repeated measurements (*READING*) of a characteristic of ten items (*ITEM),* each by three persons (*PERSON)*.

GAUGE2• Montgomery and Runger (1993). Three operators measure a quality characteristic on twenty units twice each.

GDP• The data set consists of CSO's quarterly estimates of growth rates of GDP for 1996-1997 to 2004-2005 for the following eight sectors. The variables are – *YEAR$*, *AGRICULTURE*, *MINING*, *MANUFACTURE*, *ELECTRICITY*, *CONSTRUCTION*, *TRADE*, *FINANCING*, *COMMUNITY*, *OVERALL-GDP*.

GDWTRDM• Nichols, Kane, Browning, and Cagle (1976). The U.S. Department of Energy collected samples of groundwater in West Texas as part of a project to estimate U.S. uranium reserves. Samples were taken from five different locations called producing horizons, and then measured for various chemical components. In addition, the latitude and longitude for each sample location was recorded. The variables are-

SAMPLE	The ID of the groundwater sample
LATITUDE	Latitude at which the sample was taken
LONGTUDE	Longitude at which the sample was taken
HORIZON$	Initials of producing horizon
HORIZON	ID of producing horizon
URANIUM	Uranium level in groundwater
ARSENIC	Arsenic level in groundwater
BORON	Boron level in groundwater
BARIUM	Barium level in groundwater
MOLYBDEN	Molybdenum level in groundwater
SELENIUM	Selenium level in groundwater
VANADIUM	Vanadium level in groundwater
SULFATE	Sulfate level in groundwater
TOT_ALK	Alkalinity of groundwater
BICARBON	Bicarbonate level in groundwater
CONDUCT	Conductivity of groundwater
PH	pH of groundwater
URANLOG	Log of uranium level in groundwater
MOLYLOG	Log of molybdenum level in groundwater

GRADES• The variables in this data set are marks in four quiz (*QUIZ1, QUIZ2, QUIZ3, QUIZ4)* of six students (*NAME$)* and their marks in *MIDTERM* and *FINAL* exams**.**

GROWTH• Each case in this file represents a group of plants receiving the same dose *(DOSE)* of a growth hormone. *GROWTH* is the mean growth measure for each group, and *SE* is the standard error of the mean.

HARDDIA• Taguchi (1989). The data set consists of measurements on 20 units of two characteristics of a product: Brinell hardness number (*BHN*) and circular diameter (*DIAMETER*).

HEAD• Frets (1921). The data consists of measurements on the following characteristics of two sons of 25 families. The variables are-

HLEN1	Head length of the first son
HBREAD1	Head breadth of the first son

HLEN2	Head length of the second son
HBREAD2	Head breadth of the second son

HEADDIM• Flury and Riedwyl (1988).These data are measurements of two hundred 20 year old male Swiss army personnel on the following characteristics:

MFB	Minimal frontal breadth
BAM	Breadth of angulus mandibulae
TFH	True facial height
LGAN	Length from glabella to apex nasi
LTN	Length from tragion to nasion
LTG	Length from tragion to gnathion

HEART• DASL (2005). An experiment was conducted by students at The Ohio State University in the fall of 1993 to explore the relationship between a person's heart rate and the frequency at which that person stepped up and down on steps of various heights. The response variable, heart rate, was measured in beats per minute. There were two different step heights: 5.75 inches (coded as 0), and 11.5 inches (coded as 1). There were three rates of stepping: 14 steps/min. (coded as 0), 21 steps/min. (coded as 1), and 28 steps/min. (coded as 2). This resulted in six possible height/frequency combinations. Each subject performed the activity for three minutes. Subjects were kept on pace by the beat of an electric metronome. One experimenter counted the subject's pulse for 20 seconds before and after each trial. The subject always rested between trials until her or his heart rate returned to close to the beginning rate. Another experimenter kept track of the time spent stepping. Each subject was always measured and timed by the same pair of experimenters to reduce variability in the experiment. Each pair of experimenters was treated as a block. The variables are -

ORDER	The overall performance order of the trial
BLOCK	The subject and experimenters' block number
HEIGHT	0 if step at the low (5.75") height, 1 if at the high (11.5") height
FREQUENCY	The rate of stepping, 0 if slow (14 steps/min), 1 if medium (21 steps/min), 2 if high (28 steps/min)
RESTHR	The resting heart rate of the subject before a trial, in beats per minute
HR	The final heart rate of the subject after a trial, in beats per minute

HELM• Helm (1959), reprinted by Borg and Lingoes (1987). These data contain highly accurate estimates of “distance” between color pairs by one experimental subject (CB). Variables include *A, C, E, G, I, K, M, O, Q, a*nd *S.*

HILLRACE• Atkinson (1986). The data set gives the record-winning times (*TIME*) for 35 hill races (*RACES$*) in Scotland. The distance (*DISTANCE*) travelled and the height climbed (*CLIMB*) in each race are also given. The variables are-

RACE$	Name of the Race
DISTANCE	Distance covered in miles
CLIMB	Elevation climbed during race in feet
TIME	Record time for race in minutes

HILO• These are hypothetical price data for a stock. *HIGH* is the highest price for that month *(MONTH* and *MONTH$)*, *LOW* is the low price, and *CLOSE* is the closing price at the end of the month.

HISTAMINE• Morrison and Zeppa (1963). It consists of data having a multivariate layout. In this study, mongrel dogs were divided into four groups of four. The groups received different drug treatments. The dependent variable, blood histamine in mg/ml, was measured at four times *HISTAMINE1, HISTAMINE2, HISTAMINE3* and *HISTAMINE4* after administration of the drug. The data are incomplete, since one of the dogs is missing in the last measurement.

HOSLEM• Hosmer and Lemeshow (2000). The variables are-

ID	Identification Code
LOW	Low infant birth weight
AGE	Mother's age
LWT	Mother's weight during last menstrual period
RACE	1= white, 2= black, 3= other
SMOKE	Smoking status during pregnancy
PTL	History of premature labor
HT	Hypertension
UI	Uterine irritability
FTV	Number of physician visits during first trimester
BWT	Birth weight

HOSLEMM• Hosmer and Lemeshow (2000). It already exists in SYSTAT as *HOSLEM*. Four new variables are added to it, which are fictitious: The variables are-

SETSIZE	The number of subjects in each strata (which is AGE for this analysis)
GROUP	Identity number of strata.
REC	Case number.
DEPVAR	The relative position of the case in a given matched set.

HW• It is a hypothetical data of height and weight of a group of people according to gender.

ILEA• Goldstein (1987). It is a subset of data from the Inner London Education Authority (*ILEA)*. The data consists of information about 2069 students within 96 schools. The variables are-

ACH	Measures of achievement
PFSM	The percent of students within each school who are eligible to participate in a free meal program
VRA	A verbal reasoning ability level from 1 to 3

INCOME• The data here were collected from a class of students. There are two variables. *SCORES1* represents the percent score of students in a statistics test and *INCOME* the monthly family income in thousand dollars.

INSTRDM• Huitema. (1980). This data set consists of measures of achievement on a biology exam for two groups of students. One group was simply told to study everything from a biology text in general, and the other was given terms and concepts that they were expected to master. An additional covariate, the student's aptitude, is also included in the data set. The variables are-

STUDENT	Student ID
INSTRUCT$	Type of instruction given
INSTRUCT	Coded variable for *INSTRUCT$*
APTITUDE	Student's underlying ability to learn
ACHIEVE	Student's score on the exam

IRIS• Anderson (1935). These data measure sepal length *(SEPALLEN)*, sepal width *(SEPALWID)*, petal length *(PETALLEN)*, and petal width *(PETALWID)* in centimeters for three species *(SPECIES)* of irises (1=Setosa, 2=Versicolor, and 3=Virginica).

JOHN• John (1971). These data are from an incomplete block design with three treatment factors *(A, B,* and *C)*, a blocking variable with eight levels *(BLOCK)*, and the dependent variable (*Y*).

JUDGEHILL• Judge,et al.(1988). This data set is obtained on appending data for the two models. It contains two indicator variables X11 and X21 representing the cases obtained from the first and second model respectively. X12 and X22 represent the market values of a certain product of two different companies with capital stocks X13 and X23 respectively. The dependent variable Y represents the investment figures for the two companies. The data set is fictitious.

JUICE• Montgomery (2005). The number of defective orange juice cans *(DEFECTS)* found in each of 24 samples *(SAMPLE)* of 50 juice cans. Data are collected on each of three shifts *(TIME$)* with eight samples taken for each shift *(SHIFT$)*. *SIZE* is also a variable.

JUICE1• Montgomery (2005). The following fictitious variable has been added to *JUICE*.

DEFECTS1	The number of defective orange juice cans found in each of 24 samples (*SAMPLE*) of 50 juice cans

KENTON• Neter, Kutner, Nachtsheim, and Wasserman (1996). These data comprise of unit sales of a product *(SALES)* under different types of package designs *(PACKAGE)*. Each case represents a different store.

KOOIJMAN• Kooijman (1979), reprinted in Upton and Fingleton (1990). The data consist of the locations of beadlet anemones (Actinia equina) on the surface of a boulder at Quiberon Island, off the Brittany coast, in May 1976.

KUEHL• Kuehl (2000). The original data source is Dr. S. Denise, Department of Animal Sciences, University of Arizona. A genetic study with beef animals consisted of several sires each mated to a separate group of dams. The matings that resulted in male progeny calves were used for an inheritance study of birth weights. The birth weights of eight male calves in each of five sire groups are given. The variables are - *SIRE, BIRTHW, PROGENY*, and *GR*.

LAB• Jackson (1991). The data set consists of four bivariate vector observations per laboratory Samples were tested in three different laboratories (*LAB*) using two different methods (*METHOD1*, *METHOD2*) and each *LAB* received four samples.

LABOR• U.S. Bureau of Labor Statistics. These data show output productivity per labor hour in 1977 U.S. dollars for a 25-year period *(YEAR)*. Other variables are *US, CANADA, JAPAN,* and *GERMANY* and *ENGLAND*.

LATIN• Neter, Kutner, Nachtsheim and Wasserman (1996). These data are from a Latin square design in which the response *(RESPONSE)* in each square *(SQUARE)* is from one of five days a week *(DAY)* for five weeks *(WEEK)*.

LAW• Efron and Tibshirani (1993). The law school data. A random sample of size 15 was taken from the universe of 82 USA law schools. Two variables are- average score on a national law test (*LSAT*) and average undergraduate grade-point average (*GPA*).

