User's Guide I

Quick Start & Self Help, Introduction,
Data, Tools, and Graphics

NCSS
Statistical System

Published by
NCSS
Dr. Jerry L. Hintze
Kaysville, Utah

NCSS User's Guide I

Dr. Jerry L. Hintze
Kaysville, Utah 84037

Direct inquiries to:

NCSS
329 North 1000 East
Kaysville, Utah 84037
Phone (801) 546-0445
Fax (801) 546-3907
Email: support@ncss.com

NCSS License Agreement

Important: The enclosed Number Cruncher Statistical System (NCSS) is licensed by Dr. Jerry L. Hintze to customers for their use only on the terms set forth below. Purchasing the system indicates your acceptance of these terms.

1. **LICENSE.** Dr. Jerry L. Hintze hereby agrees to grant you a non-exclusive license to use the accompanying NCSS program subject to the terms and restrictions set forth in this License Agreement.

2. **COPYRIGHT.** NCSS and its documentation are copyrighted. You may not copy or otherwise reproduce any part of NCSS or its documentation, except that you may load NCSS into a computer as an essential step in executing it on the computer and make backup copies for your use on the same computer.

3. **BACKUP POLICY.** NCSS may be backed up by you for your use on the same machine for which NCSS was purchased.

4. **RESTRICTIONS ON USE AND TRANSFER.** The original and any backup copies of NCSS and its documentation are to be used only in connection with a single computer. You may physically transfer NCSS from one computer to another, provided that NCSS is used in connection with only one computer at a time. You may not transfer NCSS electronically from one computer to another over a network. You may not distribute copies of NCSS or its documentation to others. You may transfer this license together with the original and all backup copies of NCSS and its documentation, provided that the transferee agrees to be bound by the terms of this License Agreement. Neither NCSS nor its documentation may be modified or translated without written permission from Dr. Jerry L. Hintze.

*You may not use, copy, modify, or transfer **NCSS**, or any copy, modification, or merged portion, in whole or in part, except as expressly provided for in this license.*

5. **NO WARRANTY OF PERFORMANCE.** Dr. Jerry L. Hintze does not and cannot warrant the performance or results that may be obtained by using NCSS. Accordingly, NCSS and its documentation are licensed "as is" without warranty as to their performance, merchantability, or fitness for any particular purpose. The entire risk as to the results and performance of NCSS is assumed by you. Should NCSS prove defective, you (and not Dr. Jerry L. Hintze nor his dealers) assume the entire cost of all necessary servicing, repair, or correction.

6. **LIMITED WARRANTY ON CD.** To the original licensee only, Dr. Jerry L. Hintze warrants the medium on which NCSS is recorded to be free from defects in materials and faulty workmanship under normal use and service for a period of ninety days from the date NCSS is delivered. If, during this ninety-day period, a defect in a CD should occur, the CD may be returned to Dr. Jerry L. Hintze at his address, or to the dealer from which NCSS was purchased, and NCSS will replace the CD without charge to you, provided that you have sent a copy of your receipt for NCSS. Your sole and exclusive remedy in the event of a defect is expressly limited to the replacement of the CD as provided above.

Any implied warranties of merchantability and fitness for a particular purpose are limited in duration to a period of ninety (90) days from the date of delivery. If the failure of a CD has resulted from accident, abuse, or misapplication of the CD, Dr. Jerry L. Hintze shall have no responsibility to replace the CD under the terms of this limited warranty. This limited warranty gives you specific legal rights, and you may also have other rights, which vary from state to state.

7. **LIMITATION OF LIABILITY.** Neither Dr. Jerry L. Hintze nor anyone else who has been involved in the creation, production, or delivery of NCSS shall be liable for any direct, incidental, or consequential damages, such as, but not limited to, loss of anticipated profits or benefits, resulting from the use of NCSS or arising out of any breach of any warranty. Some states do not allow the exclusion or limitation of direct, incidental, or consequential damages, so the above limitation may not apply to you.

8. **TERM.** The license is effective until terminated. You may terminate it at any time by destroying NCSS and documentation together with all copies, modifications, and merged portions in any form. It will also terminate if you fail to comply with any term or condition of this License Agreement. You agree upon such termination to destroy NCSS and documentation together with all copies, modifications, and merged portions in any form.

9. **YOUR USE OF NCSS ACKNOWLEDGES** that you have read this customer license agreement and agree to its terms. You further agree that the license agreement is the complete and exclusive statement of the agreement between us and supersedes any proposal or prior agreement, oral or written, and any other communications between us relating to the subject matter of this agreement.

Dr. Jerry L. Hintze & **NCSS**, Kaysville, Utah

Preface

Number Cruncher Statistical System (**NCSS**) is an advanced, easy-to-use statistical analysis software package. The system was designed and written by Dr. Jerry L. Hintze over the last several years. Dr. Hintze drew upon his experience both in teaching statistics at the university level and in various types of statistical consulting.

The present version, written for 32-bit versions of Microsoft Windows (95, 98, ME, 2000, NT, etc.) computer systems, is the result of several iterations. Experience over the years with several different types of users has helped the program evolve into its present form.

Statistics is a broad, rapidly developing field. Updates and additions are constantly being made to the program. If you would like to be kept informed about updates, additions, and corrections, please send your name, address, and phone number to:

User Registration
NCSS
329 North 1000 East
Kaysville, Utah 84037

or Email you name, address, and phone number to:

Sales@NCSS.COM

NCSS maintains a website at **WWW.NCSS.COM** where we make the latest edition of NCSS available for free downloading. The software is password protected, so only users with valid serial numbers may use this downloaded edition. We hope that you will download the latest edition routinely and thus avoid any bugs that have been corrected since you purchased your copy.

NCSS maintains the following program and documentation copying policy. Copies are limited to a one person / one machine basis for "BACKUP" purposes only. You may make as many backup copies as you wish. Further distribution constitutes a violation of this copy agreement and will be prosecuted to the fullest extent of the law.

NCSS is not "copy protected." You may freely load the program onto your hard disk. We have avoided copy protection in order to make the system more convenient for you. Please honor our good faith (and low price) by avoiding the temptation to distribute copies to friends and associates.

We believe this to be an accurate, exciting, easy-to-use system. If you find any portion that you feel needs to be changed, please let us know. Also, we openly welcome suggestions for additions to the system.

User's Guide I
Table of Contents

Quick Start & Self Help

1 Installation and Basics
2 Creating / Loading a Database
3 Data Transformation
4 Running Descriptive Statistics
5 Running a Two-Sample T-Test
6 Running a Regression Analysis
7 Data Window
8 Procedure Window
9 Output Window
10 Filters
11 Writing Transformations
12 Importing Data
13 Value Labels
14 Database Subsets
15 Data Simulation
16 Cross Tabs on Summarized Data

Quick Start Index

Introduction

100 Installation
101 Tutorial
102 Databases
103 Spreadsheets
104 Merging Two Databases
105 Procedures
106 Output
107 Navigator and Quick Launch

Data

115 Importing Data
116 Exporting Data
117 Data Report
118 Data Screening
119 Transformations
120 If-Then Transformations
121 Filter
122 Data Simulator
123 Data Matching – Optimal and Greedy
124 Data Stratification

Tools

130 Macros
135 Probability Calculator

Graphics

Introduction

140 Introduction to Graphics

Single-Variable Charts

141 Bar Charts
142 Pie Charts
143 Histograms
144 Probability Plots

Two-Variable Charts (Discrete / Continuous)

150 Dot Plots
151 Histograms – Comparative
152 Box Plots
153 Percentile Plots
154 Violin Plots
155 Error-Bar Charts

Two-Variable Charts (Both Continuous)

160 Function Plots
161 Scatter Plots
162 Scatter Plot Matrix
163 Scatter Plot Matrix for Curve Fitting

Three-Variable Charts

170 3D Scatter Plots
171 3D Surface Plots
172 Contour Plots
173 Grid Plots

Settings Windows

180 Color Selection Window
181 Symbol Settings Window
182 Text Settings Window
183 Line Settings Window
184 Axis-Line Settings Window
185 Grid / Tick Settings Window
186 Tick Label Settings Window
187 Heat Map Settings Window

References and Indices

References
Chapter Index
Index

User's Guide II
Table of Contents

Descriptive Statistics

200 Descriptive Statistics
201 Descriptive Tables

Means

T-Tests

205 One-Sample or Paired
206 Two-Sample
207 Two-Sample (From Means and SD's)

Analysis of Variance

210 One-Way Analysis of Variance
211 Analysis of Variance for Balanced Data
212 General Linear Models (GLM)
213 Analysis of Two-Level Designs
214 Repeated Measures Analysis of Variance

Mixed Models

220 Mixed Models

Other

230 Circular Data Analysis
235 Cross-Over Analysis Using T-Tests
240 Nondetects Analysis

Quality Control

250 Xbar R (Variables) Charts
251 Attribute Charts
252 Levey-Jennings Charts
253 Pareto Charts
254 R & R Study

Design of Experiments

260 Two-Level Designs
261 Fractional Factorial Designs
262 Balanced Incomplete Block Designs
263 Latin Square Designs
264 Response Surface Designs
265 Screening Designs
266 Taguchi Designs
267 D-Optimal Designs
268 Design Generator

References and Indices

References
Chapter Index
Index

User's Guide III
Table of Contents

Regression

Linear and Multiple Regression

300 Linear Regression and Correlation
305 Multiple Regression
306 Multiple Regression with Serial Correlation

Variable Selection

310 Variable Selection for Multivariate Regression
311 Stepwise Regression
312 All Possible Regressions

Other Regression Routines

315 Nonlinear Regression
320 Logistic Regression
325 Poisson Regression
330 Response Surface Regression
335 Ridge Regression
340 Principal Components Regression
345 Nondetects Regression

Cox Regression is found in User's Guide V in the Survival/Reliability section

Curve Fitting

Curve Fitting

350 Introduction to Curve Fitting
351 Curve Fitting – General
360 Growth and Other Models
365 Piecewise Polynomial Models

Ratio of Polynomials

370 Search – One Variable
371 Search – Many Variables
375 Fit – One Variable
376 Fit – Many Variables

Other

380 Sum of Functions Models
385 User-Written Models
390 Area Under Curve

References and Indices

References
Chapter Index
Index

User's Guide IV
Table of Contents

Multivariate Analysis

400 Canonical Correlation
401 Correlation Matrix
402 Equality of Covariance
405 Hotelling's One-Sample T2
410 Hotelling's Two-Sample T2
415 Multivariate Analysis of Variance (MANOVA)
420 Factor Analysis
425 Principal Components Analysis
430 Correspondence Analysis
435 Multidimensional Scaling
440 Discriminant Analysis

Clustering

445 Hierarchical (Dendrograms)
446 K-Means
447 Medoid Partitioning
448 Fuzzy
449 Regression
450 Double Dendrograms

Meta-Analysis

455 Means
456 Proportions
457 Correlated Proportions
458 Hazard Ratios

Forecasting / Time Series

Exponential Smoothing

465 Horizontal
466 Trend
467 Trend & Seasonal

Time Series Analysis

468 Spectral Analysis
469 Decomposition Forecasting
470 The Box-Jenkins Method
471 ARIMA (Box-Jenkins)
472 Autocorrelations
473 Cross-Correlations
474 Automatic ARMA
475 Theoretical ARMA

Operations Research

480 Linear Programming

Mass Appraisal

485 Appraisal Ratios
486 Comparables – Sales Price
487 Hybrid Appraisal Models

References and Indices

References
Chapter Index
Index

User's Guide V
Table of Contents

Tabulation

500 Frequency Tables
501 Cross Tabulation

Item Analysis

505 Item Analysis
506 Item Response Analysis

Proportions

510 One Proportion
515 Two Independent Proportions
520 Two Correlated Proportions (McNemar)
525 Mantel-Haenszel Test
530 Loglinear Models

Diagnostic Tests

Binary Diagnostic Tests

535 Single Sample
536 Paired Samples
537 Two Independent Samples
538 Clustered Samples

ROC Curves

545 ROC Curves

Survival / Reliability

550 Distribution (Weibull) Fitting
551 Beta Distribution Fitting
552 Gamma Distribution Fitting
555 Kaplan-Meier Curves (Logrank Tests)
560 Cumulative Incidence
565 Cox Regression
566 Parametric Survival (Weibull) Regression
570 Life-Table Analysis
575 Probit Analysis
580 Time Calculator
585 Tolerance Intervals

References and Indices

References
Chapter Index
Index

Chapter 1

Installation and Basics

Before You Install

1. Check System Requirements

NCSS runs on 32-bit and 64-bit Windows systems. This includes Windows ME, Windows NT 4.0, Windows 2000, Windows XP, and Windows Vista. The recommended minimum system is a Windows XP or Vista-compatible PC.

NCSS takes up about 120 MB of disk space. Once installed, ***NCSS*** also requires about 20 MB of temporary disk space while it is running.

2. Find a Home for NCSS

Before you start installing, decide on a directory where you want to install ***NCSS***. By default, the setup program will install ***NCSS*** in the *C:\Program Files\NCSS\NCSS 2007* directory. You may change this during the installation, but not after. The example data, template, and macro files will be placed in your personal documents folder (usually *C:\…\[My] Documents\NCSS\NCSS 2007*) in appropriate subdirectories. The program will save all procedure templates and macros to these folders while the program is running.

3. If You Already Own NCSS

If ***NCSS*** is already installed on your system, instruct the installation program to place this new version in a new folder (e.g. *C:\Program Files\NCSS\NCSS 2007*). All appropriate files will be copied from your old ***NCSS*** directory or replaced by updated files.

What Install Does

The installation procedure creates the necessary folders and copies the ***NCSS*** program from the installation file, called *NCSS2007SETUP.EXE*, to those folders. The files in *NCSS2007SETUP.EXE* are compressed, so the installation program decompresses these files as it copies them to your hard disk.

The following folders are created during installation:

C:\Program Files\NCSS\NCSS 2007 (or your substitute folder) contains most of the program files.

C:\Program Files\NCSS\NCSS 2007\Pdf contains printable copies of the documentation in PDF format.

C:\Program Files\NCSS\NCSS 2007\Sts contains all labels, text, and online messages.

C:\…\[My] Documents\NCSS\NCSS 2007\Data contains the database files used by the tutorials. We recommend creating a sub-folder of this folder to contain the data for each project you work on. An empty subfolder called "My Data" is created within this folder for easy storage of your personal data files. You can save the data to any folder you wish.

C:\…\[My] Documents\NCSS\NCSS 2007\Junk contains temporary files used by the program while it is running. Under normal operation, ***NCSS*** will automatically delete temporary files. After finishing ***NCSS***, you can delete any files left in this folder (but not the folder itself).

C:\…\[My] Documents\NCSS\NCSS 2007\Macros contains saved macros.

C:\…\[My] Documents\NCSS\NCSS 2007\Report is the default folder in which to save your output. You can save the reports to any folder you wish.

C:\…\[My] Documents\NCSS\NCSS 2007\Settings contains the files used to store your template files. These files are used by the ***NCSS*** template system, which is described in a later chapter.

Installing NCSS

This section gives instructions for installing ***NCSS*** on your computer system. You must use the ***NCSS*** setup program to install ***NCSS***. The files are compressed, so you cannot simply copy the files to your hard drive.

Follow these basic steps to install ***NCSS*** on your computer system.

Step	Notes
1.	Make sure that you are using a 32- or 64-bit version of windows such as Windows Me, Windows NT 4.0, Windows 2000, Windows XP, or Windows Vista.
2.	If you are installing from a CD, insert the CD in the CD drive. The installation program should start automatically. If it does not, on the Start menu, select the Run command. Enter *D:\NCSS\NCSS2007Setup*. You may have to substitute the appropriate letter for your CD drive if it is not *D*. If you are installing from a download, simply run the downloaded file (*NCSS2007Setup.exe*).
3.	Once the setup starts, follow the instructions on the screen. ***NCSS*** will be installed in the drive and folder you designate.

If Something Goes Wrong During Installation

The installation procedure is automatic. If something goes wrong during installation, delete the *C:\Program Files\NCSS\NCSS 2007* directory and start the installation process at the beginning. If trouble persists, contact our technical support staff as indicated below.

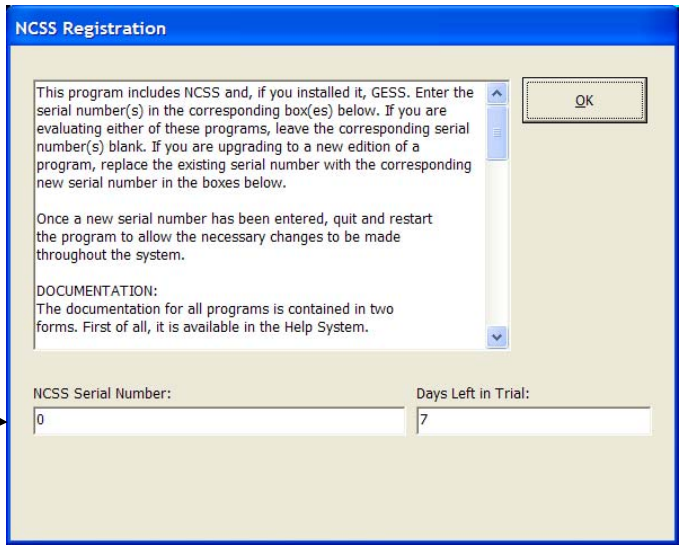
Starting NCSS

NCSS may be started using your keyboard or your mouse using the same techniques that you use to start any other Windows application. You can start ***NCSS*** by selecting **NCSS 2007** from your Start menu using standard mouse or keyboard operations.

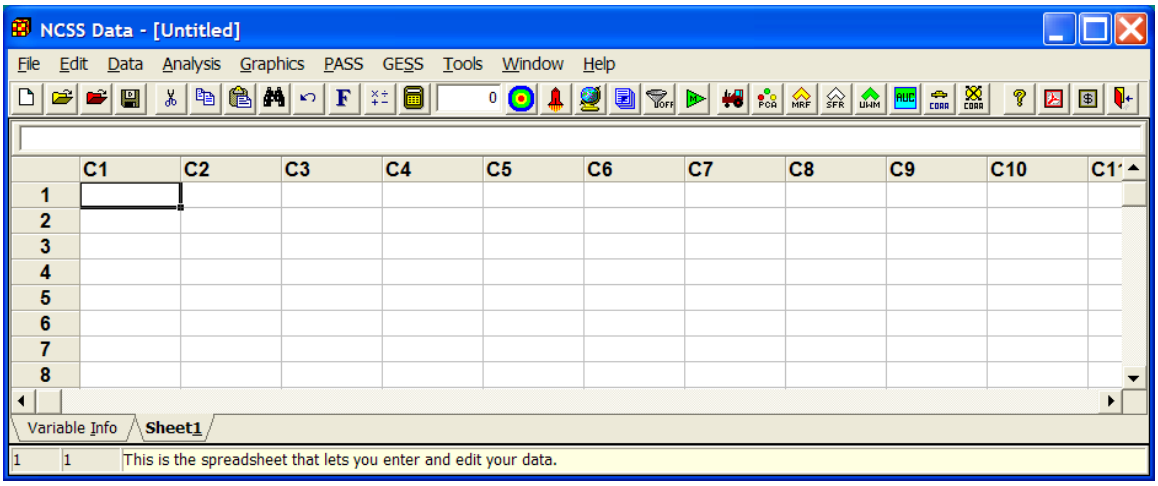
Entering Your Serial Number

The first time you run ***NCSS***, enter your serial number in the appropriate box in the pop-up window that appears. If you do not enter a serial number, ***NCSS*** will enter trial mode and you will have 7 days to evaluate ***NCSS***. When in trial mode, ***NCSS*** is fully-functional but is limited to 100 rows of data.

Enter your ***NCSS*** serial number here.



When you click **OK**, the **NCSS Data** window will appear. If you entered a serial number, you must quit and restart the program for the serial number to take effect.



The Three Main NCSS Windows

NCSS is controlled by three main windows:

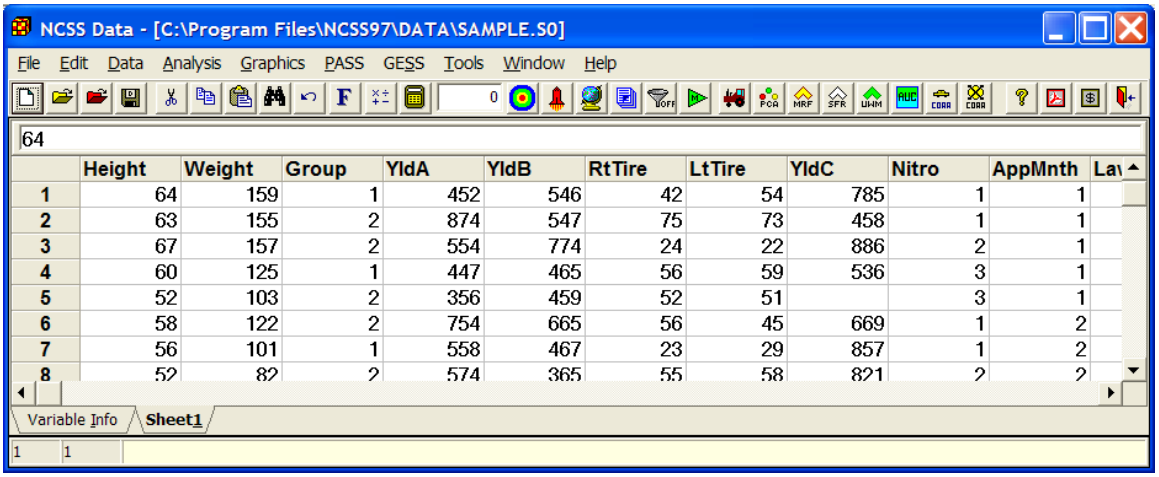
1. **Data Window**
2. **Procedure Window**
3. **Output Window**

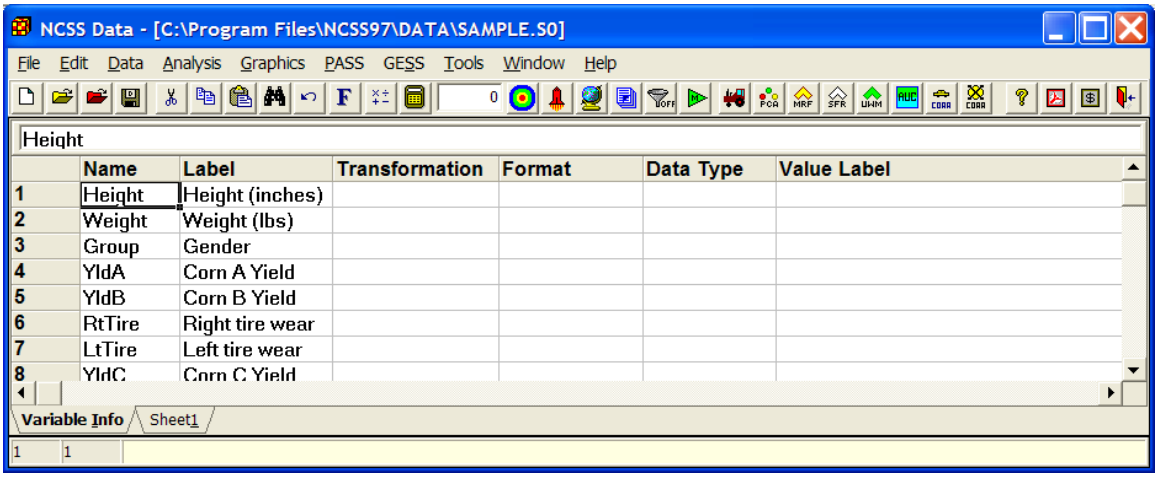
Each window has its own menu bar and tool bar. We will now briefly describe each of the three.

The NCSS Data Window

The **NCSS Data** window contains the data that is currently being analyzed. This window lets you view, modify, and save your data. It has the look and feel of a spreadsheet. This is the main *NCSS* window. Closing this window will exit *NCSS*.

Chapter 7 provides a closer look at the Data window.

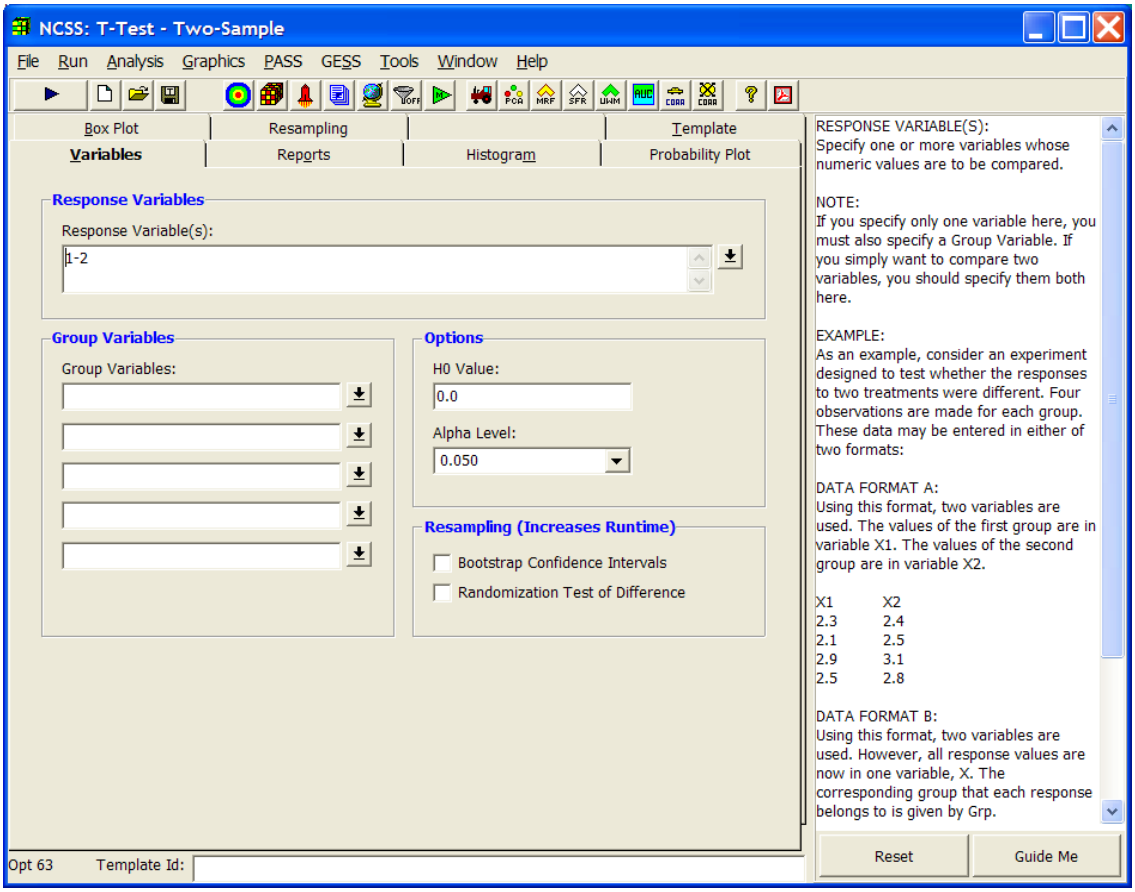




The NCSS Procedure Window

The **NCSS Procedure** windows let you set the options for a particular analysis. Whether you are running a multiple regression, an ANOVA, or a scatter plot, you will set the options of this procedure in the Procedure window. Closing this window will not exit ***NCSS***.

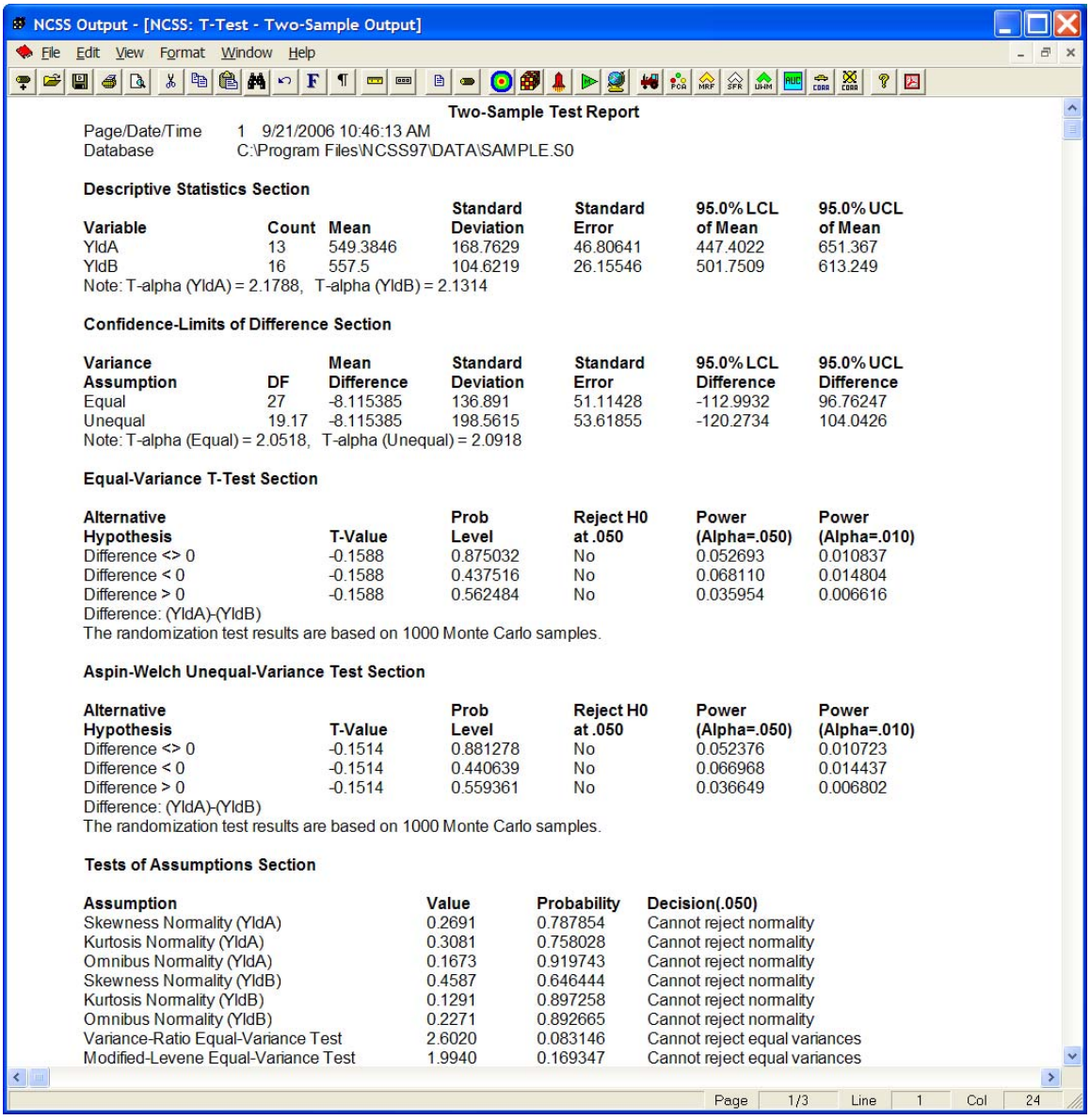
Chapter 8 provides a closer look at the Procedure window.



The NCSS Output Window

The **NCSS Output** window displays the output from the statistical and graphics procedures. It serves as a mini-word processor --- allowing you to view, edit, save, and print your output. Closing this window will not exit ***NCSS***.

Chapter 9 takes a closer look at the Output window.



Moving from Window to Window

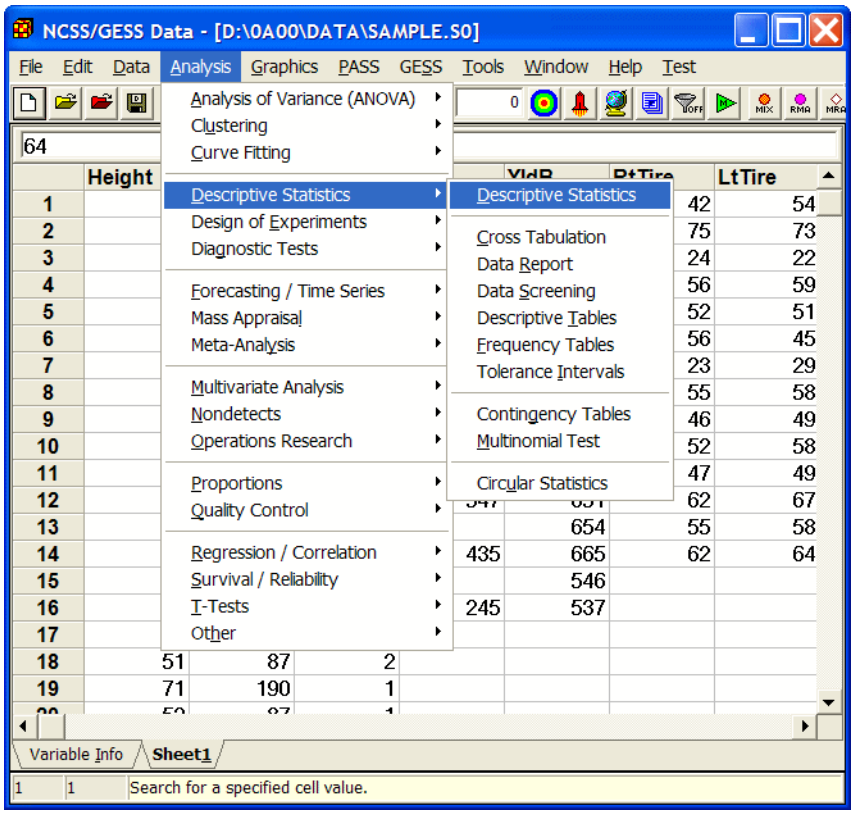
There are several ways of moving among the windows:

1. Remove the windows you are not currently using by minimizing them.
2. Arrange the windows on your screen so that all can be seen.
3. Use the task bar along the bottom of your screen.
4. Use the Windows menu.
5. Use the Navigator.
6. Use the toolbar (this is usually the quickest and easiest).

Selecting a Procedure

There are three primary ways to select a procedure:

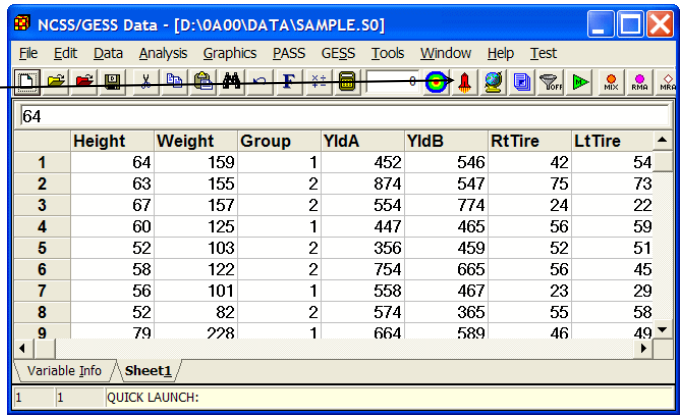
1. Select the procedure from the Analysis or Graphics menu



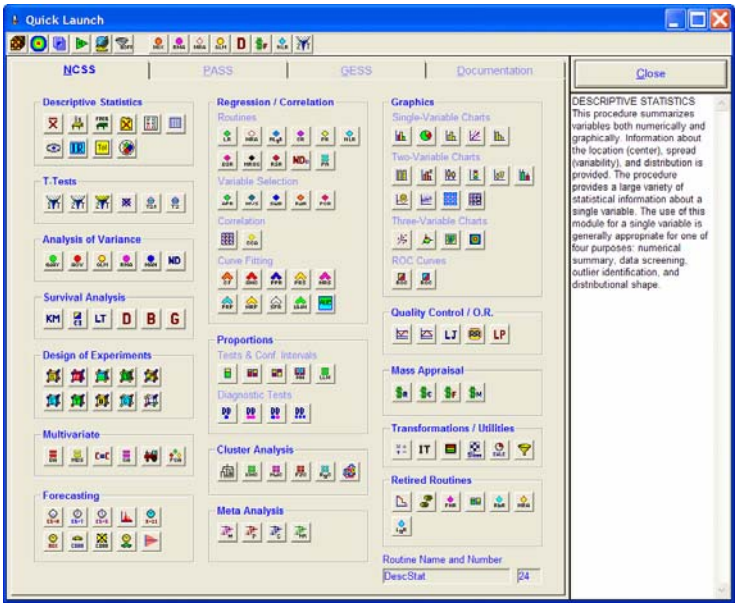
2. Click the button of the corresponding procedure from the Quick Launch buttons or from the quick access buttons of the tool bar

Click on the Quick Launch button from the tool bar.