LEAD• Ott and Longnecker (2001). The data set consists of lead concentrations (mg/kg dry weight) of 37 stations in Kenya, obtained from a geo-chemical and oceanographic survey of inshore waters of Mombasa, Kenya.

LEARN• Gilfoil (1982). These data demonstrate a quadratic function with a ceiling. They are from a study showing that inexperienced computer users prefer dialog menu interfaces while experienced users prefer command-based interfaces. *SESSION* is the session number, and *TASKS* is the number of command-based (as opposed to dialog-based) tasks initiated by the user during that session.

LEISURE• Clausen (1998). These data show a cross-classification between different leisure activities and different occupational status. The following is a list of the different activities and occupational status. The SYSTAT names are within parentheses.

Activities	**Occupational Status**
Sports Events (*Sports*)	Manual (*MANUAL*)
Cinema (*Cinema*)	Low Non Manual (*LOWNM*)
Dance/Disco (*Dance*)	High Non Manual (*HIGHNM*)
Cafe/Restaurant (*Cafe*)	Farmer (*FRAMER*)
Theatre (*Theatre*)	Student (*STUDENT*)
Art Exhibition (*Art*)	Retired (*RETIRED*)
Library (*Library*)	
Church Service (*Church*)	
Classical Music (*Classical*)	
Pop (*Pop*)	

LIFE• The data are lifetimes (*LIFE*) of 20 units of a certain equipment.

LONGLEY• Longley (1967). These data are economic data selected by Longley to illustrate computational shortcomings of statistical software. The variables are - *DEFLATOR, GNP, UNEMPLOY, ARMFORCE, POPULATN, TIME,* and *TOTAL*.

LUNGDIS• Hand, Daly, Lunn, McConway, and Ostrowski (1996). This data set consists of monthly (*MONTH$*) deaths (*DEATHS*) from lung diseases in the UK during the years (*YEAR*) 1974 to 1979.

MACHINE• These data are in the file *MACHINE* and represent the numbers *(N)* of conforming (*RESULT* is 1) and nonconforming (*RESULT* is 0) units produced by each of five machines.

MACHINE1• Milliken and Johnson (1992). An experiment was conducted by a company to compare the performances of three different brands of machines when operated by the company's own personnel. Six employees were selected at random and each of them had to operate each machine three different times. The data set consists of overall scores that take into account both the quantity and quality of the output. The variables are - *SCORE, MACHINE, OPERATOR* and *TIME*.

MACHINE2• Milliken and Johnson (1992). It is an unbalanced data set where two machines were operated by six randomly selected operators. Each operator was allowed to operate each machine at most three times.

MACK• Breslow and Day (1980). The data deals with the cases of eudiometrical cancer in a retirement community near Los Angeles. The data are reproduced in their Appendix III. The variables are-.

CANCER	
AGE	
GALL	Gallbladder disease
HYP	Hypertension
OBESE	Obesity
EST	Estrogen
DOS	Dose
DUR	Duration of conjugated estrogen exposure
NON	Other drugs

The data are organized by sets, with the case coming first, followed by four controls, and so on, for a total of 315 observations (63 * (4 + 1)).

MANOVA• Morrison (1990). These data are from a hypothetical experiment measuring weight loss in rats. Each rat was assigned randomly to one of three drugs *(DRUG),* with weight loss measured in grams for the first and second weeks of the experiment *(WEEK(1)* and *WEEK(2))*. *SEX* was another factor.

MELNMADM• Wilkinson and Engelman (1996). This data set contains reports on melanoma patients. The variables are-

TIME	The survival time for melanoma patients in days
CENSOR	The censoring variable
WEIGHT	The weight variable
ULCER	Presence or absence of ulcers
DEPTH	Depth of ulceration
NODES	Number of lymph nodes that are affected
SEX$	The sex of the patient
SEX	The stratification variable coded for analysis

METOX• Fellner (1986). The data set is about metallic oxide analysis where two types of metallic oxides, eighteen lots from the first type, and thirteen from the second were used. Two samples were drawn from each lot. A pair of chemists was randomly selected for each sample. The variables are - *TYPE, SAMPLE, CHEMIST* and *Y*.

MILK• Brownlee (1960). The data set pertains to bacteriological testing of milk. Twelve milk samples (*SAMPLE*) were tested in all six combinations of two types of bottles (*BOTTLE$*) and three types of tubes (*TUBE$*). Ten tests were run on each combination and the response was the number of positive tests in each set of ten (*Y*).

MINIWRLD• This data file is a subset of *OURWORLD*.

MINTEMP• Barnett and Lewis (1967). The data set consists of a variable *TEMP* that is annual minimum temperature (F) of Plymouth (in Britain) for 49 years.

MISSLES• Jackson (1991). These data are a covariance matrix of measures performed on 40 Nike rockets. The variables are: *INTEGRA1, PLANMTR1, INTEGRA2,* and *PLANMTR2*.

MJ006• Milliken and Johnson (1984). This data set came from an experiment that was conducted to determine how six different kinds of work tasks (*TASK*) affect a worker's pulse rate. In this experiment, 78 male workers were assigned at random to six different groups so that there were 13 workers in each group. Each group of workers was trained to perform its assigned task. On a selected day after training, the pulse rates (*PULSE*) of the workers were measured after the workers had performed their assigned tasks for one hour. Unfortunately some individuals withdrew from the experiment during the training process so that some groups contained fewer than 13 individuals. The recorded data represent the number of pulsations in 20 seconds.

MJ020• Milliken and Johnson (1984).The data set is from a paired association learning task experiment performed on subjects under the influence of two drugs. Group1 is a control (no drug), Group2 was given drug1, Group3 was given drug2 and Group4 was given both drugs. The variables are - *LEARNING* and *GROUP*.

MJ129• Milliken and Johnson (1984). The data set is from a small two-way treatment structure experiment conducted in a completely randomized design structure.

MJ166• Milliken and Johnson (1984). A bakery scientist wanted to study the effects of combining three different fats (*FAT*) with each of three different surfactants (*SURF*) on the specific volume of bread loaves (*SPVOL*) baked from doughs mixed from each of the nine treatment combinations. Four flours (*FLOUR*) of the same type but from different sources were used as blocking factors. That is, loaves were made using all nine treatment combinations for each of the four flours.

MJ173• Milliken and Johnson (1984). This is a hypothetical data set from a two-way treatment structure in a completely randomized design with treatment *T* and treatment *B* each having three levels.

MJ202• Milliken and Johnson (1984). These data are from a home economics survey experiment. *DIFF* is the change in test scores between pre-test and post-test on a nutritional knowledge questionnaire. *GROUP* classifies whether or not a subject received food stamps. *AGE* designates four age groups, and *RACE$* designates whites, blacks, and Hispanics.

MJ332• Milliken and Johnson (1984). An experiment involved 3 drugs to study the effect of each drug on heart rate of eight persons in four time periods. The variables are- *PERSON, HR, DRUG, TIME.*

MJ338• Milliken and Johnson (1984). An engineer had three environments in which to test three types of clothing. Four people (two males and two females) were put into an environmental chamber (each one was assigned one of the three environments). One male and one female wore clothing type 1, and the other male and female wore clothing type 2. The comfort score of each person was recorded at the end of one hour (*SCORE(1)*), two hours (*SCORE(2)*), and three hours (*SCORE(3)*).

MJ379• Milliken and Johnson (1984). An experimenter wanted to study the effects of three different herbicides (*HERB*) and four fertilizers (*FERT*) on the growth rate of corn. Fifteen plots of land (*PLOT*) were available for the experiment, and 5 plots were randomly assigned to each of the three herbicides. Each of the 15 plots were further divided into 4 subplots, and a different fertilizer treatment was randomly assigned to each. At the beginning of the third week, 10 plants were selected at random from each subplot And the height of each plant was measured. The average of the 10 heights (*HEIGHT*) was recorded as the measurement from the subplot. Unfortunately, before any measurement could be taken, 3 of the 15 whole plots were destroyed by excessive rainfall. Herbicide 1 had been assigned to two of those subplots and herbicide 3 to the third

MJ385• Milliken and Johnson (1984). These data form a small part of an experiment conducted to determine the effects of a drug on the scores obtained by depressed patients on a test to measure depression. Two patients were in the placebo group, and three in the drug group. The variables are- *SCORE, WEEK, PATIENT, TREAT$.*

MOTHERS• Morrison (2004). These data are hypothetical profiles on three scales of mothers *(SCALE(1)* to *SCALE(3))* in each of four socioeconomic classes *(CLASS)*. Other variables are *A$, B$, C$, A, B,* and *C*.

MRCURYDM• Lange et al. (1993). The data set consists of measurements of large-mouth bass in 53 different Florida lakes to examine the factors that influence the level of mercury contamination. Water samples were collected from which the pH level, the amount of chlorophyll, calcium, and alkalinity were measured. A sample of fish was taken from each lake, for which the age of each fish and mercury concentration in the muscle tissue was measured (older fish tend to have higher concentrations). To make a fair comparison of the fish in different lakes, the investigators used a regression estimate of the expected mercury concentration in a three-year-old fish as the standardized value for each lake. Finally, in 10 of the 53 lakes, the age of the individual fish could not be determined and the average mercury concentration of the sampled fish was used. The variables are-.

ID	Lake ID
LAKE$	Lake name
ALKLNTY	Measured alkalinity of the lake (mg/L as Calcium Carbonate)
PH	Measured PH of the lake
CALCIUM	Measured Calcium of the lake (mg/l)
CHLORO	Measured Chlorophyll of the lake (mg/l)
AVGMERC	Average mercury concentration (parts per million) in the tissue of the fish sampled from the lake
SAMPLES	Number of fish sampled in the lake
MIN	Minimum mercury concentration in sampled fish from lake
MAX	Maximum mercury concentration in sampled fish from lake
STDMERC	Regression estimate of the mercury concentration in a 3 year old fish from the lake
AGEDATA	Indicator of the availability of age data on fish sampled
LNCHLORO	Log of CHLORO

MULTIRESP• Myers & Montgomery (2002). This data set contains observations on three responses at different level combinations of two factors, time (*TIME*) and temperature (*TEMP*) of a chemical process. The three responses are yield (*YIELD*), viscosity (*VISCOSITY*) and the number-average molecular weight (*MOLWEIGHT*). The data set also contains coded versions of these variables. *X1* describes the *TIME* variables after being used coded, and *X2* describes *TEMP* after being coded.