The Quick Launch window will open.



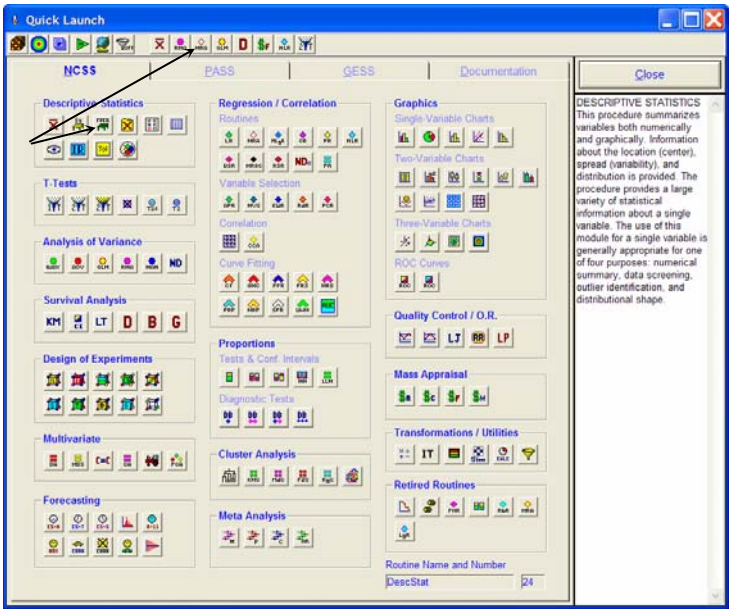
Find the desired procedure and click the corresponding button.



Buttons for the most commonly used procedures may be added to the tool bar by dragging and dropping them to the tool bar.

Eight procedure buttons are available on the tool bar.

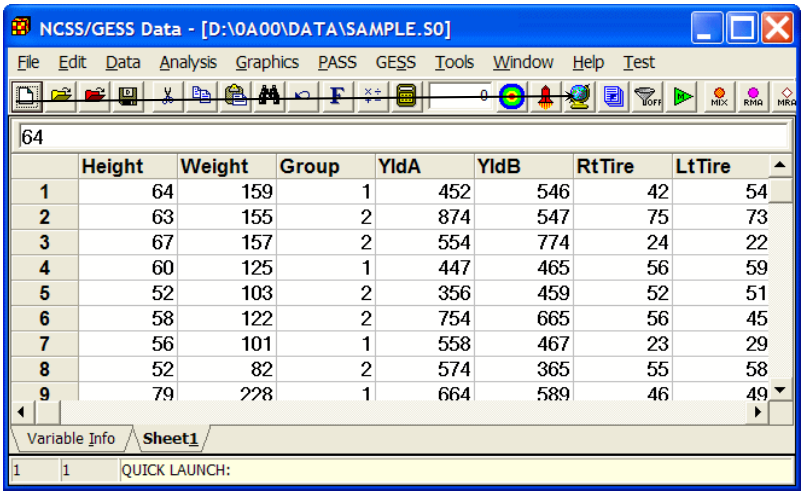
The procedure of the tool bar buttons may also be changed by right-clicking on one of the eight buttons of the tool bar and selecting the desired procedure.



3. Select the procedure using the Navigator

Click on the Navigator button from the tool bar.

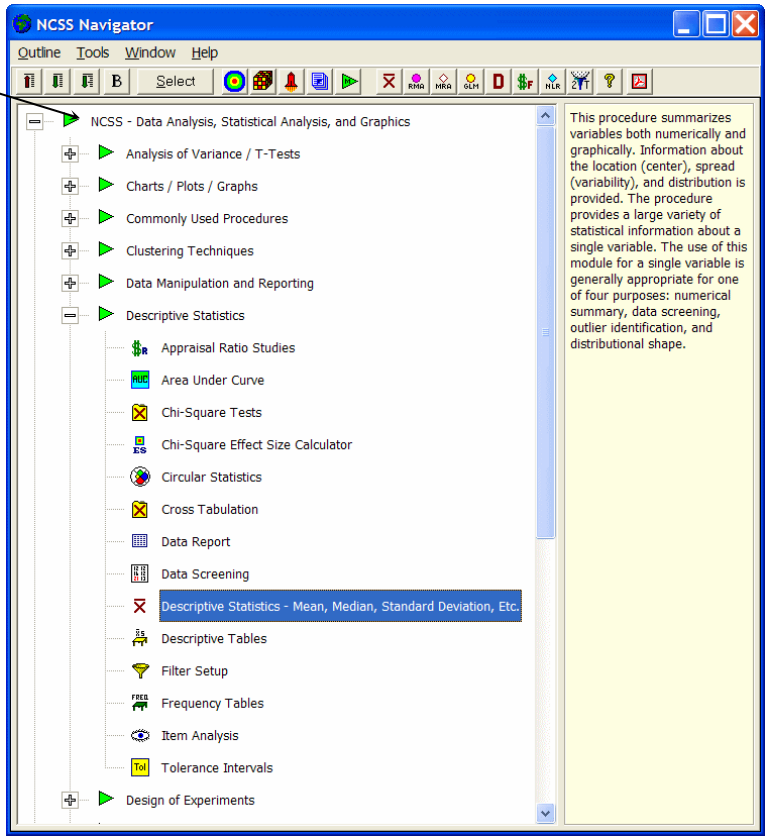
The Navigator window will open.



Click to the left of a category to open the category.

Continue until the desired procedure is reached.

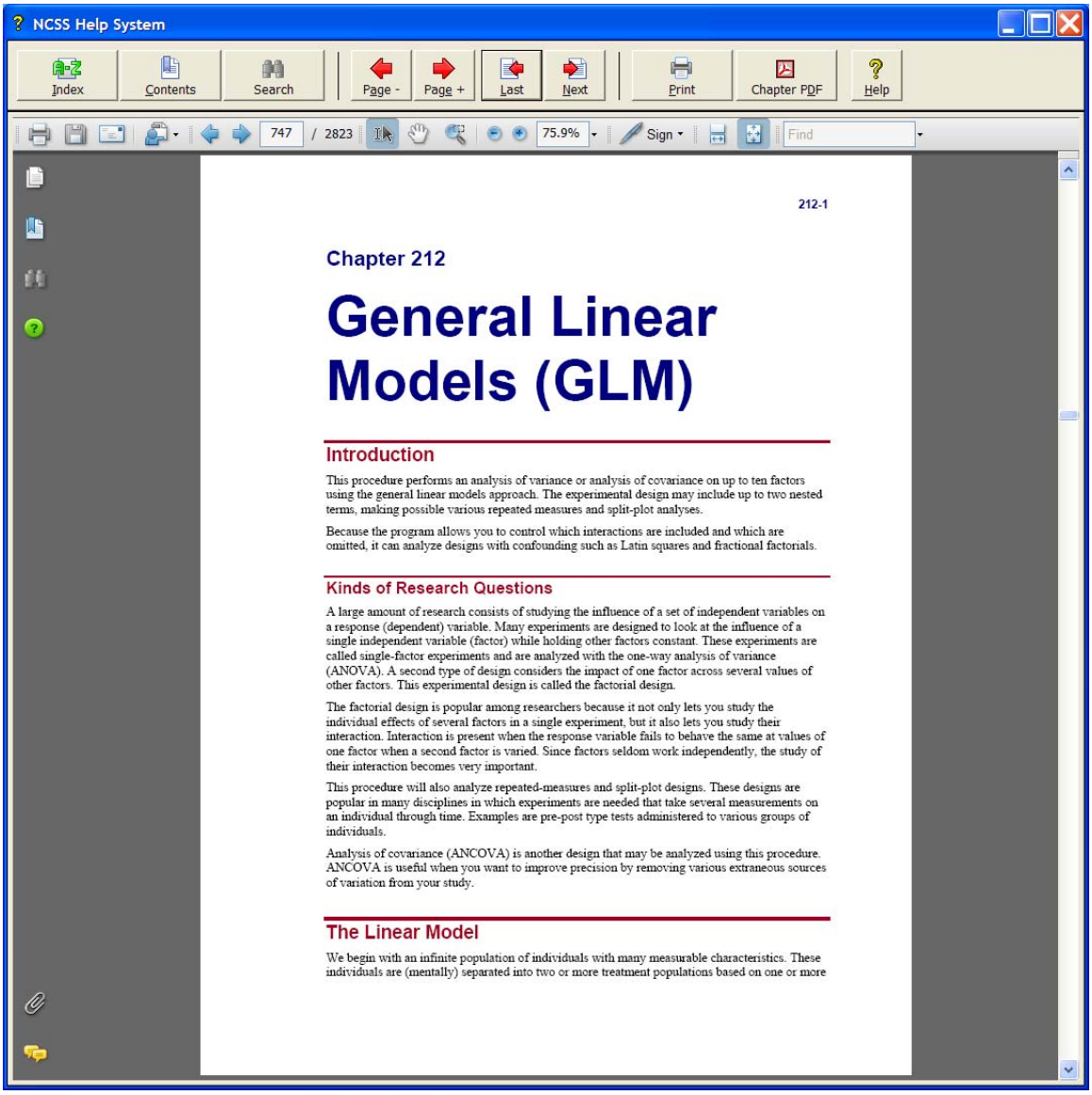
Double-click on the procedure name or symbol to open it.



Obtaining Help

Help System

To help you learn and use ***NCSS*** efficiently, the material in this manual is included in the ***NCSS*** Help System. The Help System is started from the Help menu or by clicking on the yellow '?' icon on the right side of the toolbar. ***NCSS*** updates, available for download at www.ncss.com, may contain adjustments or improvements of the ***NCSS*** Help System. Adobe Acrobat or Adobe Reader version 7 or 8 is required to view the help system. You can download Adobe Reader 8 for free by going to www.adobe.com. Adobe Reader 8 can also be installed from the *Utilities* folder on your ***NCSS*** installation CD.

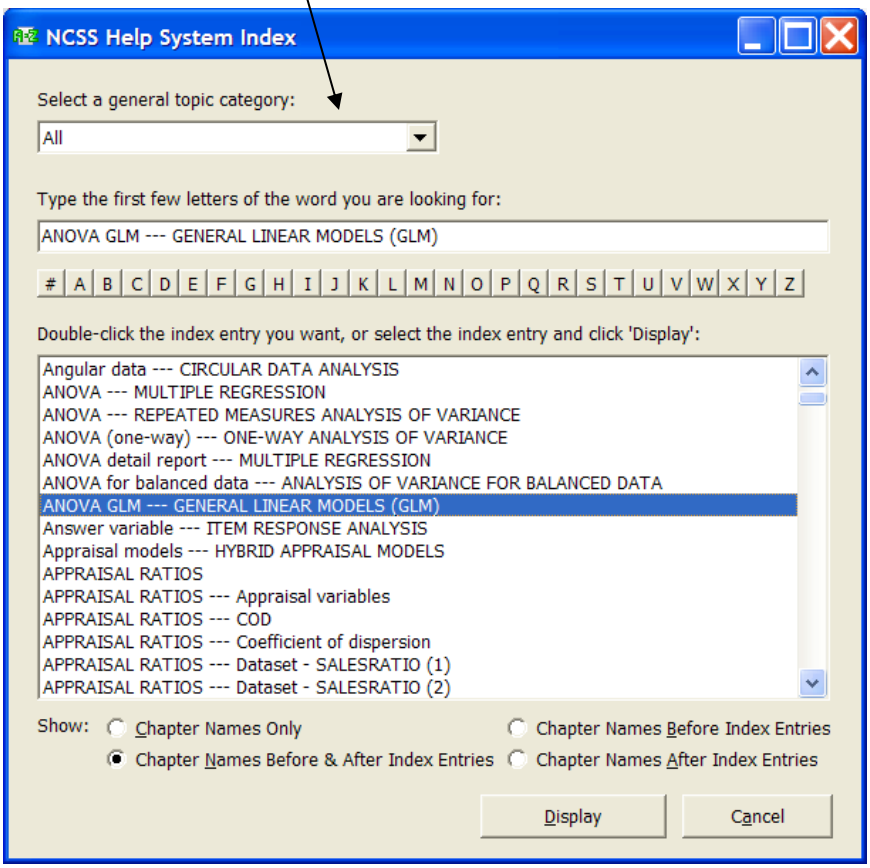


Navigating the Help System

There are a few key features of our help system that will let you use the help system more efficiently. We will now explain each of these features.

Index Window

The Index Window can be launched at any time by clicking on the Index button on the ***NCSS*** Help System display window. The index allows you to quickly locate keywords and/or statistical topics. You can narrow the list of index entries displayed by selecting a specific topic category in the uppermost dropdown box.



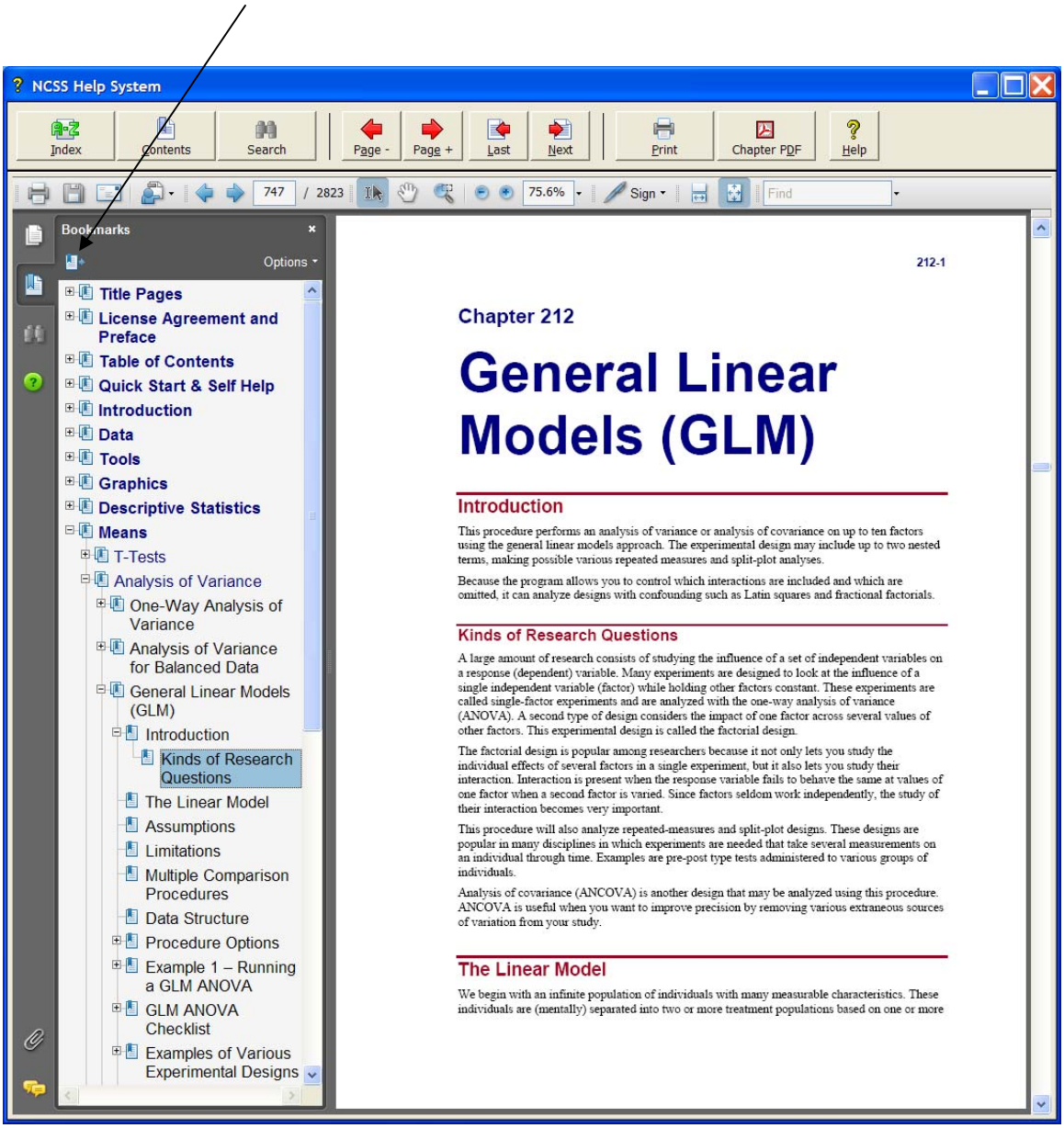
Index entries are displayed in the format

Index Entry --- CHAPTER NAME or CHAPTER NAME --- Index Entry.

You can control which entries are displayed by clicking on the radio buttons at the bottom of the window.