NAFTA• Two months before the North Atlantic Federal Trade Agreement approval and before the televised debate between Vice President Al Gore and businessman Ross Perot, political pollsters queried a sample of 350 people, asking "Are you For, Unsure, or Against NAFTA?" After the debate, the pollsters contacted the same people and asked the question a second time. Variables include *BEFORE$, AFTER$,* and *COUNT*.

NEWARK• Collected by the U.S. Government and cited in Chambers, et al. (1983). These data are 64 average monthly temperatures *(TEMP)* in Newark, New Jersey, beginning with January, 1964.

NFL• Johnson (1999). The data set is obtained from the NFL for the 1999-2000 season for those players with at least 1,500 passing attempts. It is NFL Passer Rating Data. *RATING* is based on performance standards established for completion percentage, average gain, touchdown percentage, and interception percentage. The variables are:

NAME$	Last name and first name of Quaterback
ATTEMPTS	Passing attempts
COMPLETIONS	Percentage of completions per attempt

YARDS	Average yards gained per attempt
TDS	Percentage of touchdown passes per attempt
INTS	Percentage of interceptions per attempt
RATING	NFL Ratings (rounded to the nearest 0.1)

NLS• The data used here have been extracted from the National Longitudinal Survey of Young Men (1979), containing information on 200 individuals on school enrollment.

NOTENR	School Enrollment Status (1 if not enrolled, 0 otherwise)
BLACK	A race dummy (0 for white)
SOUTH	A region dummy (0 for non-South)
EDUC	Highest completed grade
AGE	Age
FED	Father's education
MED	Mother's education
CULTURE	An index of reading material available in the home (1 for least, 3 for most)
NSIBS	Number of siblings
LW	Log10 of wage
IQ	An IQ measure
FOMY	Mean income of persons in father's occupation in 1960

OPERA• The following data are from an editorial in The New York Times (December 3, 1987). They represent the duration (*HOURS*) of various plays, films, and operas (*TITLE$*).

ORDEREDOUTPUT• Hollander and Wolfe (1999). 18 male workers are divided into three groups as receiving no information about output (Control), receiving a rough estimate (Group B), and receiving accurate information (Group C).

OURWORLD• Variables recorded for each case (country) include:

COUNTRY$	Names of the 95 countries used in this data file
URBAN	Percentage of population living in urban areas
LIFEEXPF, LIFEEXPM	Years of life expectancy for females and males
GDP$	Group variable with codes "Developed" and "Emerging"
GDP_CAP	Gross domestic product per capita in U.S. dollars
BABYMORT, BABYMT82	*BABYMORT* = infant mortality rate for 1990; *BABYMT82* = infant mortality rate in 1982
BIRTH_RT	Number of births per 1000 people in 1990
DEATH_RT	Number of deaths per 1000 people in 1990
BIRTH_82, DEATH_82	Number of births and deaths per 1000 people in 1982

B_TO_D	Birth to death ratio in 1990
HEALTH, EDUC, MIL, HEALTH84, EDUC_84 and MIL_84	Expenditures (in U.S. dollars) per person for health, education, and the military in 1990 and in 1984
POP_1983, POP_1986, POP_1990, POP_2020	Populations in millions for the years 1983, 1986, and 1990; *POP_2020* is the population projected by the United Nations for 2020
GNP_82, GNP_86	Gross national product in 1982 and 1986
RELIGION$	Expenditures grouped by the religion or personal philosophy of those who govern the country
GOV$	Type of government
LEADER$	Religion of the leaders of countries
LITERACY	Percentage of the population that can read
GROUP$	Europe, Islamic, or the New World
URBAN$	Rural or urban
MCDONALD	Number of McDonald's restaurants per country
LAT, LON	Latitude and longitude measurements of the center of the country
B_TO_D82	Birth to death ratio in 1982
LOG_GDP	Log of gross domestic product per capita
LIFE_EXP	Years of life expectancy

PAINTS• Milliken and Johnson (1992). The dataset consists of four different paints, Yellow 1, Yellow 2, White 1 and White2 that are manufactured by two different companies, where the 1 and 2 refer to the company. Each of the paint is applied on three different paving surfaces: Asphalt1, Asphalt1 and Concrete. The response is the life time measured in weeks. In original data only the cell means and error sum of squares have been reported so the following data set has been generated artificially to have the same cell means and error sum of squares as the original data. The variables are - *Y, PAINT$, PAVE$.*

PAROLE• Maltz (1984). These data record the number of Illinois parolees *(COUNT)* who failed conditions of their parole after a certain number of months *(MONTH)*. An additional 149 parolees failed after 22 months, but these are not used.

PATMISS• Hocking (2003). In an experiment a pharmaceutical company was trying to test a new medicine. Three clinics were selected at random from a large number of clinics. The drug was administered to ten randomly selected patients. However, some of the measurements from some of the clinics have not been reported. The variables are - *CLINIC* and *Y.*

PATTERN• Laner, Morris and Oldfield (1957). In a psychological experiment of visual perception, there were required 1555520 squares to color (either black with probability 0.29 or white with probability 0.71). From this a total of 1000 non-overlapping samples each containing 16 of small squares were randomly selected, and the number of black squares were counted in each case. The data set consists of the frequency distribution of this count.

PATTISON• Clarke (1987). In his 1987 JASA article, C. P. Y. Clarke discusses the data taken from an unpublished thesis by N. B. Pattinson for 13 grass samples collected in a pasture. Pattinson recorded the weeks since grazing began in the pasture *(TIME)* and the weight of grass cut from 10 randomly sited quadrants, then fit the Mitcherlitz equation:

$$\text{GRASS} = \theta_1 + \theta_2 e^{-\theta_3 \text{TIME}}$$

PDLEX1• Gujarati (1995). The data set relates to the *SALES* and *INVENTORY* of a product in 20 days.

PDLEX2• Gujarati (2003). The data set relates to the *SALES* and *INVENTORY* of a product for the United States for the period 1954-1999.

PDLEX3• Gujarati (2003). The data set relates to income-money supply model of USA for the period 1970-1999. The variables are as follows:

GDP	Gross domestic product ($, billions, seasonally adjusted)
M2	Money supply ($, billion, seasonally adjusted)
GDPI	Gross private domestic investment ($, billion, seasonally adjusted)
FEDEXP	Federal government expenditure ($, billion, seasonally adjusted)
TB6	Six-month treasury bill rate (%)

PESTICIDE• Milliken and Johnson (1992). Four chemical companies produce certain pesticides. Company A produces three such products, companies B and C produces two such products each, and company D produces four such products. No company produces a product exactly like that of another. The treatment structure is a two-way with *COMPANY$* as one factor and *PESTICIDE* as the other. To compare these we use 33 glass containers that are randomly grouped into eleven groups of three. The pesticides are assigned randomly to the groups. The assigned pesticide is applied to the inside of each box in its group. A box with 400 mosquitoes and soil with bluegrass is put inside each container and the number of live mosquitoes in each box was counted after 4 hours (*Y*).

PESTRESIDUE• Kuehl (2000). A comparison was made among two standard pesticide methods to compare and test the amount of residue left on cotton plant leaves is the same for the two methods (*METHOD*). To test these six batches (*BATCH*) of plants were sampled from the field. Two plants were used in the experiment from each batch. Thus, there were twelve plants in the experiment (*SAMPLE*). The plants inside each batch were from the same field plot. Method one was applied to three randomly selected batches, and the remaining three batches were given method two. The amounts of residue on the leaves were measured after a specified amount of time for each of the twelve plants (*Y*).

PHONECAL• Rousseeuw and Leroy (1987). The data set, which comes from the Belgian Statistical survey, describes the number of international phone calls from Belgium in years 1950-1973. The variables are:

X Years

Y Number of phone calls

PHOSPHOR• Hocking (1985). The data set is about the concentration of phosphorus in the wash water. The aim of the investigation is to determine how the concentration varies with the types of detergent and washing machines. The experiment was carried out with four different types of detergents, three different types of machines, and seven laundromats. The laundromats had different numbers of machines, but each laundromat had only machines of a single type. Thus, laundromats are nested inside machine types. The machines within each laundromat were divided into four groups of roughly equal sizes, and the four types of detergent were allocated to them. The response is the average amount of phosphorus in grams per liter from daily one-hour samples over a seven day period. The variables are - *Y, N, MACHINE, LAUNDRY, DETERG*

PHYSICAL• Crowder and Hand (1990). The data set shows three groups of diabetic patients and one control group (*GROUP*). The response variable is observed at 12 time points and the corresponding variables are *X1, X2 & Y1* through *Y10*, respectively.

PISTON• Taguchi,El Sayed, Hslang (1989). This data set consists of diameter differences (*DIA*) between the cylinder and the piston of a six-cylinder engine. The sample was selected from a month's (*MONTHS*) production of an automobile manufacture unit.

PLANKS• Netmaster Statistics Courses. After drying beech wood the humidity level at any given point inside a plank typically depends on the depth of the point. To study the relation between the humidity levels (measured as a percentage) the depth , and twenty different randomly selected beech planks were measured for humidity level at five depths and three widths. The variables are - *PLANK, WIDTH, DEPTH* and *HUMIDITY*.

PLANTS• SYSTAT created this file to demonstrate regression with ecological or grouped data. The variables are: *CO2, SPECIES,* and *COUNT*.

PLOTS• The split plot design is closely related to the nested design. In the split plot, however, plots are often considered a random factor. Thus, different error terms are constructed to test different effects. Here is an example involving two treatments: *A* (between plots) and *B* (within plots). The numbers in the cells are *YIELD* of the crop within plots. These data also use *PLOT, PLOT(1),* and *PLOT(2)* as variables.

POLAR• These data show the highest frequency *(FREQ)* (in 1000's of cycles per second) perceived by a subject listening to a constant amplitude sine wave generator oriented at various angles relative to the subject *(ANGLE)*.

POLYNOM. The following variables were created in SYSTAT using the equations

$X = u + i - 10$

$Y = 2 + 3*X + 4*X^2 + 5*X^3 + 500*z$

where *u* is a uniform random variable, *i* is an index running from 1 to 20, and *z* is a standard normal random variable. The variable *ESTIMATE* was estimated from a cubic regression model. Finally, the variables *UPPER* and *LOWER* were computed. *UPPER* corresponds to two standard errors above the estimated value and *LOWER* corresponds to two standard errors below.

POWER• Ott and Longnecker (2001). The data set consists of deviations from target power (*POWER*) using monomers from three different suppliers (*SUPPLIER*) with a total number of 27 cases.