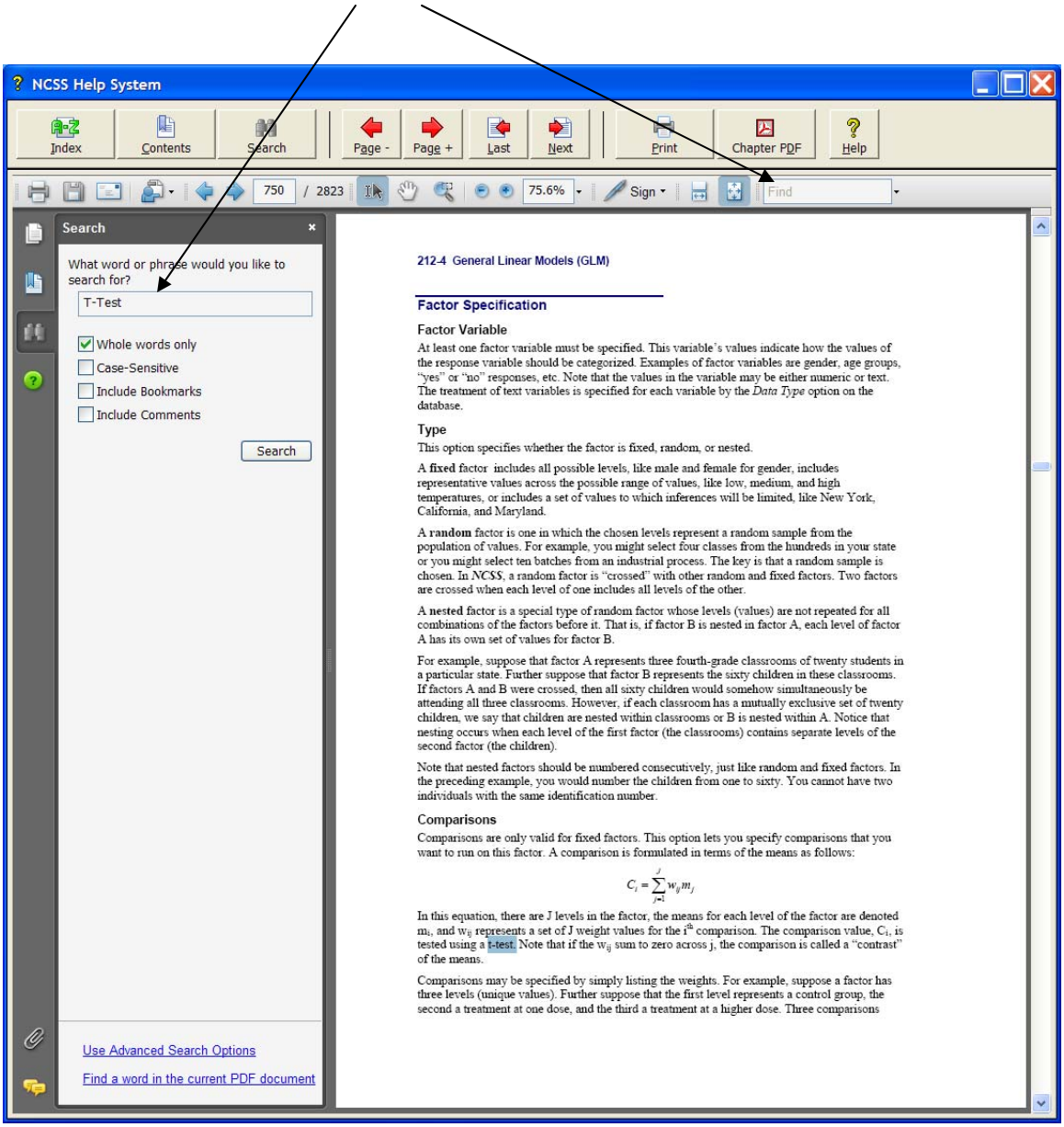
Contents Window

Clicking on the Contents button opens the Contents (Bookmarks) Window of the viewer. From this window you can expand the table of contents to view nested headings. You can click on the "Expand Current Bookmark" icon to instantly find the bookmark location for the currently-displayed page in the help document.



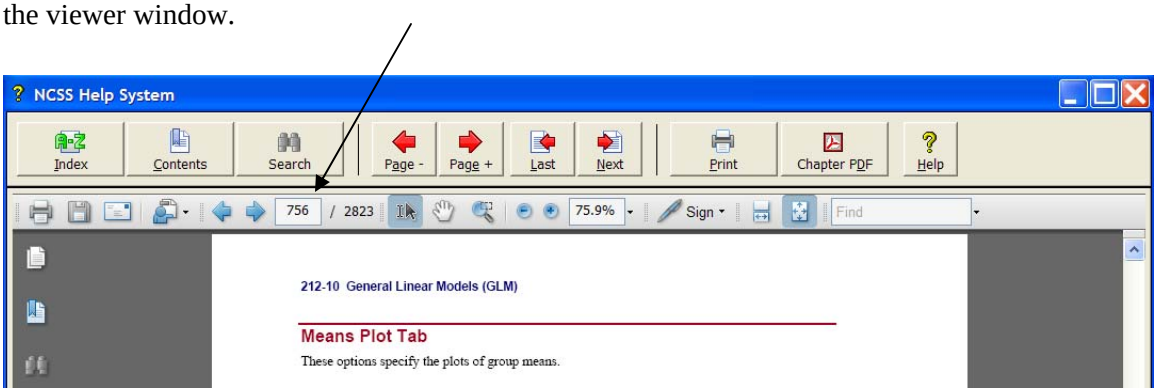
Search Window

Clicking on the Search button opens the Search Window of the viewer. From this window you can search the entire help system for any word or phrase. A search can also be initiated from the Find box in the viewer toolbar.



Printing the Documentation

To print pages from the documentation, click on the **Print** button on the ***NCSS*** Help System toolbar. This will launch the Adobe Reader print dialogue screen. You can choose to print a single page or a range of pages from the help file. When entering page numbers, remember to use the PDF file page numbers (e.g., 750-756) and not the page numbers found in the document pages (e.g., 212-4 to 212-10 is not a valid page range). The Adobe Reader page numbers can be seen in the viewer window.

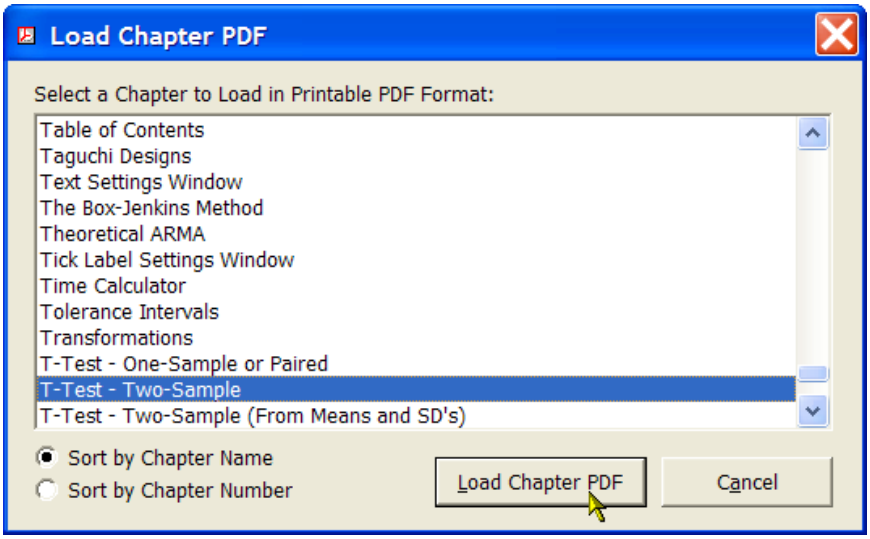


If you are using Adobe Reader 7, then the page numbers are found at the bottom of the viewer window.

One of the benefits of the ***NCSS*** Help System is the ease with which you can print any chapter or topic from the electronic help manual. To print a single chapter or topic using your default PDF viewer, take the following steps:

1. Click on the **Chapter PDF** icon in the ***NCSS*** Help System toolbar.

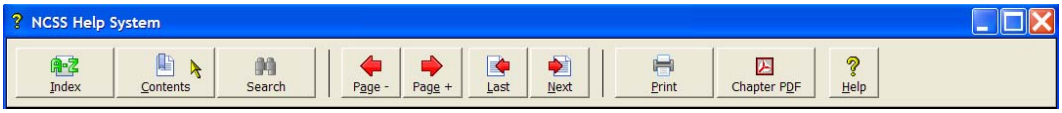
2. Choose the chapter you would like to print from the list and click **Load Chapter PDF**. This will launch the individual chapter PDF in a separate window using your default PDF viewer (e.g., Adobe Reader).



3. Use the **Print** function of your PDF viewer to print the entire chapter or individual pages from the chapter.

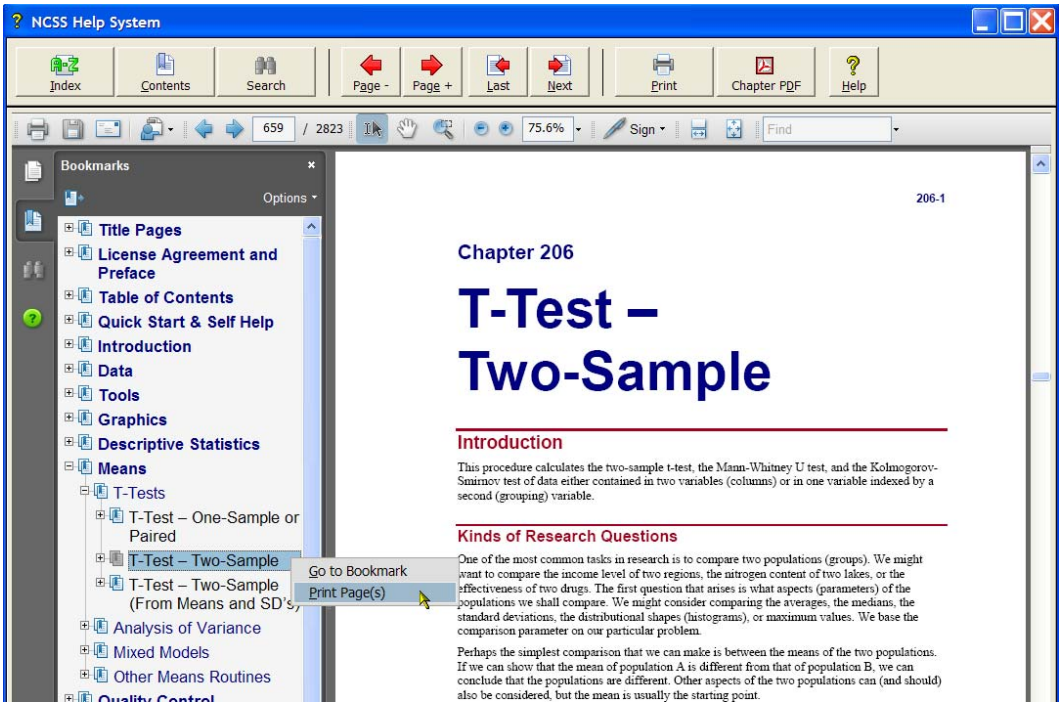
If you have Adobe Reader 8 or later, you can print entire chapters using an alternative method as follows (**This will not work with Adobe Reader 7**):

1. Open the Contents (Bookmarks) Window by clicking on the **Contents button** at the top of the ***NCSS*** Help System display window.



2. Expand the bookmarks to display the chapter or topic name you wish to print (e.g., the Two-Sample T-Test Chapter). Then, **highlight** the chapter name, **right-click** on the highlighted selection (or select Options in the panel above), and select **Print Page(s)**. This will automatically print only the pages from the selected chapter.

   CAUTION: When you click Print Page(s), the command is sent to the printer automatically without any intermediate Print Setup window being displayed. Make sure that you have selected only the topic you want before clicking Print Page(s).



If you do not want to print the entire chapter, continue to expand the bookmark tree to the topic you wish to print before completing step 2. The Print Page(s) command prints all pages containing bookmarks that are nested within the highlighted bookmark.

Technical Support

If you have a question about ***NCSS***, you should first look to the printed documentation and the included Help system. If you cannot find the answer there, look for help on the web at www.ncss.com/support.html. If you are unable to find the answer to your question by these means, contact ***NCSS*** technical support for assistance by calling (801) 546-0445 between 8 a.m. and 5 p.m. (MST). You can contact us by email at support@ncss.com or by fax at (801) 546-3907. Our technical support staff will help you with your question.

If you encounter problems or errors while using ***NCSS***, please view our list of recent corrections before calling by going to www.ncss.com/release_notes.html to find out if you problem or error has been corrected by an update. You can download updates anytime by going to http://www.ncss.com/download.html. If updating your software does not correct the problem, contact us by phone or email.

To help us answer your questions more accurately, we may need to know about your computer system. Please have pertinent information about your computer and operating system available.

Chapter 2

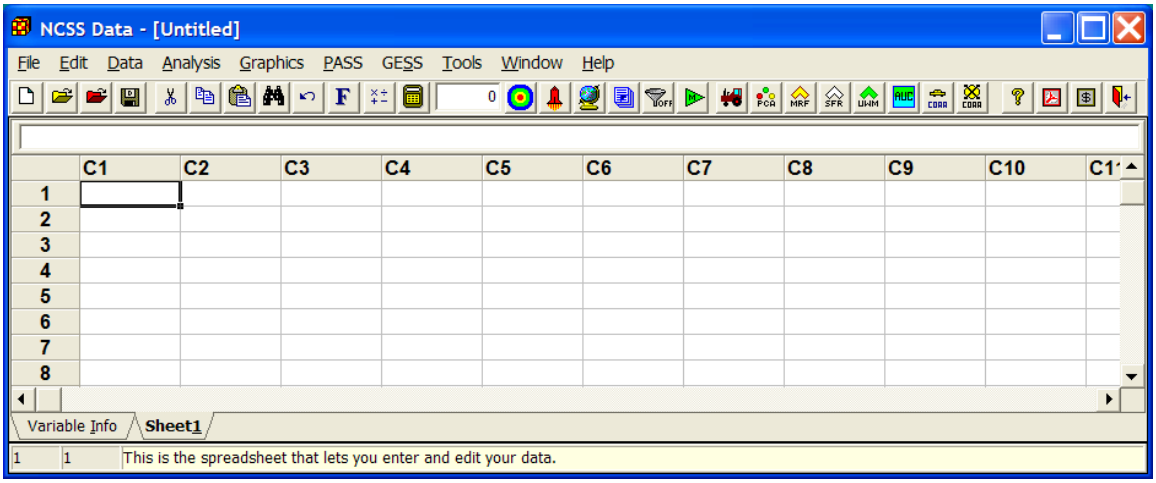
Creating / Loading a Database

About This Chapter (Time: 21 minutes)

This chapter will show you how to enter, label, save, and print a database. You will find complete coverage of these subjects in *NCSS User's Guide I*.

Starting NCSS

Select **All Programs** > **NCSS** from the Windows Start menu to start ***NCSS***. After starting, the ***NCSS*** spreadsheet program appears with an empty datasheet. We will begin by entering data into this datasheet.



Brain Weight Data

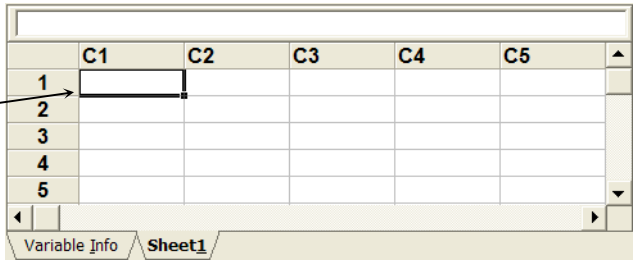
The following data give the body weight in kilograms and the brain weight in grams of various mammals. This chapter will show how to enter these data into an ***NCSS*** database and perform basic database operations such as saving and printing.

Mammal Name	Body Weight	Brain Weight
African Elephant	6654	5712
Asian Elephant	2547	4603
Giraffe	529	680
Horse	521	655
Cow	465	423
Gorilla	207	406
Pig	192	180
Jaguar	100	157
Man	62	1320
Chimpanzee	52	440
Gray Wolf	36	120
Kangaroo	35	56
Baboon	11	179
Red Fox	4	50
Cat	3	26

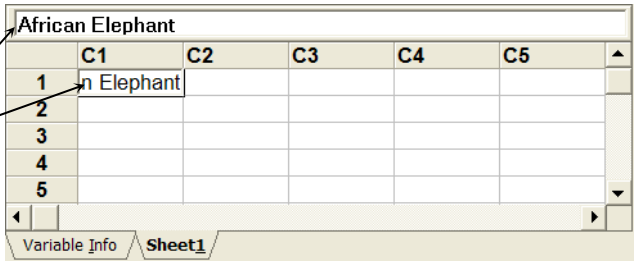
Entering a Column of Data

Take the following steps to enter the brain weight data into ***NCSS***:

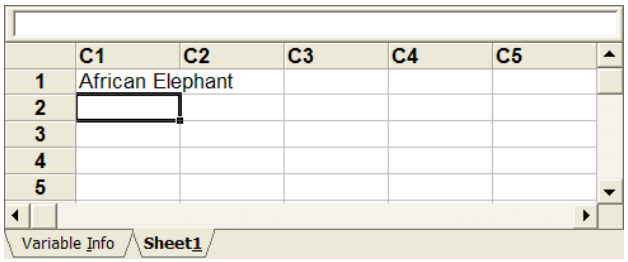
1. Position the cursor in the upper-left cell. This is done by clicking in the cell just below the title **C1**.



2. Type **African Elephant**. Notice that as you type, the characters appear in two places: in the cell where you are typing and in the cell-edit box at the top of the sheet.

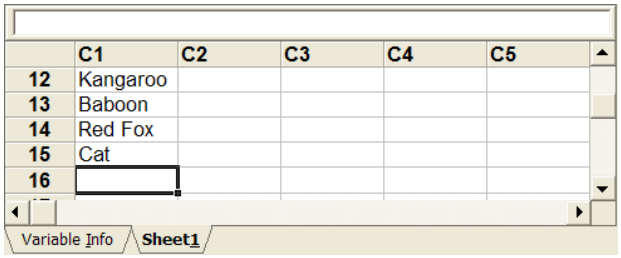


3. Press **Enter**. The text is entered and the cell cursor (the dark border around the cell) moves down to the next cell.