PRENTICE• Prentice (1973). This is a survival time data of 137 advanced lung cancer patients. The data file contains following variables:

TRTMNT	Two treatments; 1= standard, 2= test
SURVTIME	Survival time measured from the start of the treatment for each patient
STATUS	Censoring status where 1= censored, 0 = failed
TMRTYPE	Types of tumor: 1= squamous, 2= small, 3= adeno, and 4 =large
KSCORE	Karnofsky score is a performance status assigned to the patient at the time of diagnosis
AGE	Age of the patient
MONTHS	Diagnostic period
THERAPY	Prior therapy status where 0 = no prior therapy and 10 = with prior therapy

PROCESS• Breyfogle (2003). The data set consists of the number of units checked and the number of defects found in 10 operations step in a production process.

PULPFIBER• Lee (1992). The data set contains 62 measurements on the properties of pulp fibers and the paper made from them.

Four types of pulp fiber characteristics are:

X1	Arithmetic fiber length
X2	Long fiber fraction
X3	Fine fraction

X4 Zero span tensile

The four paper properties are:

Y1 Breaking length
Y2 Elastic modulus
Y3 Stress at failure
Y4 Burst strength

PUMPFAILURES• Gaver and O'Muircheartaigh (1987). The data set consists of the number of failures (*F*) and times of observation (*T*) for 10 pump systems at a nuclear power plant.

PUNCH• Cornell (1985). These data measure the effects of various mixtures of watermelon *(WATERMELN)*, pineapple *(PINEAPPL)*, and orange juice *(ORANGE)* on taste ratings by judges *(TASTE)* of a fruit punch.

QUAD• Cook and Weisberg (1990). The data set is from a function, which reaches its maximum at –b/2c; however, for the data given by Cook and Weisberg, this maximum is close to the smallest *X*. In other words, little of the response curve is found to the left of the maximum.

QUAKES• The Open University (1981). The data set consists of *TIME* in days between successive serious earthquakes worldwide.

QUESTABILITY• Gibbons and Chakraborti (2003). In raising small children's ability, an important factor is to develop their ability to ask questions in groups. A study of group size and number of questions asked by preprimary children in a classroom atmosphere was conducted with a familiar person after dividing the 46 children into 4 groups: Group1 (24 children), Group2 (12 children), Group3 (6 children), and Group4 (4 children). The total number of questions asked (*QUESTIONS*) by all children of each group is recorded for 30 minutes on each of eight different days (*BLOCK*).

RAINFALL• Lee (1989*).* This is a data set of December rainfall (*Y*) on November rainfall (*X*) from 1971 to 1980.

RANSAMPLE• The data set consists of 100 random observations on (*X, Y, Z*) where *X* follows the standard normal distribution, *Y* given *X* follows normal distribution with mean *X* and standard deviation 1, *Z* given (*X*,*Y*) follows normal distribution with mean *X* and *Y* and standard deviation 1. The data set is generated by using SYSTAT.

RATGROWTH • Milliken and Johnson (1992). This experiment involved studying the effect of a dose of a drug on the growth of rats. The data set consists of the growth of fifty rats, where ten rats were randomly assigned to each of the five doses of the drug. The weights were obtained each week for eleven weeks. The variables are - *DOSE, RAT, WEEK, WEIGHT*

RATS• Morrison (2004). For these data, six rats were weighed at the end of each of five weeks *(WEIGHT(1)* to *WEIGHT(5))*.

RCITY• Adapted from a Swiss Bank pamphlet: These data include 46 international cities *(CITY$)*, the name of continental region *(REGION$)*, average working hours per week *(WORKWEEK)*, working time (in minutes) to buy a hamburger and a large portion of french fries *(BIG_MAC)*, average cost (in U.S. dollars per basket) of a basket of goods and services *(LIVECOST)*, net hourly earnings *(EARNINGS)*, and percentage of taxes security paid by worker *(PCTTAXES)*.

REACT• These data involve yields of a chemical reaction *(YIELD)* under various combinations of four binary factors (*A*, *B*, *C*, and *D*). Two reactions are observed under each combination of experimental factors, so the number of cases per cell is two.

REGORTHO• The data set consists of 25 random observations on (X,Y) with X2 = X^2, X3 = X^3, X4=X^4 and X5=X^5,where X follows normal distribution with mean 5 and standard deviation 1,Y given X follows normal distribution with mean 1-X+X^2 and standard deviation 1.The data set is generated by using SYSTAT. The variables in this data set are *X, Y, X2, X3, X4, X5*.

REPEAT1• Winer (1971). These data contain two grouping factors *(ANXIETY* and *T ENSION)* and one trial factor *(TRIAL(1)* to *TRIAL(4))*.

REPEAT2• Winer (1971). This data set has one grouping factor *(NOISE)* and two trial factors (period and dial). The trial factors must be entered as dependent variables in a MODEL statement, so the variables are named *P1D1*, *P1D2*,..., *P3D3*. For example, *P1D2* means a score in the {period1, dial2} cell.

RIESBY• Reisby et al. (1977) studied the relationship between desipramine and imipramine levels in plasma in 66 depressed patients classified as either endogenous or nonendogenous. After receiving a placebo for one week, the researchers administered a dose of imipramine each day for four weeks, recording the imipramine and desipramine levels at the end of each week. At the beginning of the placebo week and at the end of each week (including the placebo week), patients received a score on the Hamilton depression rating scale. A diagnosis of endogenous or non-endogenous depression was made for each patient. Although the total number of subjects in this study was 66, the number of subjects with all measures at each of the weeks fluctuated: 61 at week 0 (start of placebo week), 63 at week 1 (end of placebo week), 65 at week 2 (end of ¯first drug treatment week), 65 at week 3 (end of second drug treatment week), 63 at week 4 (end of third drug treatment week), and 58 at week 5 (end of fourth drug treatment week).The variables are- *ID, HAMD, CONSTANT, WEEK, ENDOG, ENDOGWK*.

RLONGLEY• Longley (1967). The data were originally used to test the robustness of least-squares packages to multicollinearity and other sources of ill conditioning. The variables in his data set are *TOTAL, DEFLATOR, GNP, UNEMPLOY, ARMFORCE, POPULATN,* and *TIME*.

ROCKET• Components *A, B,* and *C* are mixed to form a rocket propellant. The elasticity of the propellant *(ELASTIC)* was the dependent variable. The other variable is *RUN*.

ROHWER• Timm (2002). The data set is based on the performance of 32 kindergartens in three standardized tests, peabody picture vocabulary test (*PPVT*), Raven progressive matrices test (*RPMT*) and a student achievement test (*SAT*).The independent variables are, named (*N*), still (*S*), named still (*NS*), named action (*NA*), sentence still (*SS*).

ROTATE• Metzler and Shepard (1974). These data measure reaction time in seconds *(RT)* versus angle of rotation in degrees *(ANGLE)* in a perception study. The experiment measured the time it took subjects to make “same” judgments when comparing a picture of a three dimensional object to a picture of possible rotations of the object.

ROTHKOPF• Rothkopf (1957). These data are adapted from an experiment by Rothkopf in which 598 subjects were asked to judge whether Morse code signals presented two in succession were the same. All possible ordered pairs were tested. For multidimensional scaling, the data for letter signals is averaged across sequence and the diagonal (pairs of the same signal) is omitted. The variables are *A* through *Z*.

RYAN• Ryan (2002). *Y1* and *Y2* are the control variables and *SAMPLE* is the sample identifier.

SALARY• These data compare the low and high salaries of executives in a particular firm.The variables are- *SEX*, *EARNINGS,* and *COUNT*.

SCHOOLS• Neter, Kutner, Nachtsheim and Wasserman (2004). These data comprise a nested design where two teachers from each of three different schools are rated. *SCHOOL* indicates the school that the case describes. Each teacher variable *(TEACHER(1–3))* represents a different school; a value of “1” indicates teacher 1 for that school, “2” indicates teacher 2 for that school, and “0” indicates that the teacher does not teach at that school. *LEARNING* measures the teacher’s effectiveness (the higher, the better).

SCORES• Hand at al. (1996). The data set shows the results of 10 students sitting 14 examination papers for a degree in Statistics. Each result is a percentage. The variables are: *TEST1….TEST14*.

SERUM• Crowder and Hand (1990). The data set consists of the antibiotic serum levels with two types of drugs applied to the same group of volunteers in two phases at different time points (*TIME1, TIME2, TIME3, TIME6*).

SICKDATE• The data file lists the diagnosed date of each patient's illness (*DIAGDATE*) and the date each died (*MORTDATE*). These dates are listed in day-of-the-century format.

SIMUL1 and SIMUL2• These data contain three variables: *Y*, *I*, and *J*. *Y* is generated from $N(0, 1.5^2)$.

SLEEPDM• Allison and Cicchetti (1976). This data set contains information from a study on the effects of physical and biological characteristics and sleep patterns influencing the danger of a mammal being eaten by predators. The study includes data on the hours of dreaming and non-dreaming sleep, gestation age, and body and brain weight for 62 mammals. The variables are-

SPECIES$	Type of species
BODY	Body weight of the mammal in kg
BRAIN	Brain weight of the mammal in g
SLO_SLP	Number of hours of nondreaming sleep
DREAM_SLP	Number of hours of dreaming sleep
TOTAL_SLEEP	Number of hours of total sleep
LIFE	The life span in years
GESTATE	The gestation age
PREDATION	Index of predation as a quantitative variable
EXPOSURE	Index of exposure as a quantitative variable
DANGER	Danger index as a quantitative variable (based on the above two indices)

SMOKE• Greenacre (1984). The data comprise a hypothetical smoking survey in a company. The variables are: *STAFF, SMOKE, FREQ.*

SOCDES• Strahan and Gerbasi (1972). The 20-item version of the Social Desirability Scale was administered as embedded items in another test to 359 undergraduate students in psychology. The social desirability items were scored for the "social desirability" of the response and coded as 0's and 1's in this SYSTAT data set.

SOFTWARE1• Musa (1979). The data set consists of failure times (*TIME)* (in CPU seconds, measured in terms of execution time) of a real-time command and control software system. The variable *INTER* contains inter-failure times.

SOIL• Zinke and Stangenberger. These data were taken from a compilation of worldwide carbon and nitrogen soil levels for more than 3500 scattered sites. The full data set is available at the U.S. Carbon Dioxide Information Analysis Center (CDIAC) site on the World Wide Web. The subset included in SYSTAT pertains to the continental U.S. Duplicate measurements at single sites are averaged.