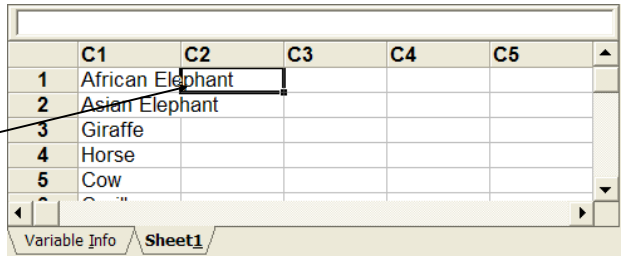


4. Type **Asian Elephant**.
Press **Enter**.
Type **Giraffe**.

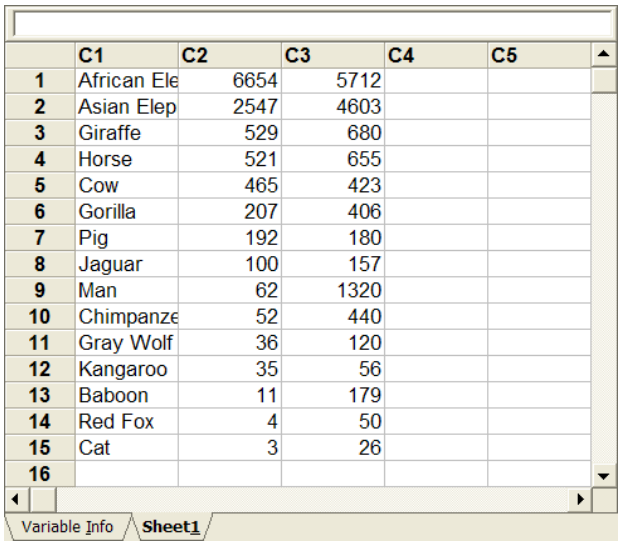
Continue until you finish entering all of the names.



5. Using the vertical scroll bar or the **Page Up** key, reposition the screen so that you can begin entering data in the second column. Click in the first row of column two. This will highlight this cell.



6. Type in the second and third columns of numbers. The completed table should appear as shown.



	C1	C2	C3	C4	C5
1	African Ele	6654	5712		
2	Asian Elep	2547	4603		
3	Giraffe	529	680		
4	Horse	521	655		
5	Cow	465	423		
6	Gorilla	207	406		
7	Pig	192	180		
8	Jaguar	100	157		
9	Man	62	1320		
10	Chimpanze	52	440		
11	Gray Wolf	36	120		
12	Kangaroo	35	56		
13	Baboon	11	179		
14	Red Fox	4	50		
15	Cat	3	26		
16					

Variable Info Sheet1

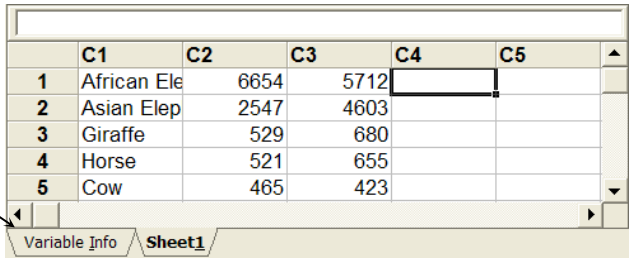
To cancel an entry, you can press the **Esc** key. If you have already pressed Enter, you can choose Undo from the Edit menu.

Labeling a Variable

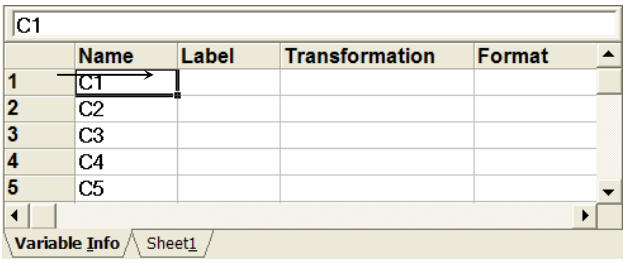
In ***NCSS***, a column of data is called a variable. Each variable has a column number and a name. The name is the label at the top of the column. The name of the variable will be displayed in all statistical reports and graphs that you generate, so it is important to name variables so that they will be remembered.

In a new database, the variables receive the default names C1, C2, C3, etc. Hence, you have just entered data into variables C1, C2, and C3. We will now show you how to change the names of these variables.

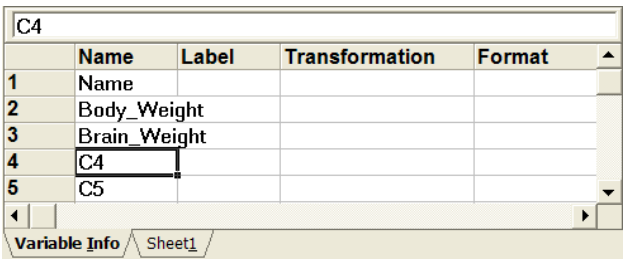
1. Click on the **Variable Info** tab.



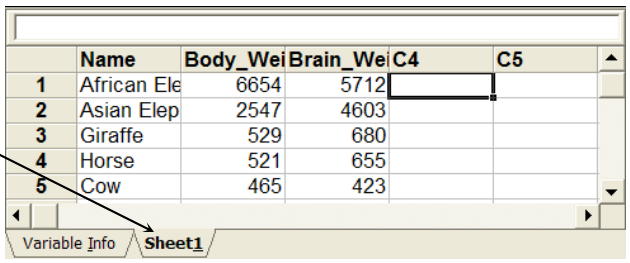
2. Click in the **C1** cell. This will position the cell cursor in that cell. (The cell cursor may already be there.)



3. Type **Name**.
Press **Enter**.
Type **Body_Weight**.
(Use the underscore, not the minus sign in these names.)
Press **Enter**.
Type **Brain_Weight**.
Press **Enter**.

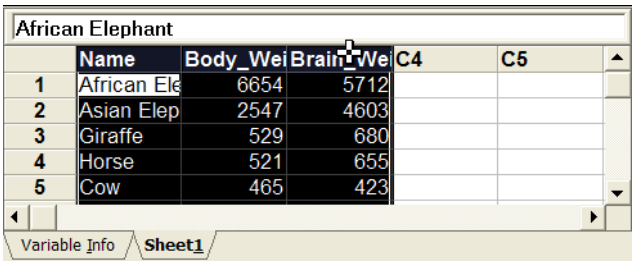


4. Click on the **Sheet1** tab. This will return you to a view of the data. The screen should appear like this.

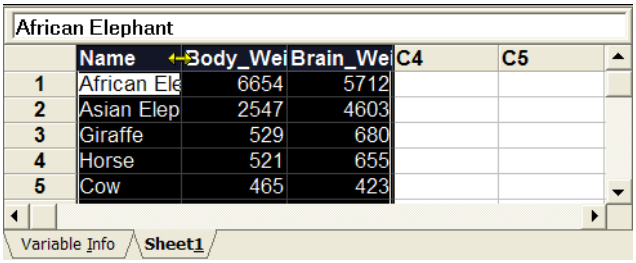


The final step is to widen the columns so that the complete names and labels are shown.

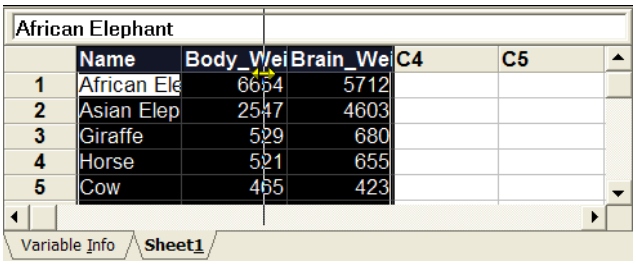
5. Drag the mouse from the **Name** heading to the **Brain_Weight** heading. This is done by pressing the left mouse on the heading **Name** and, without letting up, moving the mouse pointer to the heading **Brain_Weight** and finally letting up on the mouse. All three columns (headings and data) will be darkened.



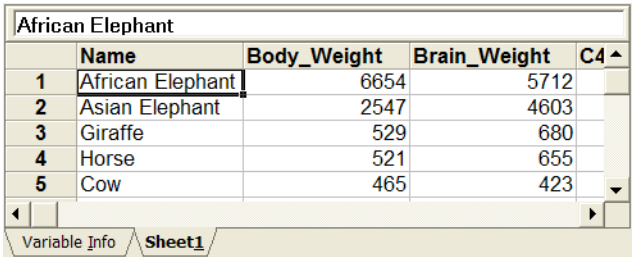
6. Now, position the mouse between the two columns. The mouse pointer will change to a two directional arrow.



7. Drag the mouse to the right until you are almost to the next border and let go of the mouse button. The columns will be widened, showing the complete variable names (column headings) and animal names.



8. Click on a cell in the body of the table to cancel the selection (the reverse video).

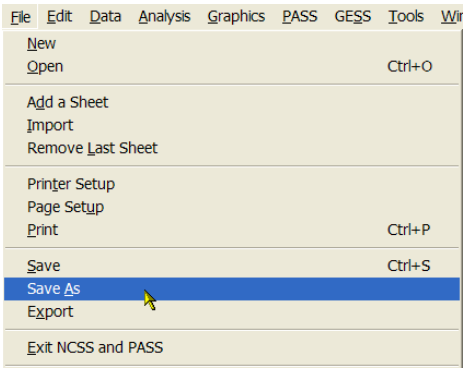


Variable names are used throughout the program to identify which columns of data to analyze. A variable name must begin with a letter (not a number); should contain only letters, numbers, and the underscore; and should not contain blanks. For correct formatting on reports, variable names should be less than fourteen characters, although there is no maximum length.

Saving Your Database

As you enter data, it is stored in your computer's temporary memory but not on your hard disk. If the computer loses power, you lose your data. We will now show you how to save the data to your hard disk.

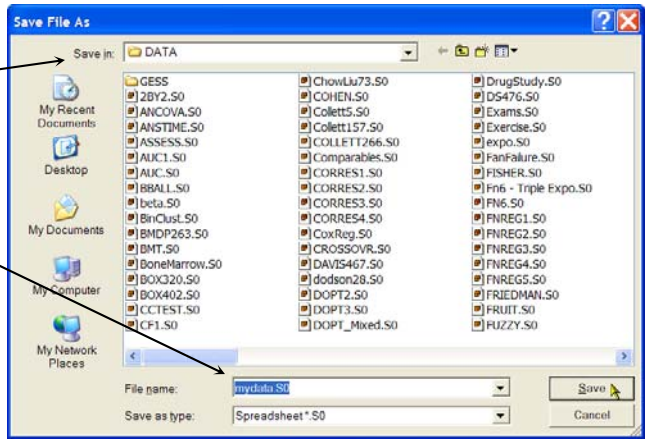
1. Select ***Save As*** from the File menu of the Data window.



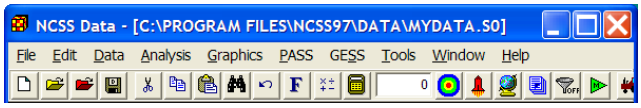
2. Select the **DATA** directory in your NCSS directory.

3. Enter **mydata.s0** in the File Name box.

4. Click **Save** button.



The database is stored as two files on your hard disk. If you use Windows Explorer to view the Data directory, you will find that you have created two files: **mydata.s0** and **mydata.s1**. The name of the database is now displayed at the top of the data window.

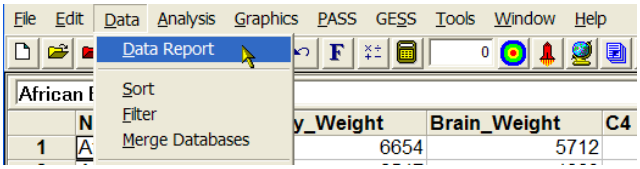


An ***NCSS*** database name must end with the file extension ".s0". Hence, a valid file name would have numbers, spaces, and letters followed by the extension ".s0". For example, you might use "abc.s0".

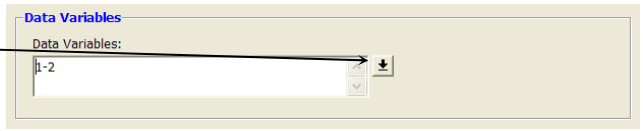
Printing Your Database

You will often want to create a printout of the data you have entered. We will now show you how this is accomplished.

1. Select **Data Report** from the Data menu. The Data Report procedure appears. This window allows you to control the format of your report.

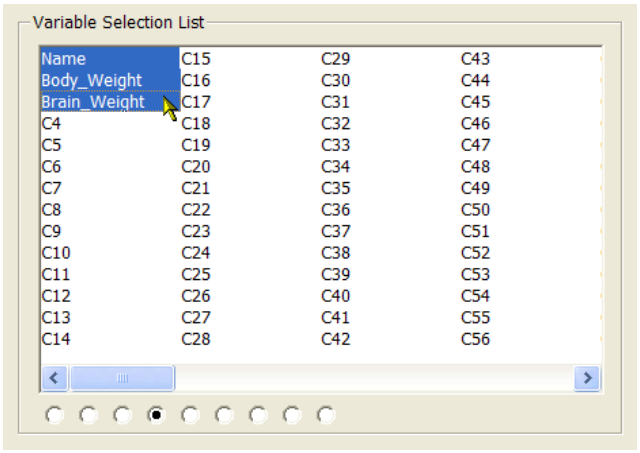


2. Click on the button to the right of the Data Variables line.

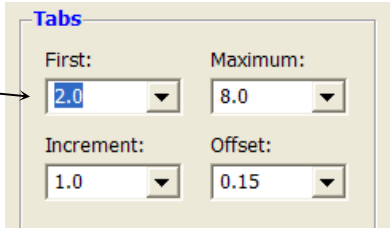


3. Select the first three variable names: **Name, Body_Weight,** and **Brain_Weight.**
Press **Ok**.

These variable names will appear in the Data Variables box.



4. Enter **2.0** in the **First** box of the **Tabs** section.



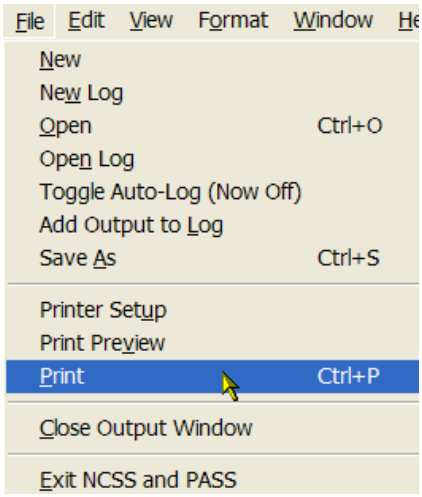
5. Press the **Run** button on the left of the toolbar at the top of the window.

The report will appear as shown below.

Data List Section

Row	Name	Body_Weight	Brain_Weight
1	African Elephant	6654	5712
2	Asian Elephant	2547	4603
3	Giraffe	529	680
4	Horse	521	655
5	Cow	465	423
6	Gorilla	207	406
7	Pig	192	180
8	Jaguar	100	157
9	Man	62	1320
10	Chimpanzee	52	440
11	Gray Wolf	36	120
12	Kangaroo	35	56
13	Baboon	11	179
14	Red Fox	4	50
15	Cat	3	26

6. Finally, select **Print** from the File menu of the Output window. This will display the Print dialog box from which you can print the report. Or, highlight the text using the mouse, and cut and paste the report directly into a word processor or slide show.

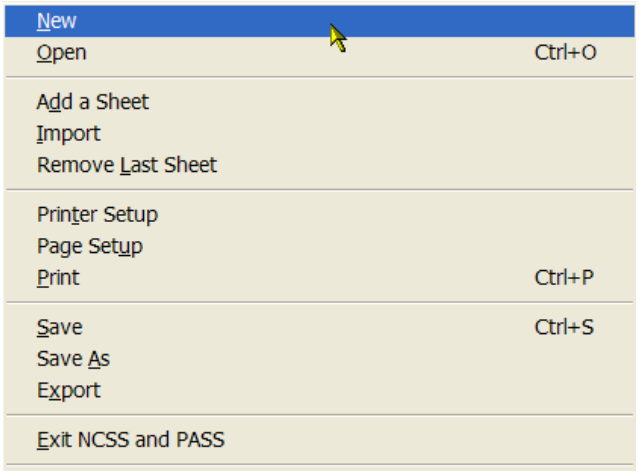


Congratulations! You have successfully entered and printed a set of statistical data. Analyzing these data using the various statistical procedures will not be much more difficult.

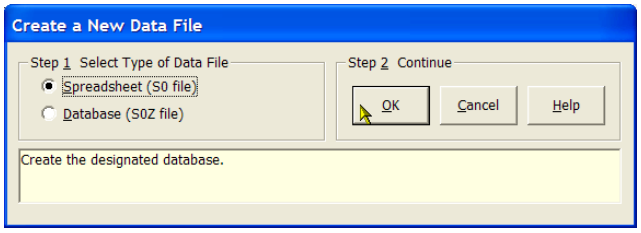
Clearing the Database

As you move from analysis to analysis, you will often have to clear the data screen so that new data may be entered. This is done as follows. (Of course, you should save your data before clearing it!)

1. Select **New** from the File menu of the Data window. Use the Windows menu to transfer from the Output window to the Data window. If you have not previously saved your data, choosing New will cause the program to ask you if you want to save the current datasheet before it is cleared.



2. Click **OK.** This will clear the screen and present you with an empty file just like when you start the program.

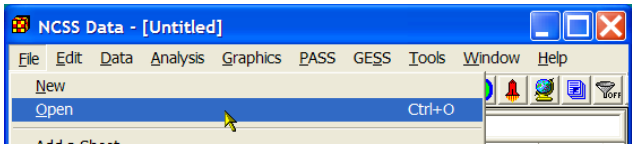


NCSS maintains two data formats. The spreadsheet (S0-type) format is for routine databases of fewer than 16,384 rows and 256 columns. The database (S0Z-type) format is for databases with more than 16,384 rows and/or 256 columns.