LAT	Sample site latitude
LON	Sample site longitude
STATISTC$	Mean
CARBON	Carbon content in kg/m^2

NITRO	Nitrogen content in kg/m^2
ELEV	Sample site elevation in meters

SPECTRO• Lindberg et al. (1983).The data set was used to fit a spectrographic model to help determine the amounts of three compounds present in samples from the Baltic Sea: Lignin Sulfonate: pulp industry pollution (*LS*), Humic Acids: natural forest products (*HA*), and optical whitener from detergent (*DT*). The data set consists of 16 samples of known concentrations of *LS*, *HA* and *DT*, with spectra based on 27 frequencies. (or equivalently, wavelengths)

SPECTROMETERS. Two mass spectrometers (*SPECTROMTR$*) were compared for accuracy in measuring the ratio of ^{14}N to ^{15}N. Three plots of land (*PLOT*) treated with ^{15}N were used and from every plot two soil samples (*SAMPLE*) were taken. Each sample had two observations. The response variable *RATI*O is the ratio of ^{14}N to ^{15}N multiplied by 1000.

RATIO	Ratio of two soil measurements.
SPECTROMTR$	ID of a spectrometer (A, B).
PLOT	Plot number.
SAMPLE	Sample number

SPIRAL• These data consist of a spiral in three dimensions with the variables *X*, *Y*, *Z, R,* and *THETA*.

SPLINE• Brodlie (1980). These data are *X* and *Y* coordinates taken from a figure in Brodlie's discussion of cubic spline interpolation.

SPNDMONY• Chatterjee, Hadi and Price (2000). In this data set, *SPENDING* is consumer expenditures, and *MONEY* is money stock in billions of dollars in each quarter of the years 1952–1956 *(DATE).*

STRESS• Brown (2006), adapted from Folkman& Lazarus (1970), Tobin, Holroyd, Reynolds, & Wigal (1989). The data set is a covariance matrix of 12 manifest variables which represents four distinctive ways of coping with stressful events of 275 college undergraduates. The variables are-

P1-P3	Problem solving
C1- C3	Cognitive restructuring
E1-E3	Express Emotions
S1- S3	Social Support

SUBWORLD• The data in the file *SUBWORLD* are a subset of cases and variables from the *OURWORLD* file.

SUBWRLD2• The dataset is a transformation of *SUBWORLD* data set. The variables are standardized and sorted in descending *GDP_CAP* order and transformed them to log base 10 units to symmetrize the distributions before they are standardized. only cases, with values for all the variables have been included.

SUB_OURWORLD• It's a subset of data set *OURWORLD* in SYSTAT. The variables are:

CTEDUC	Expenditure (in US dollars) per person for education in the city
CTHEALTH	Expenditure (in US dollars) per person for health in the city
RUEDUC	Expenditure (in US dollars) per person for education in rural area
RUHEALTH	Expenditure (in US dollars) per person for health in rural area[2]

SUNSPTDM• Andrews and Herzberg (1985). The data set consists of a calculated relative measure of the daily number of sunspots compiled from the observations of a number of different observatories.

YEAR	The year the observations
JAN-DEC	The relative measure of sunspots for the indicated month
ANNUAL	The mean relative measure of sunspots for the entire year

SURVEY2• In Los Angeles (circa 1980), interviewers from the Institute for Social Science Research at UCLA surveyed a multiethnic sample of 256 community members for an epidemiological study of depression and help-seeking behavior among adults (Afifi and Clark, 2004). The CESD depression index was used to measure depression. The index is constructed by asking people to respond to 20 items: "I felt I could not shake off the *blues...,*" "My sleep was *restless,*" and so on. For each item, respondents answered "less than 1 time per day" (score 0); "1 to 2 days per week" (score 1); "3 to 4 days per week" (score 2), or "5 to 7 days" (score 3). Responses to the 20 items were summed to form a *TOTAL* score. Persons with a CESD *TOTAL* greater than or equal to 16 are classified as depressed. Variables include:

ID	Subject identification number
SEX	1 = male; 2 = female
AGE	Age in years at last birthday
MARITAL	1 = never married; 2 = married; 3 = divorced; 4 = separated; 5 = widowed
EDUCATN	1 = less than high school; 2 = some high school; 3 = finished high school; 4 = some college; 5 = finished bachelor's degree; 6 = finished master's degree; 7 = finished doctorate
EMPLOY	1 = full time; 2 = part time; 3 = unemployed; 4 = retired; 5 = houseperson; 6 = in school; 7 = other
INCOME	Thousands of dollars per year
SQRT_INC	Square root of income

RELIGION	1 = Protestant; 2 = Catholic; 3 = Jewish; 4 = none; 6 = other
BLUE to DISLIKE	Depression items
TOTAL	Total CESD score
CASECONT	0 = normal; 1 = depressed (CESD ≥ 16)
DRINK	1 = yes, regularly; 2 = no
HEALTHY	General health? 1 = excellent; 2 = good; 3 = fair; 4 = poor
CHRONIC	Any chronic illnesses in last year? 0 = no; 1 = yes

SURVEY3• Marascuilo and Levin (1983) and Cohen (1988). This is a fictitious data set consisting of responses of 640 men (*COUN T*) to the question "Does a woman have the right to decide whether an unwanted birth can be terminated during the first three months of pregnancy?" The response alternatives were cross-tabulated with religion. *RELIGION$* and *RESPONSE$* are represented by ordinal numbers in the data.

SWEAT• Johnson and Wichern (2002). The data set consists of perspiration measurements from 20 healthy females, on three variables, sweat rate (*SWEAT_RATE*), sodium content (*SODIUM*), and potassium content (*POTASSIUM*).

SWETSDTA• Swets, Tanner, and Birdsall (1961) and reported by Swets and Pickett (1982). This example shows frequency data for two detectors in a study. Each of the subjects in the experiment used a six-category rating scale (*RATING*) to indicate his or her confidence that a signal was present on each of 597 trials when the signal was present, and on 591 randomly-mixed trials on which the signal was not present. The *COUNT* variable shows the number of times a subject gave a particular rating to a given signal state. The identifier *SUBJ* is a numeric variable in this case.

SYMP• The dataset consists of 18 representative symptoms that have been taken and tallied for how many times they have occurred together in 50 diseases. The variables *DIM1* and *DIM2* are the coordinates in two dimensions after performing the multidimentional scaling on the cooccurrences of symptoms for 50 diseases. The other variables *LYME, MALARIA, YELLOW, RABIES* and *FLU* (5 among the 50 diseases) are the dichotomous variables which indicate weather a particular symptom is present or not.

TABLET• Netmaster Statistics Courses. An experiment was undertaken to compare two methods, HPLC and NIR, to ascertain the amount of active content in tablets. The tests have been applied to the same set of ten tablets, breaking each tablet into two halves, and applying one method to each half. The resulting data consists of the following variables - *TABLET, HPLC* and *NIR*.

TABLET2• The data set is the indexed form of data set *TABLET*.

TARGET• The data set is hypothetical. It describes the success of an arrow throwing machine to hit the target. The variables in the data set are:

NOOFTRAILS	Number of trails
NOOFEVENTS	Number of events
HEIGHT	Height (cms) at which the machine is placed
FORCE	Force (newton) applied to hit the target

TEACH• Mickey et al. (2004). The data set contains the two teaching methods and three teachers. Each teacher uses each teaching method with four different batches of students. The performance of each batch is measured by the average score of the batch in a common examination. The variables are - *SCORE, TEACHER* and *METHOD*.

TEACHER• Timm (2002). The data set was obtained at the University of Pittsburgh by J. Raffaele to analyze the reading comprehension and reading rate of students. The teachers were nested within classes. The classes were noncontract and contract classes. The variables are-

CLASSES$	Types of classes
TEACHERS$	Teachers
READRATE	Reading rate
READCOMPRE	Reading comprehension

TETRA• These data are from a bivariate normal distribution. Variables include *X*, *Y* and *COUNT* (frequency).

THREAD• Taguchi et al. (1989). The data set consists of the tensile strength (*STRENGTH*), in kilograms per millimeter squared, of thread samples, collected every day for two months (*MONTH*) of production.

TRANSAMSTERDAM• Franses and Dick van Dijk (2000). The data utilized the index of the stock markets in Amsterdam (EOE). The exchange rate is Dutch guilder. The sample period for the stock index runs from January 6, 1986 until December 31, 1997. The original series is sampled 5 days in a week. The variables are:

AMSTEOE : Daily indices of stock data of Amsterdam in Netherlands. There is 5 days in a week, opening date 1/06/1986 ending date 12/31/1997.

TRAMSTOCK: Simple difference transforms series of AMSTEOE.

TIME : Time is sample case number

TRIAL• These data contain six variables, *X (1)... X(5),* and *SEX$*.

TVFSP• Hedeker and Gibbons (1996). The data set is from the Television School and Family Smoking Prevention and Cessation Project. Hedeker and Gibbons looked at the effects of two factors on tobacco use for students in 28 Los Angeles schools. One factor involved the use of a social-resistance curriculum or not. The other factor was the presence or absence of a television intervention. Crossing these two factors yields four experimental conditions, which were randomly assigned to the schools. Students were measured on tobacco and health knowledge both before and after the introduction of the two factors.

TYPING• These data show the average speeds of typists in three groups, using typing speed *(SPEED)* and a character or numeric code for the machine used *(EQUIPMNT$)*.

US• State and Metropolitan Area Data Book (1986), Bureau of the Census; The World Almanac (1971).

POPDEN	People per square mile
PERSON	FBI-reported incidences, per 100,000 people, of personal crimes (murder, rape, robbery, assault)
PROPERTY	Incidences, per 100,000 people, of property crimes (burglary, larceny, auto theft)
INCOME	Per capita income
SUMMER	Average summer temperature
WINTER	Average winter temperature
LABLAT	Latitude in degrees at the center of each state
LABLON	Longitude at the center of each state
RAIN	Average inches of rainfall per year

USCORR• The data set is a correlation matrix among 16 variables from the *USSTATES* data file**.** Following are the variable names:

ACCIDENT	*CARDIO*	*CANCER*	*PULMONAR*	*PNEW_FLU*
DIABETES	*LIVER*	*VIOLRATE*	*PROPRATE*	*AVGPAY*
TEACHERS	*TCHRSAL*	*MARRIAGE*	*DIVORCE*	*HOSPITAL*
DOCTOR				

USCOUNT• Taken from the *US* data. These data are the means of *PERSON* (personal crimes) and *PROPERTY* (property crimes) within *REGION$*. The *COUNT* variable shows the number of states over which the means were computed.

USINCOME• These data are on the average income *(INCOME)* of a few regions. The variables are *DIVISION$*, *COUNT*, *INCOME*.

USSTATES• State and Metropolitan Area Data Book (1986). The variables are -

REGION and REGION$	Divide the country into four regions
DIVISION and DIVISION$	Divide the country into nine regions
LANDAREA	Land area in square miles, 1980
POP85	1985 population in thousands
ACCIDENT	Number of deaths by accident per 100,000 people
CARDIO	Number of deaths from major cardiovascular disease per 100,000 people
CANCER	Number of deaths from cancer per 100,000 people
PULMONAR	Number of deaths from chronic obstructive pulmonary disease per 100,000 people
PNEU_FLU	Number of deaths from pneumonia and influenza per 100,000 people
DIABETES	Number of deaths from diabetes mellitus per 100,000 people
LIVER	Number of deaths from chronic liver disease and cirrhosis per 100,000 people
DOCTOR	Number of active, nonfederal physicians per 100,000
HOSPITAL	Number of hospitals per 100,000 in 1988
MARRIAGE	Number of marriages in thousands in 1989
DIVORCE	Number of divorces and annulments in thousands in 1989
TEACHERS	Number of teachers in thousands
TCHRSAL	Average salary for teachers for the 1990 year
HSGRAD	Number of public high school graduates in the 1982–83 school year
AVGPAY	Average annual pay for a worker in 1989
TOTALSLE	Total sale
VIOLRATE	Violent crime rate per 100,000 people in 1989
PROPRATE	Rate of property crimes per 100,000 people in 1989
PERSON	Number of persons who commit crimes
POP90	Population in thousands in 1990 as cited in the *New York Times*
ID$	Name of each state in the United States
COUNT	Number associated with the state
MSTROKE and FSTROKE	Risk of stroke per 100,000 males and females (adjusted to weight each state's various age groups equally)
INCOME89	Median household income in 1989
INCOME	Income in 1991
BUSH, PEROT, and CLINTON	Vote count in 1000 for each candidate in the 1992 presidential election

ELECVOTE	Number of electoral votes each state received in the 1992 presidential election
PRES_88$	Number of electoral votes each state received in the 1988 presidential election
GOV_93$	Newly elected governor's political party in each state after winning the 1993 gubernatorial races
GOV_92$	Winning political parties in the 1992 gubernatorial races
POVRTY91	Census Bureau's estimate of the percentage of Americans living below the poverty level in 1991
POVRTY90	Poverty estimates for 1990
TORNADOS	Number of tornados per thousand square miles from 1953 to 1991
HIGHTEMP	Average high temperature
LOWTEMP	Average low temperature
RAIN	Average annual rainfall
SUMMER	Average summer temperature
WINTER	Average winter temperature
POPDEN	Population density
LABLON, LABLOT	Longitude and latitude at the center of the state according to the *World Almanac and Book of Facts* (1992), Pharo Books, New York
GOVSLRY	Salaries for U.S. governors

USVOTES• This data file breaks down the votes for *CLINTON, BUSH,* and *PEROT* by *DIVISION$*.

VOLTAGE• Montgomery and Peck (2002). The data set contains observations on the battery voltage drop (VOLTAGE) of a guided missile motor over the time of the missile flight (TIME).

WATERQUALITY• Databook (2005). The data file contains measurements of several physio-chemical properties of water, in five different cities. The variables used are CHLORIDES and SULPHATES.

WESTWOOD• Neter, Kutner, Nachtsheim and Wasserman (1996). A spare part is manufactured by the Westwood Company once a month. The lot sizes manufactured vary from month to month because of differences in demand. These data show the number of man-hours of labor for each of 10 lot sizes manufactured. The variables are *PROD_RUN, LOT_SIZE,* and *MAN_HRS*.

WILL• Williams (1986). *RESPONSE* is the dependent variable, *LDOSE* is the logarithm of the dose (stimulus), and *COUNT* is the number of subjects with that response.

WILLIAMS• Cochran and Cox (1957). These data are from a crossover design for an experiment studying the effect of three different feed schedules *(FEED)* on milk production by cows *(MILK)*. The design of the study has the form of two 3 x 3 Latin squares. *PERIOD* represents the period. *RESIDUAL* indicates the treatment of the preceding period. Other variables include number assigned to the cow *(COW)* and the Latin square number *(SQUARE)*.

WILLMSDM• Hubert (1984). This data set contains the results of a bioassay conducted to determine the concentration of nicotine sulfate required to kill 50% of a group of common fruit flies. The experimenters recorded the number of fruit flies that are killed at different dosage levels. The variables are-

RESPONSE	The dependent variable, which is the response of the fruit fly to the dose of nicotine sulfate (stimulus)
LDOSE	The logarithm of the dose
COUNT	The number of fruit flies with that response

WINER• Winer (1971). The data are from a design with two trials *(DAY(1–2))*, one covariate *(AGE)*, and one grouping factor *(SEX)*.

WORDS• Caroll, Davies, and Richmond (1971). The data set contains the most frequently used words (*WORD$*) in American English. Three measures have been added to the data. The first is the (most likely) part of speech (*PART$*). The second is the number of letters (*LETTERS*) in the word. The third is a measure of the meaning (*MEANING*). This admittedly informal measure represents the amount of harm done to comprehension (1 = a little, 4 = a lot) by omitting the word from a sentence.

WORLD• Global mapping. The variables include *MAPNUM, MAXLAT, MINLAT, MINLON, MAXLON, LABLAT, LABLON,* and *COLOR$*.

WORLD95M• For each of 109 countries, 22 variables were culled from several 1995 almanacs—including life expectancy, birth rate, the ratio of birth rate to death rate, infant mortality, gross domestic product per capita, female and male literacy rates, average calories consumed per day, and the percentage of the population living in cities.

WORLDDM• Wilkinson, Blank, and Gruber (1996). This data set contains 1990 information on 30 countries including birth and death rates, life expectancies (male and female), types of government, whether mostly urban or rural, and latitude and longitude. The variables are-

COUNTRY$	Country name
BIRTH_RT	Number of births per 1000 people in 1990
DEATH_RT	Number of deaths per 1000 people in 1990
MALE	Years of life expectancy for males
FEMALE	Years of life expectancy for females

GOV$	Type of government
URBAN$	Rural or urban
LAT	Latitude of the country's centroid
LON	Longitude of the country's centroid

YOUTH• Harman (1976). It is a correlation matrix, consisting of measurements recorded for 305 females aged seven to seventeen: height, arm span, length of forearm, length of lower leg, weight, bitrochanteric diameter (the upper thigh), torso girth, and torso width.

References

Afifi, A. A. and Azen, S. P. (1974). *Statistical analysis*: *A computer oriented approach.* New York: Academic Press.

Afifi, A. A., May, S., and Clark, V. (2004). *Computer-aided multivariate analysis*, 4th ed. New York: Chapman & Hall.

Akima, H. (1978). A method of bivariate interpolation and smoth surface fitting for irregularly distributed data points. *ACM Transactions on Mathematical software*.

Allison and Cicchetti (1976). Sleep in mammals: Ecological and constitutional correlates. *Science*, 194, 732—734.

Anderson, E. (1935). The irises of Gaspe peninsula. *Bulletin of the American Iris Society*, 59, 2—5.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A collection of problems from many fields for the student and research worker*. New York: Springer-Verlag.

Ansfield, F., Klotz, J. and the central Oncology Group (1977). A phase III study comparing the clinical utility of four regiments of 5-fluorouracil. *Cancer,* 39, 34—40.

Atkinson, A. C. (1986). Aspects of diagnostic regression analysis, *Statistical Science*, 1, 397—402.

Automotive Industry Action Group (1995). *Statistical process control (SPC) reference manual.* Chrysler Corporation, Ford Motor Company, General Motors Corporation.

Barnett, V. D. and Lewis, T. (1967) A study of low-temperature probabilities in the context of an industrial problem. *Journal of the Royal Statistical Society*, Series A, 130, 177—206.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: John Wiley & Sons.

Beckman, R. J., Nachtsheim, C. J. and Cook, D. J. (1987). Diagnostics for mixed model analysis of variance. *Technometrics*, 29, 413-426.

Belsley, D. A., Kuh, E., and Welesh, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: John Wiley & Sons.

Bennett, R. M. and Desmarais, R. N. (1975). Curve fitting of aeroelastic transient response data with exponential functions. In *Flutter Testing Techniques*. Report of a conference held at Dayton Flight Research Center, Edwards, CA, October 9—10, 1975. Washington, DC: NASA. Pp. 43—58.

Birkes, D. and Dodge, Y. (1993). *Alternative methods of regression*. New York: John Wiley & Sons, pp. 177—183.

Bishop, Y. V. V., Fienberg, S. E., and Holland, F. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Bliss, C. I. (1967). *Statistics in biology*. New York: McGraw-Hill.

Borg, I. and Lingoes, J. (1987). *Multidimensional similarity structure analysi*s. New York: Springer Verlag.

Box, G. E. P., Jenkins, G. M, and Reinsel, G. (1994). *Time series analysis: Forecasting & control*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. I. (1984). *Classification and regression trees*. Belmont, Calif.: Wadsworth.

Breslow, N.and Day, N. E. (1980). *Statistical methods in cancer research, Vol II: The design and analysis of cohort studies.* Lyon: IARC.

Breyfogle, F. W. III (2003). *Implementing six sigma: Smarter solution through statistical methods*. 2nd ed. New York: John Wiley & Sons.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods*. Springer-Verlag.

Brodlie, K. W. (1980) *A review of methods for curve and function drawing, in Mathematical Methods in Computer Graphics and Design*, pp 1-37. Academic Press, New York and London.

Brownlee, K. A. (1960). *Statistical theory and methodology in science and enginnering*. New York: John Wiley & Sons.

Cameron, E. and Pauling, L. (1978). Supplemental ascorbate in the supportive treatment of cancer: Reevaluation of prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Sciences,* USA, 75, 4538—4542.

Carey, J. R., Liedo, P. Orozco, D., and Vaupel, J. W. (1992), "Slowing of Mortality Rates at Older Ages in Large Medfly Cohorts," *Science*, 258, 457—461.