Loading a Database

Take the following steps to load the brain weight data into ***NCSS***:

1. Select **Open** from the File menu of the Data window. The File Open window will appear.

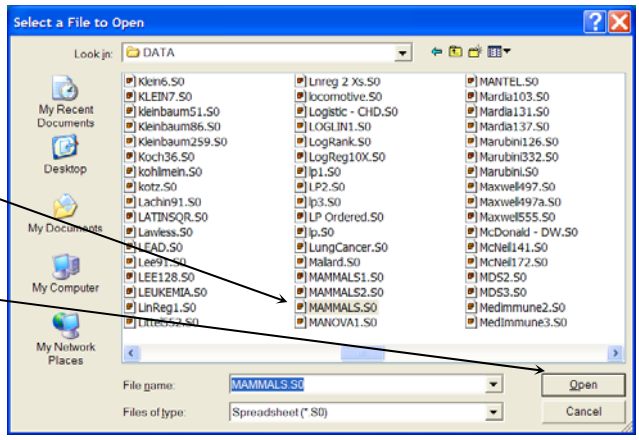


2. Double click the **Data** subdirectory to select it.

3. Double click **MAMMALS.S0** in the list of available files.

4. Click the ***Open*** button.

This will load the MAMMALS database into the Data window.



Chapter 3

Data Transformation

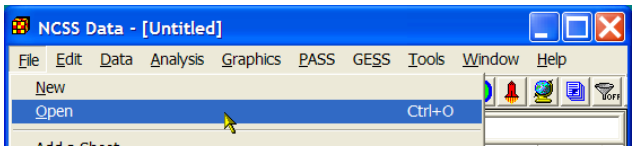
About This Chapter (Time: 13 minutes)

This chapter continues the introduction to the ***NCSS*** system by taking you through examples of using transformations to create new variables. Specifically, you will be shown how to calculate percentages and how to recode the values of a variable.

Loading the MAMMALS Database

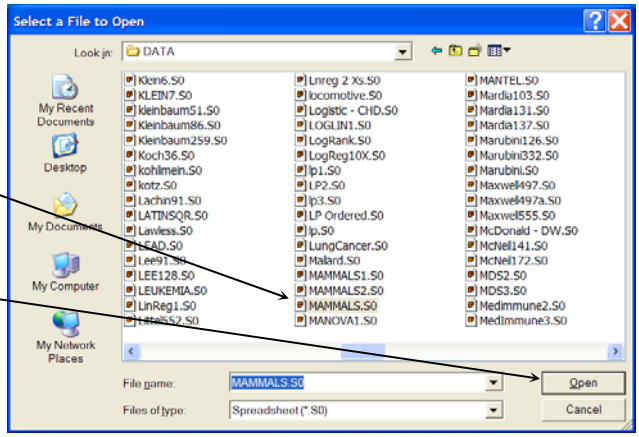
Take the following steps to load the brain weight data into ***NCSS***:

1. Select **Open** from the File menu of the Data window. The File Open window will appear.



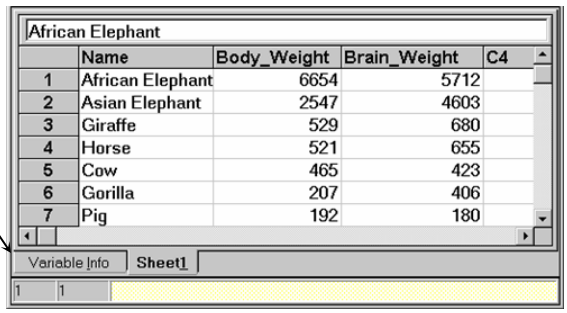
2. Double click the **Data** subdirectory to select it.
3. Double click **MAMMALS.S0** in the list of available files.
4. Click the ***Open*** button.

   This will load the MAMMALS database into the Data window.

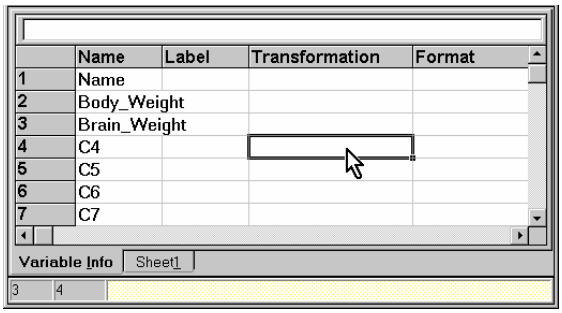


Creating a Percentage Variable

1. Click on the **Variable Info** tab. This will position you in the Variable Info datasheet.



2. In the **Transformation** column, click on the fourth cell down--the one in the **C4** row. This will position the spreadsheet cursor in this cell. This is where the transformation will be entered.

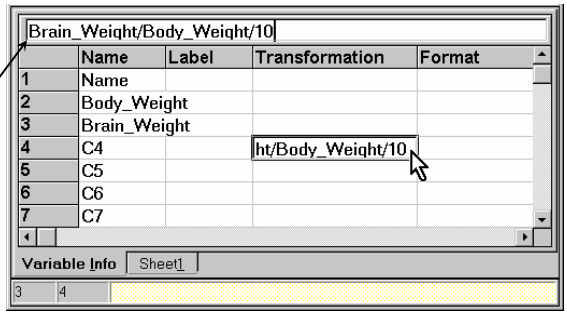


3. Type in the transformation expression: **Brain_Weight/Body_Weight/10**. (Be sure to type the underscores!)

Press **Enter**.

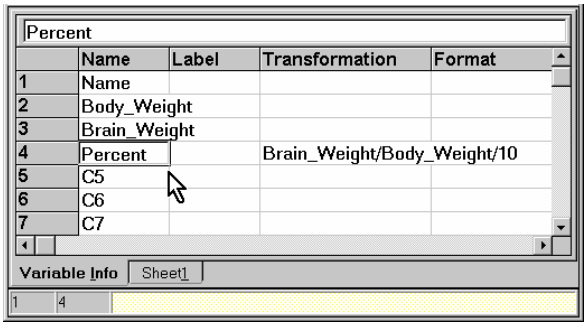
Notice that you edit the transformation in the edit bar at the top of the spreadsheet.

This step enters the new transformation expression, but does not change the data. The data are not generated until the spreadsheet is manually recalculated.



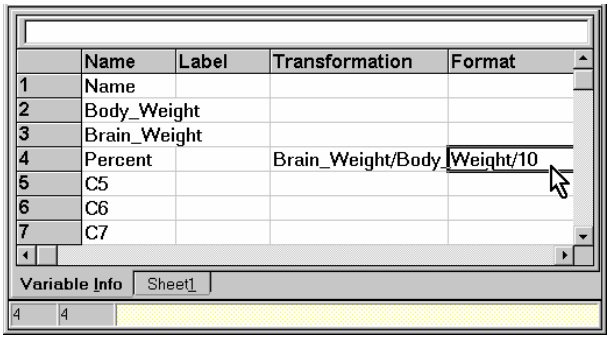
4. Click on **C4** in the Name column and type **Percent** and press **Enter**.

This renames the variable from the default of C4 to a new value that better describes the data in this column.



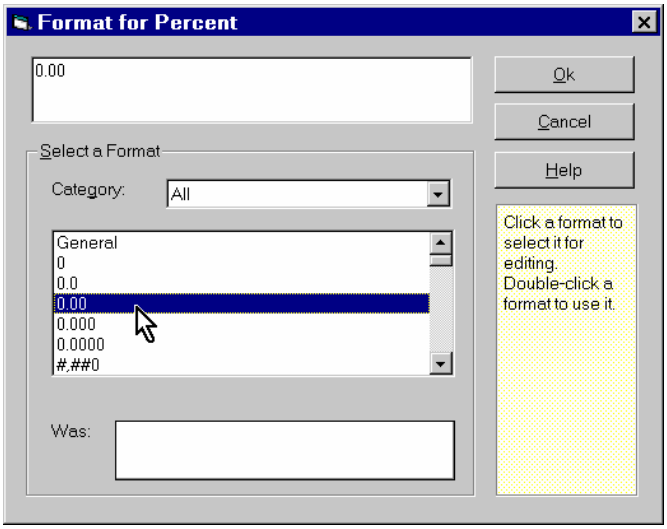
5. Double click on the cell in the fourth row in the **Format** column.

This will display the Format window for editing the format of the Percent variable. Note that this format will not influence the internal precision of the data.

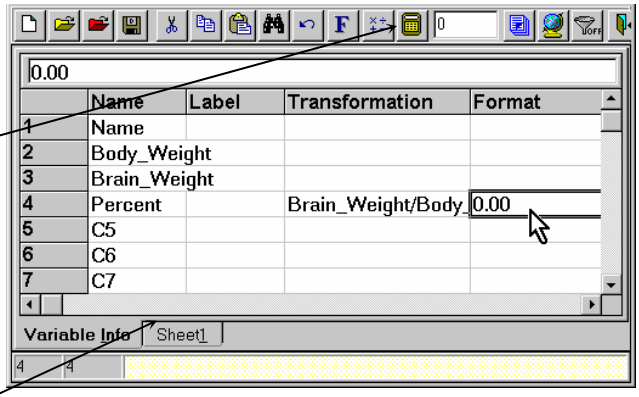


6. Select the two-decimal format: 0.00 from the list.

Although it is not necessary to reformat the numbers, it will make
viewing them much easier.



The completed screen will appear like this.



7. Click the **Apply Transformation** button on the toolbar.

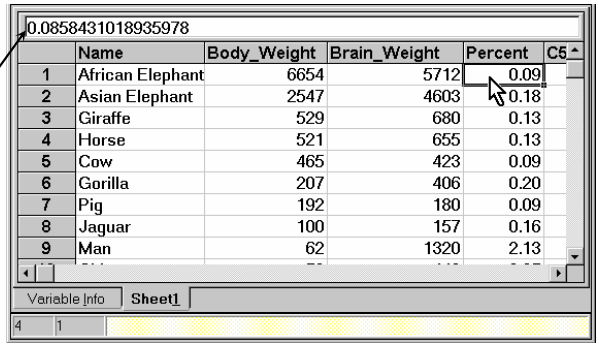
This will cause all transformations
to be recalculated.

8. Click the **Sheet1** tab.

The final result appears like this.

Notice the new column of data in the Percent variable's column.

By clicking on a cell, you can see that the data is actually stored in double precision.



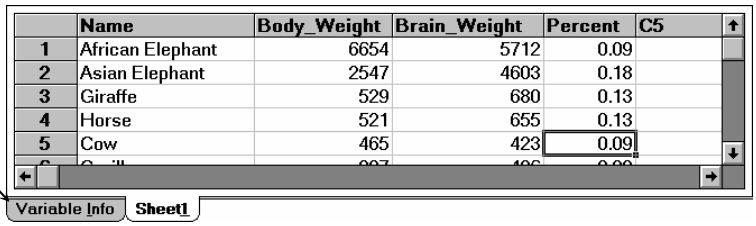
If you change or add data to either ***Body_Weight*** or ***Brain_Weight***, the ***Percent*** variable's values will not be automatically recalculated. You must recalculate the database using the **Apply Transformation** button or the **Recalc All** option of the Data menu.

Also remember that these changes are not automatically saved on your hard disk. If you want a permanent copy of a database with new transformations, you must save this modified version of the database using the Save option of the File menu in the Data window.

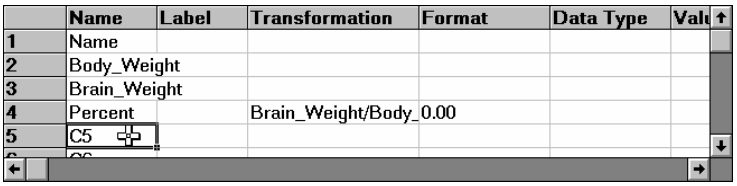
Recoding a Variable

It is often necessary to recode the values of a variable. As an example, we will recode the body weights to form a new variable as follows. Animals with a body weight less than 100 kg will receive a value of 1. Animals with a body weight greater than or equal to 100 kg will receive a value of 2. The transformation formula that will accomplish this is ***(Body_Weight>=100)+1***. The expression inside the parentheses results in a "1" if it is true or "0" if it is false. We will call the new variable ***SizeGroup***.

1. Click the **Variable Info** tab.

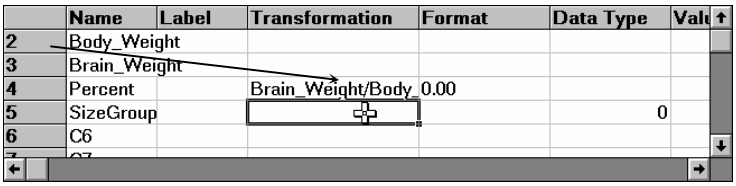


2. Click the **C5** name.

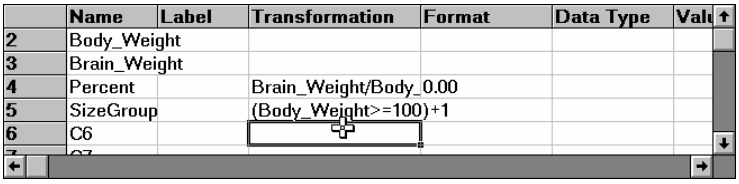


3. Type **SizeGroup** and press **Enter**.

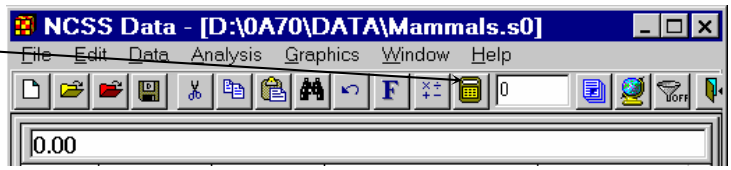
4. Click in the cell in the fifth row and third column.



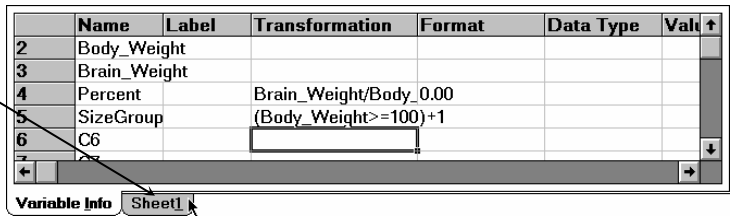
5. Type **(Body_Weight>=100)+1** and press **Enter**.



6. Press the **Apply Transformations** button to generate the new values.



7. Click on the **Sheet1** tab to return to the data.



The final result appears like this.

	Name	Body_Weight	Brain_Weight	Percent	SizeGrou	C6
1	African Elephant	6654	5712	0.09	2	
2	Asian Elephant	2547	4603	0.18	2	
3	Giraffe	529	680	0.13	2	
4	Horse	521	655	0.13	2	
5	Cow	465	423	0.09	2	
6	Gorilla	207	406	0.20	2	
7	Pig	192	180	0.09	2	
8	Jaguar	100	157	0.16	2	
9	Man	62	1320	2.13	1	
10	Chimpanzee	52	440	0.85	1	
11	Gray Wolf	36	120	0.33	1	
12	Kangaroo	35	56	0.16	1	
13	Baboon	11	179	1.63	1	
14	Red Fox	4	50	1.25	1	
15	Cat	3	26	0.87	1	

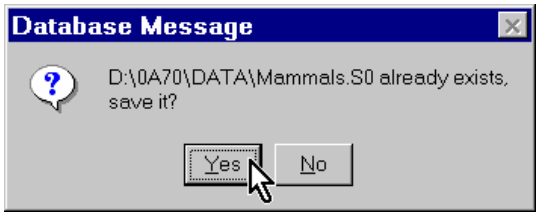
Saving the Changes

It is important to save changes to your database. Take the following steps to do this.

1. Choose ***Save*** from the File menu of the Data window.

2. Click ***Yes***.

   The MAMMALS database on your hard disk will be replaced with the revised edition.



Chapter 4

Running Descriptive Statistics

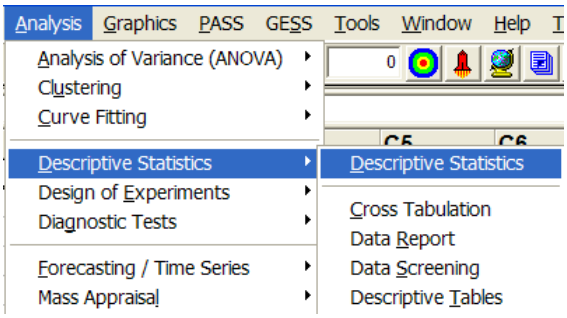
About This Chapter (Time: 8 minutes)

This chapter continues the introduction to the **NCSS** system by taking you through an example of using **NCSS** to obtain descriptive statistics.

Running Descriptive Statistics

In this section, you will generate descriptive statistics (mean, standard deviation, etc.) on the Body_Weight variable in the MAMMALS data. To begin, start **NCSS** and load the MAMMALS database. Detailed instructions for doing this are at the beginning of Chapter 3. After the database is loaded, follow these steps to run the procedure:

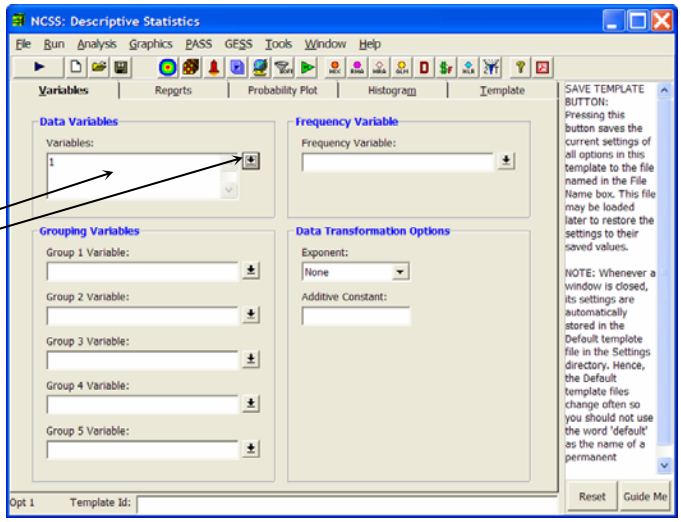
1. Click on the Analysis menu. From the Descriptive Statistics menu, select **Descriptive Statistics**.



4-2 Quick Start – Running Descriptive Statistics

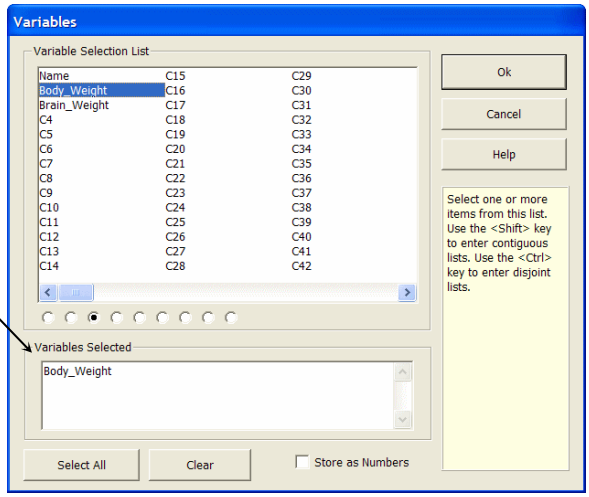
The Descriptive Statistics window will appear.

The next step is to select the variables to be analyzed.



2. Double click in the **Variables** box or click the small button to the right of this box.

This will cause the Variables window to appear.

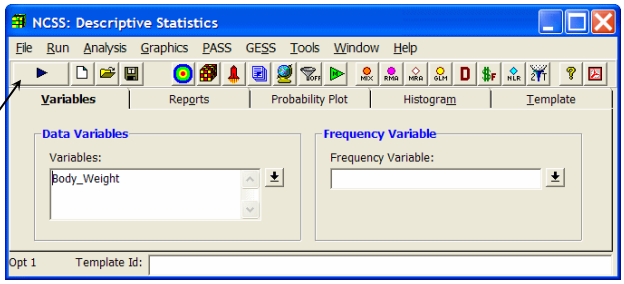


3. Click on **Body_Weight** in the Variable Selection List box.

The variable will appear in the Variables Selected box.

4. Click **Ok**.

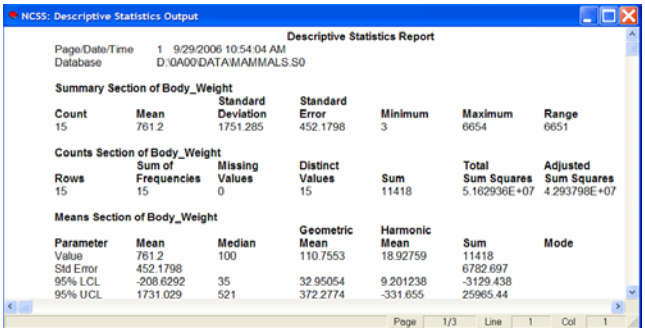
The procedure window reappears. Note that the Variables option now has a value of ***Body_Weight***. This is the name of the variable that was selected.



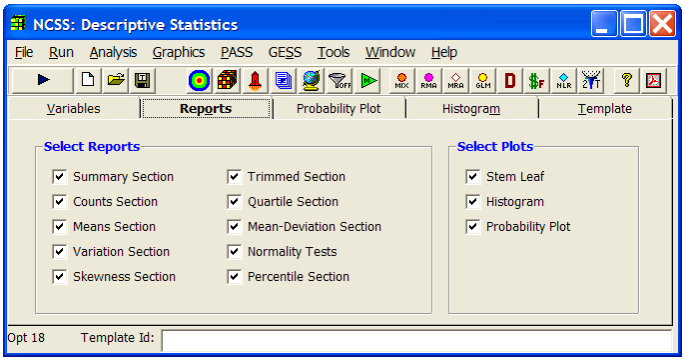
5. Press the **Run** button to run the procedure and generate the following output report.

The results are displayed in NCSS's word processor.

You can scroll through the output using the scroll bars. You can enlarge this window by double-clicking the title bar--the bar at the top containing the words NCSS: Descriptive Statistics Output.

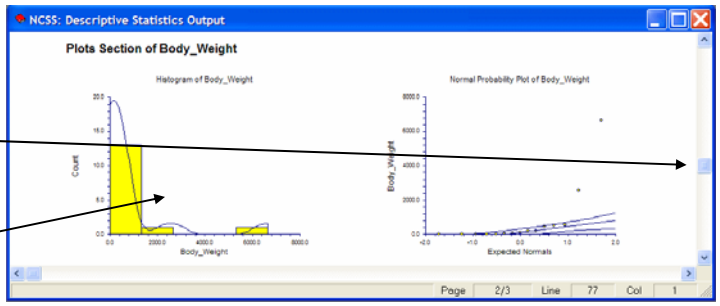


Don't be intimidated by the amount of output. The default descriptive statistics report contains much more information than would normally be used. You can generate only those reports you want by making appropriate selections on the Reports panel of the Descriptive Statistics window.

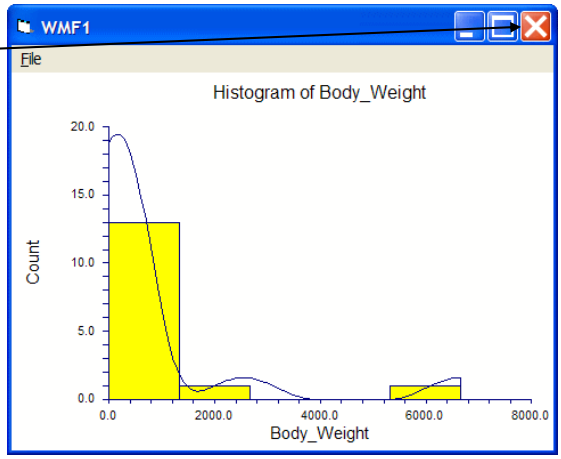


We will now show how to view the graphics in more detail.

6. Scroll down through the output until reach the histogram.
7. Double-click the histogram to obtain a full-screen version of the histogram.



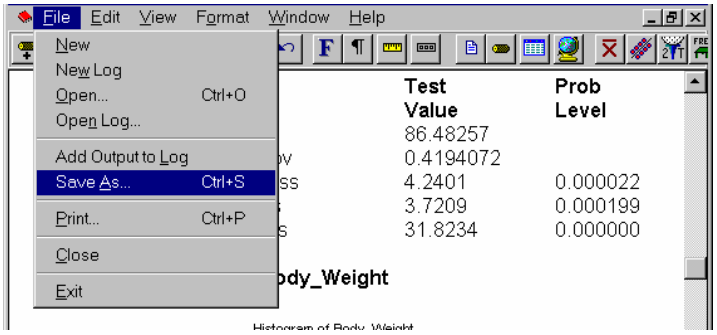
8. After viewing the graph, close it by clicking the close button.



Saving the Output

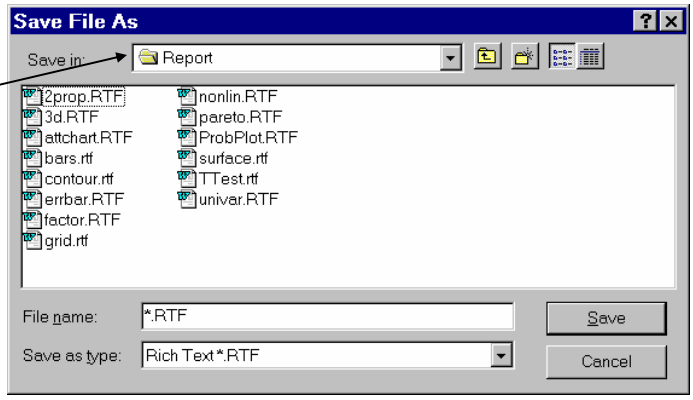
We will now show you how to save the output so that it can be imported into your word processor.

1. Select ***Save As*** from the File menu of the Output window.



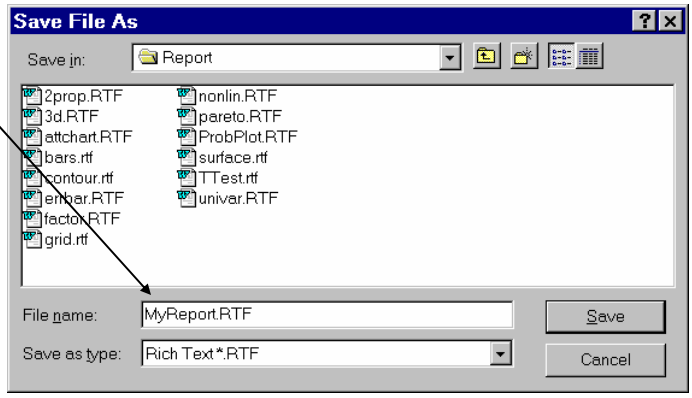
This will bring up the Save File As dialog box.

Switch the current directory to the **Report** subdirectory, which is provided as a convenient place to save your reports.



2. Type **myreport.rtf** in the File name box.

3. Click ***Save*** to save the report.



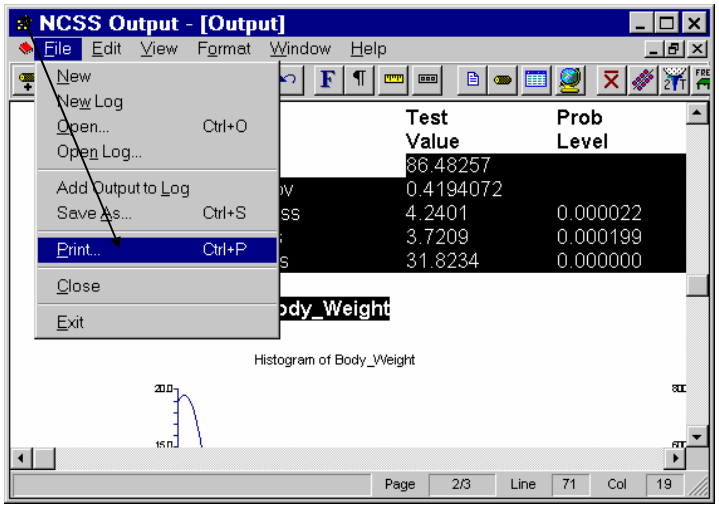
Note that the three-character extension “rtf” is very important. RTF stands for rich text format. Other programs, such as Microsoft Word and WordPerfect, recognize files with this extension as importable. Hence, using this extension makes sure that other programs will be able to import your report files.

As an exercise, run your word processor and load the **myreport.rtf** file.

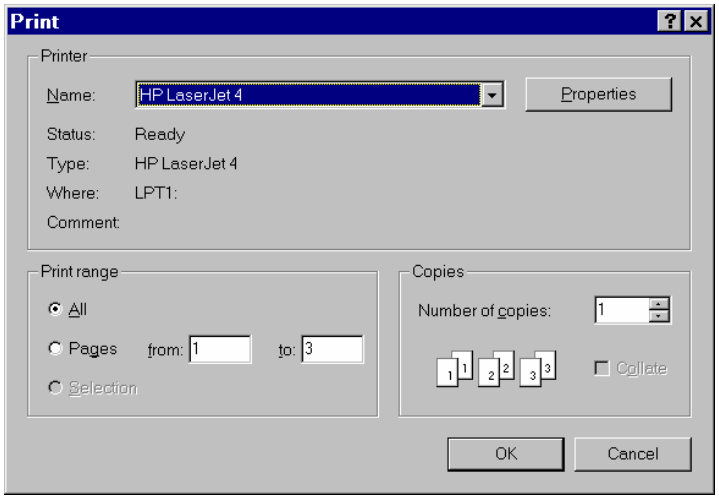
Printing the Output

We will now show you how to print the output.

1. Select ***Print*** from the File menu.



This will bring up the Print dialog box.



You can select which pages you want to print.

2. Click ***OK*** to begin printing.

Chapter 5

Running a Two-Sample T-Test

About This Chapter (Time: 6 minutes)

This chapter continues the introduction to the **NCSS** system by taking you through an example of using **NCSS** to run a two-sample t-test.

Running a Two-Sample T-Test

In this section, you will conduct a two-sample t-test on data in the MAMMALS1 database. To begin, start **NCSS** and load the MAMMALS1 database (be careful to load MAMMALS1, not MAMMALS). Detailed instructions for loading a file are at the beginning of Chapter 3.

Remember to load the database from the Data window.

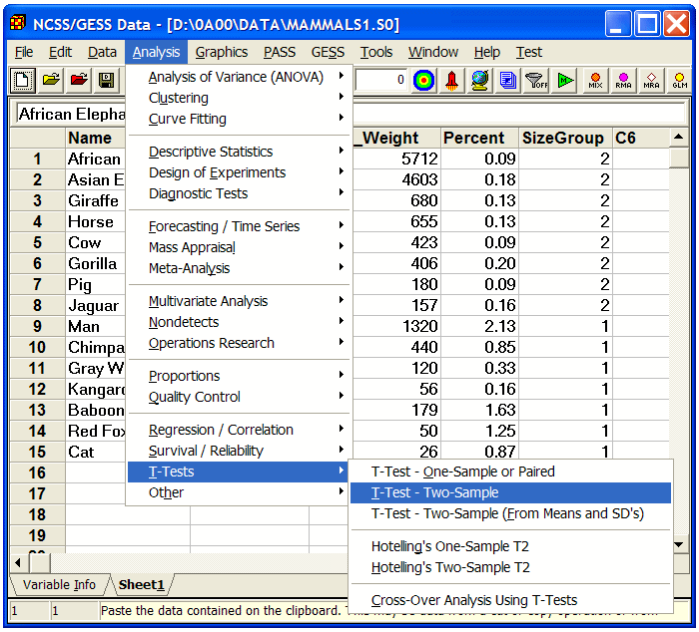
In this example, we will compare the average percent brain weight of small mammals (those under 100 kg in weight) to the same average for large mammals. That is, the response variable will be **Percent** and the grouping variable will be **SizeGroup.**

	Name	Body_Weight	Brain_Weight	Percent	SizeGrou	C6
1	African Elephant	6654	5712	0.09	2	
2	Asian Elephant	2547	4603	0.18	2	
3	Giraffe	529	680	0.13	2	
4	Horse	521	655	0.13	2	
5	Cow	465	423	0.09	2	
6	Gorilla	207	406	0.20	2	
7	Pig	192	180	0.09	2	
8	Jaguar	100	157	0.16	2	
9	Man	62	1320	2.13	1	
10	Chimpanzee	52	440	0.85	1	
11	Gray Wolf	36	120	0.33	1	
12	Kangaroo	35	56	0.16	1	
13	Baboon	11	179	1.63	1	
14	Red Fox	4	50	1.25	1	
15	Cat	3	26	0.87	1	
16						
17						

After the database is loaded, follow these steps to run the procedure:

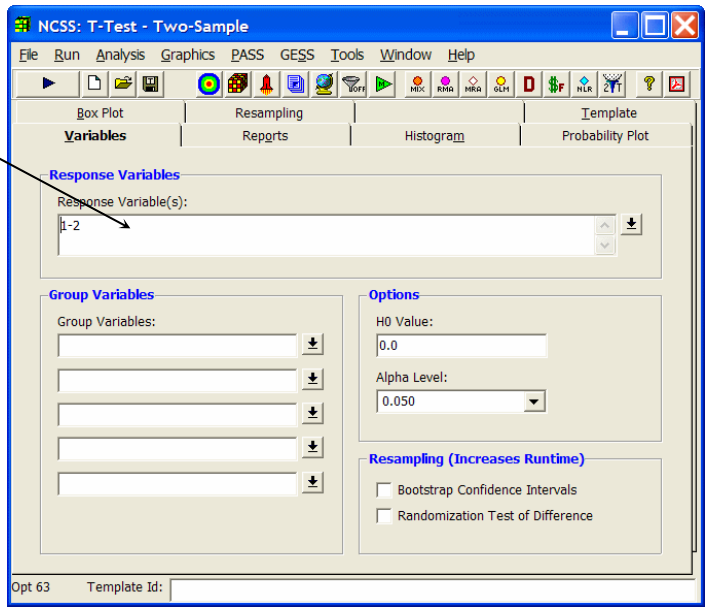
5-2 Quick Start – Running a Two-Sample T-Test

1. From the T-Tests submenu of the Analysis menu, select **T-Test - Two-Sample**.



The Two Sample Tests procedure window will appear.

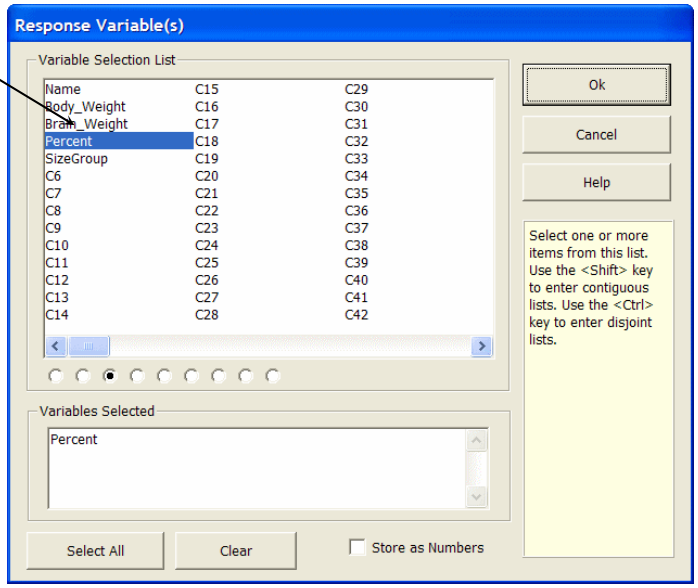
2. Double click in the **Response Variables** box.



This will cause the Response Variables selection window to appear.