Caroll, J. B., Davies, P., and Richmond. B. (1971). *The word frequency book.* Boston, Mass.: Houghton-Mifflin.

Chambers, J. M., Cleveland, W. S., Kleiner, B., Tukey, P. A. (1983). *Graphical methods for data analysis*. Duxbery Press, Boston.

Chatterjee, S., Hadi, A. S., and Price, B. (2000). Regression analysis by example. 3rd ed., New York: John Wiley & Sons.

Clarke, C. P. Y. (1987). Approximate confidence limits for a parameter function in nonlinear regression. *Journal of the American Statistical Association,* 85, 544—551.

Clausen, S. E. (1998). Applied correspondence analysis: An introduction. *University Paper*

Series on Quantitative Application in Social Science, 7-121. Thousand Oaks, CA: Sage.
Cleveland, W. S. (1993). *Visualizing Data*. Summit, NJ: Hobart Press.
Cochran, W. G. and Cox, G. (1957). *Experimental designs*. New York: John Wiley & Sons.
Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement,* 12, 425—434.
Cohen, P. and Brook, J. (1987). Family factors related to the persistence of psychopathology in childhood and adolescence. *Psychiatry*, 50, 332-345.
Conover, W. J. (1999). *Practical nonparametric statistics*. 3rd ed. New York: John Wiley & Sons, pp. 371—373.
Cook, R. D. and Weisberg, S. (1990). Confidence curves in nonlinear regression. *Journal of The American Statistical Association,* 85, 544—551.
Cornell, J. A. (1985). Mixture Experiments. In Koltz, S. and Johnson, N. L. (Eds.). *Encyclopedia of Statistical Sciences, Vol. 5,* 569—579. New York: John Wiley & Sons.
Cox, D. R. (1970). *The analysis of binary data.* New York: Halsted Press.
Crowder, M. J. and Hand, D. J. (1990). *Analysis of repeated measures*. London: Chapman & Hall.
DASL (2005). Available at:
http://lib.stat.cmu.edu/DASL/Stories/SteppingandHeartRates.html
Databook (2005). Available at:
http://stats.unipune.ernet.in/Databook/DatasetsPUNE/Waterquality.xls
Davis, D. J. (1977). An analysis of some failure data. *Journal of the American Statistical Association*, 72, 113-150.
Devor, R. E., Chang, T., Sutherland, J. W. (1992). *Statistical Quality Design and Control* New York: MacMillan.
Draper, N. R. and Smith, H. (1998). *Applied regression analysis*, 3rd ed., New York: John Wiley & Sons.
Duncan, O. D., Haller, A. O., and Portes, A. (1971). Peer influence on aspirations, a reinterpretation. *Casual Models in Social Sciences,* H. M. Blalock, ed. 219—244. Aldine-Atherstone.
Efron, B. and Tibshirani, R. (1993). *An Introduction to the bootstrap*. Chapman and Hall, New York, London.
Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology*, 38, 467—474.
Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics*, 28, 51-60.
Fisher, R. A. (1935). *The design of experiments*. 7th ed. New York: Hafner.
Fisher, R. A. (1936). The use of multiple measurments in taxonomic problems. *Annals of Eugenics,* 7, 179-188.
Flury, B, and Riedwyl, H. (1988). *Multivariate statistics: A practical approach*. London: Chapman and Hall.

Franses, P. H., and Dick van Dijk. (2000). *Non-linear time series models in empirical finance*. Cambridge University Press Datastream.

Frets, G. P. (1921). Heredity of head form in man. *Genetica,* 3, 193—384.

Gaver, D. P. and O'Muircheartaigh, I. G. (1987). Robust empirical bayes analysis of event rates, *Technometrics,* 29, 1—15.

Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric statistical inference*, 4th ed., Boca Raton, Florida: CRC Press.

Gilfoil, D. M. (1982). Warming up to computers: A study of cognitive and affective interaction overtime. In *Proceedings: Human factors in computer systems.* Washington, D. C.: Association for Computing Machinery.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

Greco, W. R., Priore, RL, Sharma, M., Korytnyk, W. (1982). ROSFIT: An enzyme kinetics nonlinear regression curve fitting package for a microcomputer. *Computers and Biomedical Research,*15, 39—45.

Green, P. F. and Carmone, F. J. (1970). *Multidimensional Scaling and related technique in marketing analysis*. Boston, MA: Allyn and Bacon.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. New York: Academic Press.

Gujarati, D. N. (1995). *Basic Econometrics*, 3th ed. New York: McGraw-Hill.

Gujarati, D. N. (2003). *Basic Econometrics*, 4th ed. New York: McGraw-Hill.

Hand, D. J., Daly, F., Lunn A. D., McConway, K. J. and Ostrowski, E. (Editors) (1996). *A handbook of data sets*. London: Chapman & Hall.

Harman, H. H. (1976). *Modern factor analysis*. 3rd ed., Chicago: University of Chicago Press.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: John Wiley & Sons.

Hedeker, D. and Gibbons, R. D. (1996). MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine*, 49, 229-252.

Helm, C. E. (1959). A multidimensional ratio scaling analysis of color relations. *Technical Report,* Princeton University and Educational Testing Service, June 1959.

Hocking, R. R. (1985). *The analysis of linear models*. Monterrey, CA: Brooks-Cole.

Hocking, R. R. (2003) *Methods and Applications of Linear Models*, Second Edition, John Wiley & Sons.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods*, 2nd ed. New York: John Wiley & Sons.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression* 2nd ed. New York: John Wiley & Sons.

Hubert J. J. (1984). *Bioassay*. Second Edition. Dubuque, Iowa: Kendall Hunt.

Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York: John Wiley & Sons.

Jackson, J. E. (2003). *A user's guide to principal components*, John Wiley & Sons.

Jobson, J. D. (1992*). Applied multivariate data analysis, Vol II: Categorical and multivariate methods*. New York: Springer-Verlag.

John, P. W. M. (1971). *Statistical design and analysis of experiments.* New York: MacMillan.

Johnson, R. A. and Wichern, D.W. (2002). *Applied multivariate statistical analysis*, 5th ed. Engelwood Cliffs, N. J.: Prentice Hall.

Johnson, R. W. (1999). *The official NFL 1999 Record & Fact Book*. New York: Workman Publishing, 435.

Judge, G. G., Griffiths, W. E., Lutkepohl, H., Hill, R. C. and Lee, T. C. (1988). *Introduction to the theory and practice of econometrics*, 2nd ed., New York: John Wiley & Sons, pp. 275—318, pp. 453—454.

Kooijman, S. A. L. M. (1979). The description of point patterns. In R. M. Cormack and J. K. Ord (eds.),. *Spatial and Temporal Analysis in Ecology.* Fairland, Md.: International Co-operative Publishing House, pp. 305—332.

Kuehl, R. O. (2000). *Design of experiments: statistical principles of research design and analysis*. New York: Duxbury Thomson Learning.

Laner, S., Morris, P. and Oldfild, R. C. (1957). A random pattern screen. *Quarterly Journal of Experimental Psychology*, 9, 105—108.

Lange, T. R., Royals, H. E., and Connor, L. L. (1993). *Transactions of the American Fisheries Society*.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor analysis as a statistical method*. 2nd ed. New York: American Elsevier Publishing Company.

Lee, J. (1992). *Relationships Between Properties of Pulp-Fibre and Paper*, unpublished doctoral thesis. University of Toronto, Faculty of Forestry.

Lee, P. M. (1989). *Bayesian statistics: An introduction*, London: Edward Arnold. p. 179.

Lindberg, W., Persson, J. A. and Wold, S. (1983). Partial least squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate. *Analytical Chemistry*, 55, 643-648.

Long, L. H. (ed.) (1971). *The world almanac*. New York: Doubleday.

Longley, J. (1967). An appraisal of least squares program for the electronic computer from the point of view of the user manual. Journal of American Statistical Association, 62, 819—841.

Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy. *Biometrics*, 18, 455—477.

MacGregor, G. A., Markandu, N. D., Roulston, J. E., and Jones, J. C. (1979). Essential hypertension: Effect of an oral inhibitor of angiotensin-converting enzyme. *British Medical Journal*, 2, 1106-1109.

McFadden, D. (1979). Quantitative methods for analyzing travel behavior of individuals: Some recent developments. In D. A. Hensher and P. R. Stopher (eds.): *Behavioral Travel Modelling*. London: Croom Helm.

Maltz, M. D. (1984). *Recidivism.* New York: Academic Press.

Marascuilo, L. A., and Levin, J. R. (1983). *Multivariate statistics in the social sciences*. Monterey, Calif.: Brooks/Cole.

Mels, G. and Koorts, A. S. (1989). *Casual Models for various job spects.* SAIPA*,*
24, 144—156.

Mendenhall, W., Beaver, R. J., and Beaver, B. M. (2002). *A brief introduction to probability and statistics. Pacific Grove, CA:* Duxbury. p. 424.

Messina, W. S. (1987). *Statistical quality control for manufacturing managers.* New York: John Wiley & Sons.

Metzler, J., and Shepard, R. N. (1974). *Transformational studies of the internal representation of three-dimensional objects*. Hillsdale, NJ: Erlbaum.

Mickey, R. M., Dunn, O. J., and Clark, V. A. (2004). Applied statistics: Analysis of variance and regression. New York: John Wiley & Sons.

Milliken, G. A. and Johnson, D. E. (1984). Analysis of messy data, Vol. 1: *Designed Experiments.* New York: Van Nostrand Reinhold.

Milliken, G. A. and Johnson, D. E. (1992). *Analysis of messy data: Designed experiments*, Vol I. Chapman and Hall.

Montgomery, D. C., Peck, E. A. and Vining G. G. (2001). *Introduction to linear regression analysis*, 3rd edition. New York: John Wiley & Sons.

Montgomery, D. C., Peck E. A., and Vining G. G. (2006). *Introduction to linear regression analysis*, 4th ed. Hoboken, N. J.: Wiley-Interscience.

Montgomery, D. C. and Runger, G. C. (1993). Gauge capability and designed experiments. Part 1: Experimental design models and variance component estimation, *Quality Engineering*, 6(1), 115.

Montgomery, D. C. (2005). *Introduction to statistical quality control*. 5th ed. New York: John Wiley & Sons.

Morrison, A. S., Black, M. M., Lowe, C. R., MacMahon, B., and Yuasa, S. Y. (1990). Some international differences in histology and survival in breast cancer. International Journal of Cancer, 11, 261-267.

Morrison, D. F. (2004). *Multivariate statistical methods*. 4th ed. Pacific Grove CA: Duxbury Press.

Morrison, K. J. and Zeppa, R. (1963). Histamine-introduced hypothesion due to morphine and arfonad in the dog. *Journal of Surgical Research* 3, 313-317.

Musa, J. D. (1979) *Software reliability data*. Data and Analysis Centre for Software, Rome Air Development Center, Rome, NY.

Myers, R. H. and Montgomery, D. C. (2002). *Response surface methodology*, 2nd ed. New York: John Wiley & Sons.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (2004). *Applied linear regression models*. Homewood, IL: Irwin.

Netmaster Statistics Courses. *Available at*: http://www.dina.kvl.dk/~per/Netmaster/courses/st113/Data/datafiles/planks.txt

Nichols, C. E., Kane, V. E., Browning, M. T., and Cagle, G. W. (1976). *Northwest Texas pilot geochemical survey*, Union Carbide, Nuclear Division Technical Report (K/UR-1)

Ott, R. L. and Longnecker, M. (2001). *Statistical methods and data analysis*, 5th edition. Pacific Grove, CA: Duxbury. p. 223.

Pearson, K. and Lee, A. (1903). On the laws of inheritance in man. I. Inheritance of physical characters. *Biometrika*, 2, 357—462.

Prentice, R. L. (1973). Exponential survival with censoring and explanatory variables. *Biometrika*, 60, 279-288.

Rao, C. R (2002). *Linear Statistical Inference and its Application*, 2nd ed., John Wiley & Sons.

Reisby, N., Gram, L. F., Bech, P., Nagy, A., Petersen, G.O., Ortmann, J., Ibsen, I., Dencker, S. J., Jacobsen, O., Krautwald, O., Sondergaard, I., and Christiansen, J. (1977). Imipramine: clinical effects and pharmacokinetic variability, *Psychopharmacology* 54, 263-272.

Robinson, D. (1987). Estimation and use of variance components. *The Statistician*, 36, 3-14.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired associate learning tasks. *Journal of Experimental Psychology*, 53, 94—101.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*, New York: John Wiley & Sons.

Ryan, T. P. (2002). *Statistical methods for quality improvement.* New York: John Wiley & Sons.

Schiffman, S. S., Reynolds, M. L., and Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods and applications.* New York: Academic Press.

Simonoff, J. S. (2003). *Analyzing categorical data*. New York: Springer-Verlag.

Smith, G. M. (2001). *Statistical process control and quality improvement*. Upper Saddle River, NJ: Prentice—Hall. p. 474.

Stouffer, S. A., Guttmann, L., Suchman, E. A., Lazarsfeld, P. F., Staf, S.A., and Clausen, J. A. (1950). *Measurement and prediction*. Princeton, N. J.: Princeton University Press.

Strahan, R. and Gerbasi, K. C. (1972). Short, homogeneous versions of the Crowne-Marlowe social desirability scale. *Journal of Clinical Psychology,*28, 191-193.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of diagnostic systems*. New York: Academic Press.

Swets, J. A, Tanner, W. P., and Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, 68, 301-340.

Taguchi, G., El Sayed, E. A., and Hslang, T.(1989*). Quality engineering in production systems*. New York: McGraw-Hill. pp. 32—41.

The Open University (1981) S237: *The Earth: Structure, composition and evolution*.

Thomson, A. and Randall-Maciver, R. (1905) *Ancient Races of the Thebaid*. Oxford: Oxford University Press.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer- Verlag.

Walser, P. (1969). Untersuchung über die Verteilung der Gerburtstermine bei dermehrgebärenden Frau, *Helvetica Paediatrica Acta, Suppl. XX* ad vol. 42, fasc. 3, 1-30.

Wheaton, B., Muthen, B., Alwin, D. F., and Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological methodology* D. R. Heise (Ed.), 84—136. San Francisco: Jossey-Bass.

Wilkinson, L. (1975). *The effect of involvement on similarity and preference structures*. Unpublished dissertation, Yale University.

Wilkinson, L. (1988). *SYSTAT. The system for statistics.* Evanston, IL: Systat, Inc.

Wilkinson, L. (2005). *The grammer of graphics. 2nd ed.* New York: Springer-Verlag..

Wilkinson, L., Blank, G., and Gruber, C. (1996). *Desktop data analysis with SYSTAT*. Upper Saddle River, N. J.: Prentice-Hall.

Wilkinson L. and Engelman, L. (1996), SYSTAT 7.0: New Statistics, pp. 235, SPSS Inc.

Williams, D. A. (1986). Interval estimation of the median lethal dose. *Biometrics*, 42, 641–645.

Winer, B. J. (1971). *Statistical principles in experimental design*. 2nd ed., New York: Mc-Graw Hill.

Winer B. J., Brown D. R., and Michels K. M. (1991). *Statistical principles in experimental design*, 3rd ed. New York: McGraw-Hill.

Wludyka, P. S. and Nelson, P. R. (1997). An analysis-of-means-type test for variances from normal populations. *Technometrics*, 39:3, 274-285.

Acronym & Abbreviation Expansions

A
ABS - absolute value
ACF - autocorrelation function
ACT - actuarial life table
AD test - Anderson Darling test
AIC - Akaike information criterion
AID - automatic interaction detection
ALT - alternative
ANCOVA - analysis of covariance
ANOVA - analysis of variance
AR - autoregressive
ARCH - Autoregressive Conditional Heteroskedasticity
ARIMA - autoregressive integrated moving average
ARL - average run length
ARMA - autoregressive moving average
ARS - adaptive rejection sampling
ASCII - American Standard Code for Information Interchange
ASE - asymptotic standard error
AVG - average

B
BC - Bray-Curtis similarity measure
BFGS - Broyden-Fletcher-Goldfarb-Shannon
BHHH - Berndt-Hall-Hall-Housman
BIC - Bayesian information criterion
BMP - Windows bitmap
BOOT - bootstrap

C
C&RT - classification and regression trees
CCF - cross-correlation function
cdf/CF - cumulative distribution function
CFA - confirmatory factor analysis
CGM - Computer graphics metafile: binary or clear text
CI - confidence interval
COL/col - column
CONV - convergence
COV - covariance
Cp - process capability index
Cpk-Process capability index for off-centered process
CR - confidence region
CRN - Cauchy random number
CSV - comma separated values
CV - coefficient of variation
CVI - cross validation index

D
DBF - Dbase files
dep. - dependent
DEVI - deviates (observed values - expected values)
df - degrees of freedom
DIM - dimension
DOS - disc operating system
DPMO - defects per million opportunities
DPU - defects per unit
DTA - Stata files

DWASS - Dwass-Steel-Chritchlow-Fligner pairwise comparisons test
DWLS - distance weighted least-squares

E
EM - expectation-maximization
EMF - Windows enhanced metafile
EWMA - exponentially weighted moving average

F

G
GARCH - Generalized Autoregressive Conditional Heteroskedaticity
GG - Greenhouse Geisser
GIF - Graphics Interchange Format
GLM - generalized linear models
GLS - generalized least-squares
GMA - geometric moving average
GN - Gauss-Newton method

H
H & L - Hosmer and Lemeshow
H-L trace - Holding-Lawley trace
HTML - hyper text markup language

I
IIDMC - independently and identically distributed Monte Carlo
IMPSAMPI - importance sampling integration
IMPSAMPR - importance sampling ratio
IndMH - Independent Metropolis-Hastings
INDSCAL - individual differences scaling
INITSAMP - initial sample
ITER - iterations

J
JB - Jarque-Bera
JMP - JMP v3.2 data files
JPEG/JPG - joint photographic experts group

K
K-M - Kaplan-Meier
K-S test - Kolmogorov-Smirnov test
KS1 - one sample Kolmogorov-Smirnov tests
KS2 - two sample Kolmogorov-Smirnov tests

L
LAD - least absolute deviations
LCL - lower control limit
LMS- least median of squares
LM Test - Lagrange Multiplier Tes
LR - likelihood ratio
LRDEV - likelihood ratio of deviate
LW - Lawless and Wang

M
MA - moving average
MAD - mean absolute deviation
MANCOVA - multivariate analysis of covariance
MANOVA - multivariate analysis of variance
MAX - maximum
MC Test - McLeod-Li Test
MCMC - Markov Chain Monte Carlo
MDS - multidimensional scaling
MIN - minimum
M-H- Metropolis-Hastings
ML - Maximum Likelihood
MLA - maximum likelihood analysis
MLE - maximum likelihood estimate
MML - maximum marginal likelihood
MS - mean squares
MSE - mean square error

MTW - MINITAB v11 data files
MU2 - Guttman's mu2 monotonicity coefficients

N
NR - Newton-Raphson

O
OC - operating characteristic
ODBC - open database capture and connectivity
OLS - ordinary least-squares

P
PACF - partial autocorrelation function
PCA - process capability analysis
PCF - iterated principal axis factoring
pdf - probability density function
PLS - partial least squares
pmf - probability mass function
PNG - Portable Network Graphics
PVAF/p.v.a.f. -- present value annuity factor
p-value - probability value

Q
QC - quality control

R
R & R - repeatability and reproducibility
RAMONA - Reticular Action Model or Near Approximation
ROC - receiver operating characteristic
RSE- robust standard errors
RSM- response surface methods
RTF - rich text format

S
SAV - SPSS files
SBC - Schwarz's Bayesian information criterion
sc - scale
SC - set correlation
SD - standard deviations
sd2/sas7bdat - SAS v9 files
SE/se/S.E. - standard error
SETCOR - Set and Canonical Correlations
SQL - structured query language
SQRT/SQR - square-root
SRWR - sum of rank weighted residuals
SS - sum of squares
SSCP - sum of squares and cross products
SYC/CMD - SYSTAT command Files
SYZ/SYD/SYS - SYSTAT data files
SYO - SYSTAT output files

T
TLOSS - Taguchi's Loss Function
TOL - tolerance
TSLS - Two-Stage Least Squares
TSQ chart - Hotelling's T^2 chart
TXT - text format

U
U chart - chart showing defects per unit
UCL - upper control limit
USL - upper specification limit
UTL - upper tolerance limit

V
VAR - variance
VIF - variance inflation factor

W
WMF - Windows metafile

X
XLS - excel format
X-MR chart - Individuals and moving range chart
XPT/TPT - SAS transport files
XTAB - Crosstabulations

Y

Z

Index

D

G

H

I

J

K

L

W