3. Click on the **Percent** item in the Variable Selection List box.

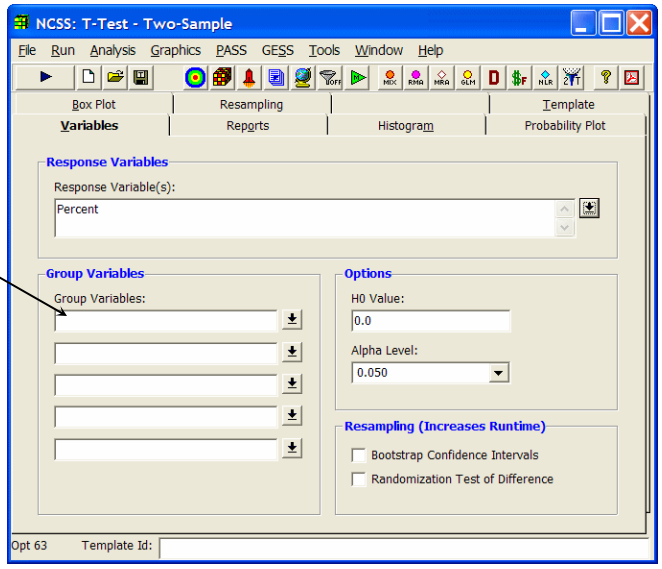
4. Click **Ok**.



The Response Variables now has the entry **Percent**. This is the variable that was selected.

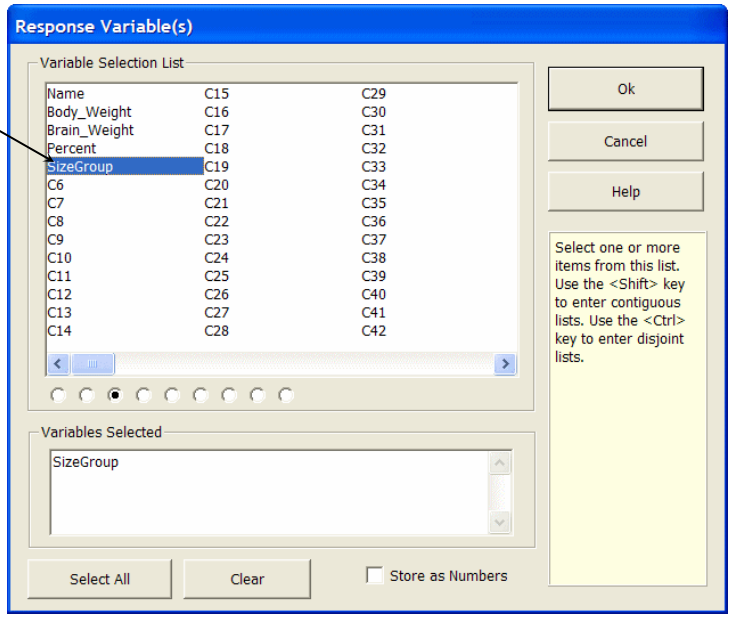
5. Double click the top **Group Variable** box.

This is the grouping variable. The average percent of those rows with a SizeGroup value of 1 (small animals) will be compared with the average percent of those rows with a SizeGroup value of 2 (large animals).

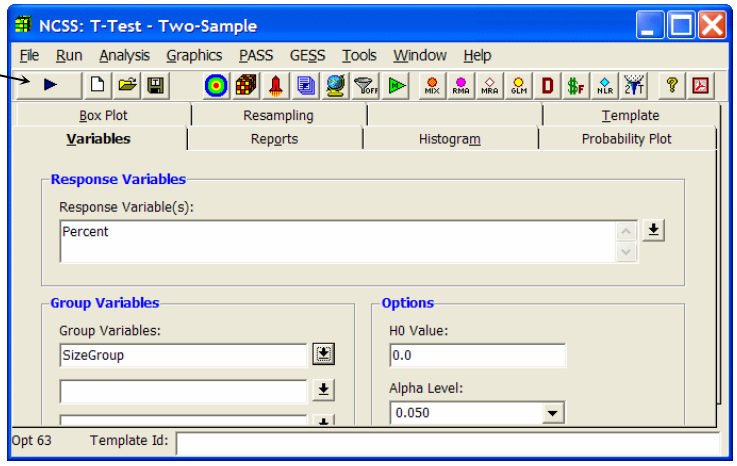


6. Select **SizeGroup** from the list of available variables.

7. Click ***Ok***.



8. Click the ***Run*** button to run the analysis.



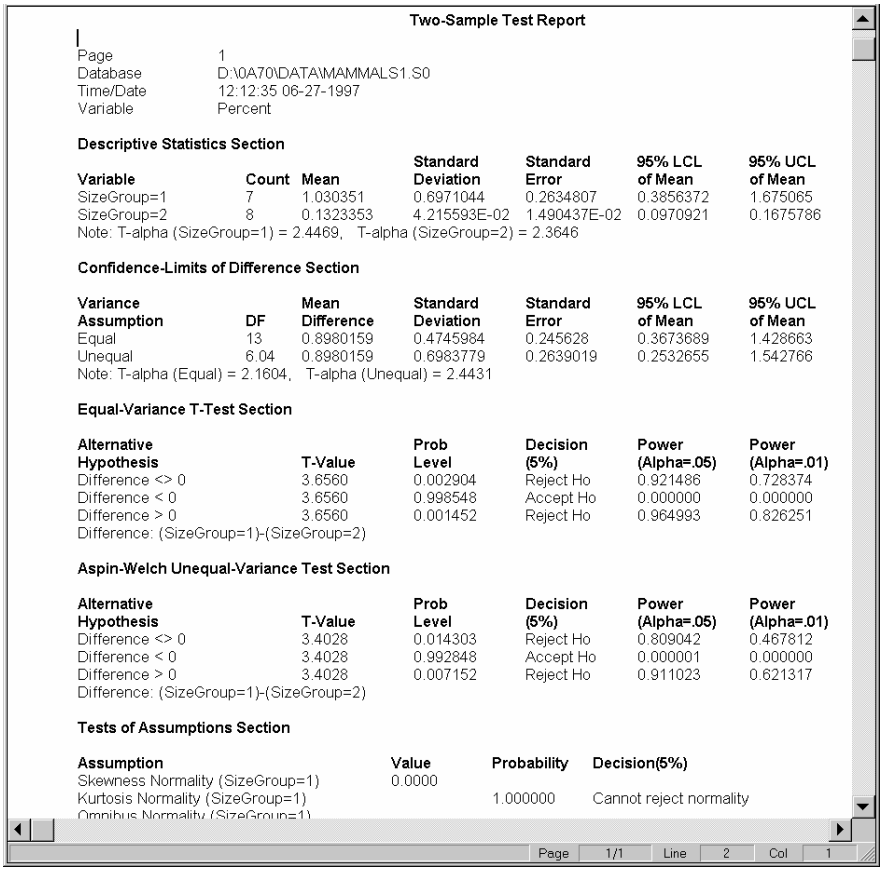
The results are displayed in ***NCSS***'s word processor.

The T-Test compares the mean percent of two groups. Often, all you will need is the t-value and associated probability level. These are contained in the Equal-Variance T-Test Section. In this case the T-value is 3.6560 and the probability level is 0.002904. Hence we reject the null hypothesis that means are equal.

A quick glance at the means of the two groups shows that the mean percent for small animals is 1.03 and for large animals is 0.13. Hence the two percentages are an order of magnitude apart!

The T-Test chapter of the User's Guide goes into much more detail on how to perform a T-Test analysis.

At this point, you could save or print the t-test report.



Two-Sample Test Report

Page 1
Database D:\0A70\DATA\MAMMALS1.S0
Time/Date 12:12:35 06-27-1997
Variable Percent

Descriptive Statistics Section

Variable	Count	Mean	Standard Deviation	Standard Error	95% LCL of Mean	95% UCL of Mean
SizeGroup=1	7	1.030351	0.6971044	0.2634807	0.3856372	1.675065
SizeGroup=2	8	0.1323353	4.215593E-02	1.490437E-02	0.0970921	0.1675786

Note: T-alpha (SizeGroup=1) = 2.4469, T-alpha (SizeGroup=2) = 2.3646

Confidence-Limits of Difference Section

Variance Assumption	DF	Mean Difference	Standard Deviation	Standard Error	95% LCL of Mean	95% UCL of Mean
Equal	13	0.8980159	0.4745984	0.245628	0.3673689	1.428663
Unequal	6.04	0.8980159	0.6983779	0.2639019	0.2532655	1.542766

Note: T-alpha (Equal) = 2.1604, T-alpha (Unequal) = 2.4431

Equal-Variance T-Test Section

Alternative Hypothesis	T-Value	Prob Level	Decision (5%)	Power (Alpha=.05)	Power (Alpha=.01)
Difference <> 0	3.6560	0.002904	Reject Ho	0.921486	0.728374
Difference < 0	3.6560	0.998548	Accept Ho	0.000000	0.000000
Difference > 0	3.6560	0.001452	Reject Ho	0.964993	0.826251

Difference: (SizeGroup=1)-(SizeGroup=2)

Aspin-Welch Unequal-Variance Test Section

Alternative Hypothesis	T-Value	Prob Level	Decision (5%)	Power (Alpha=.05)	Power (Alpha=.01)
Difference <> 0	3.4028	0.014303	Reject Ho	0.809042	0.467812
Difference < 0	3.4028	0.992848	Accept Ho	0.000001	0.000000
Difference > 0	3.4028	0.007152	Reject Ho	0.911023	0.621317

Difference: (SizeGroup=1)-(SizeGroup=2)

Tests of Assumptions Section

Assumption	Value	Probability	Decision(5%)
Skewness Normality (SizeGroup=1)	0.0000		
Kurtosis Normality (SizeGroup=1)		1.000000	Cannot reject normality
Omnibus Normality (SizeGroup=1)			

Page 1/1 Line 2 Col 1

5-6 Quick Start – Running a Two-Sample T-Test

Chapter 6

Running a Regression Analysis

About This Chapter (Time: 10 minutes)

This chapter continues the introduction to the **NCSS** system by taking you through an example of regression analysis. Regression techniques analyze the relationship between a dependent (Y) variable and one or more independent (X) variables. **NCSS** has regression procedures for many different situations.

Running a Regression Analysis

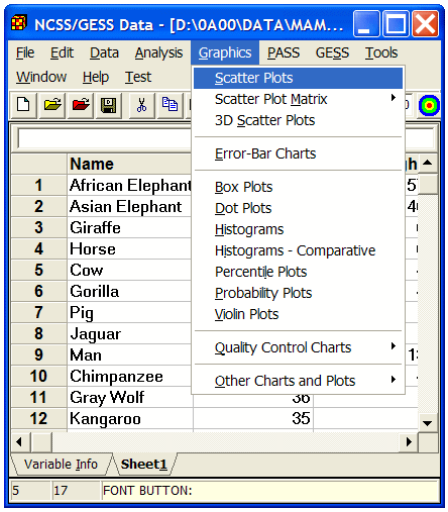
In this section, you will conduct a regression analysis using the MAMMALS1 database. To begin, start **NCSS** and load the MAMMALS1 database. Detailed instructions for loading a database are at the beginning of Chapter 3.

In this example we will investigate the relationship between ***Brain_Weight*** (dependent variable) and ***Body_Weight*** (independent variable).

	Name	Body_Weight	Brain_Weight	Percent	SizeGrou	C6
1	African Elephant	6654	5712	0.09	2	
2	Asian Elephant	2547	4603	0.18	2	
3	Giraffe	529	680	0.13	2	
4	Horse	521	655	0.13	2	
5	Cow	465	423	0.09	2	
6	Gorilla	207	406	0.20	2	
7	Pig	192	180	0.09	2	
8	Jaguar	100	157	0.16	2	
9	Man	62	1320	2.13	1	
10	Chimpanzee	52	440	0.85	1	
11	Gray Wolf	36	120	0.33	1	
12	Kangaroo	35	56	0.16	1	
13	Baboon	11	179	1.63	1	
14	Red Fox	4	50	1.25	1	
15	Cat	3	26	0.87	1	
16						
17						

Creating a Scatter Plot

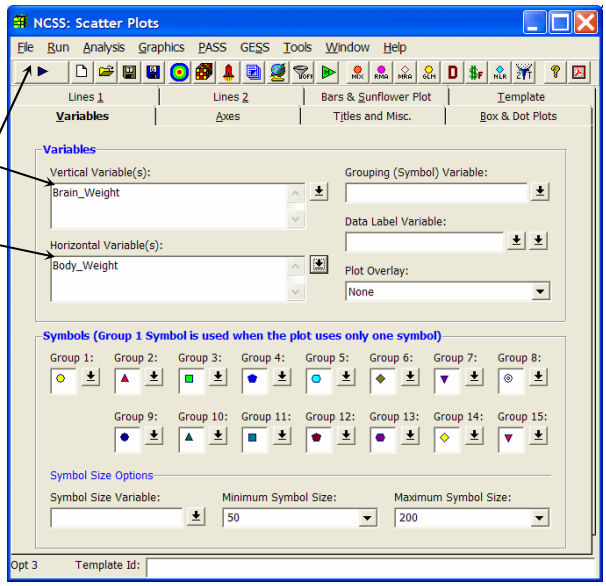
The first step in a regression analysis is to plot the data.



1. From the Graphics menu, select **Scatter Plots**.

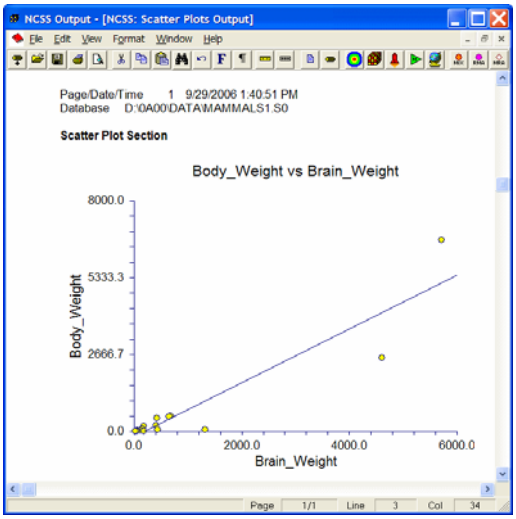
   The Scatter Plot window will appear.

2. Click in the ***Vertical Variable(s)*** box.
3. Enter ***Brain_Weight***.
4. Click in the ***Horizontal Variable(s)*** box.
5. Enter ***Body_Weight***.
6. Click the ***Run*** button on the toolbar.



The scatter plot shown at the right will appear. In order for regression analysis to be applied, the points in the plot should fall along an imaginary straight line.

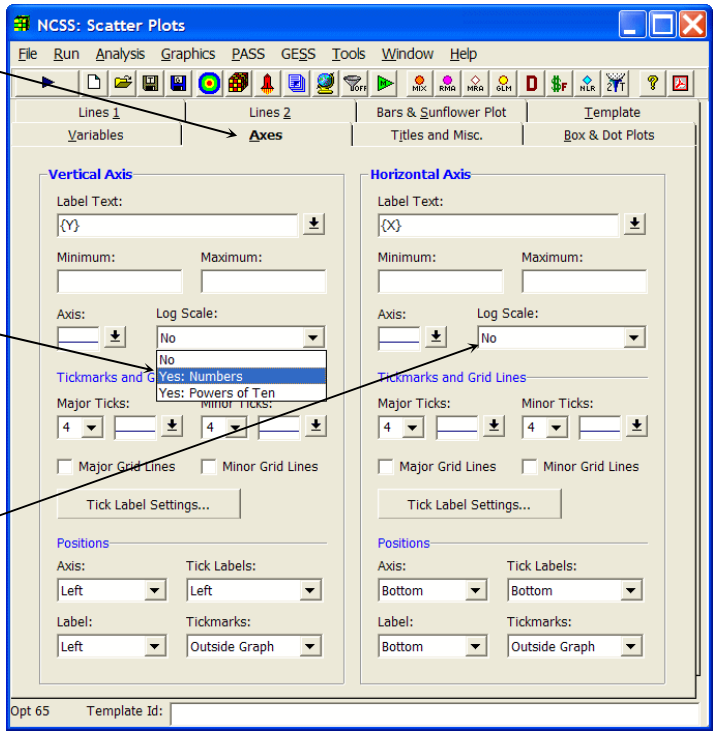
While studying the plot, notice that all but two of the points are clustered in the lower left-hand corner. You cannot tell whether the points fall along a straight line. This suggests that a logarithmic scale should be used to display the data. This will be done next.



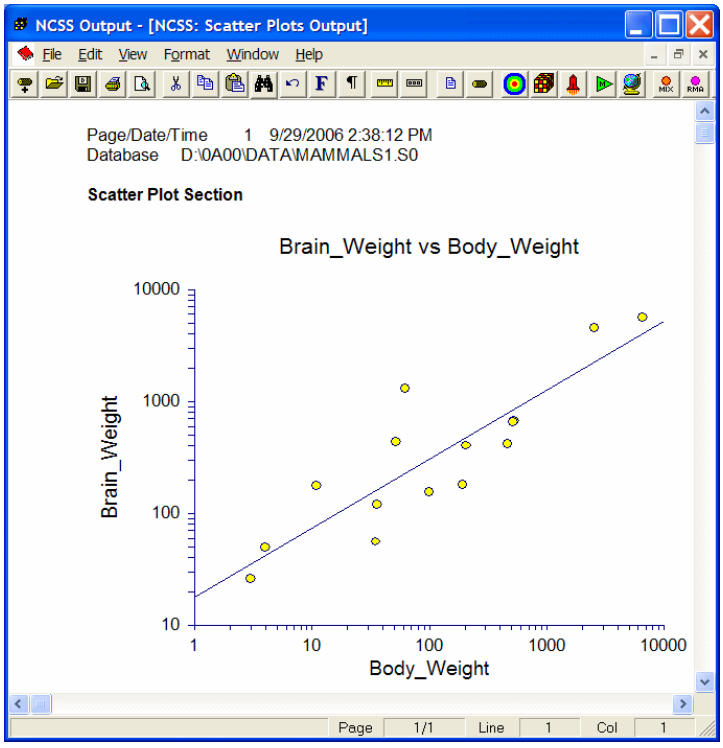
7. Press the **Axes** tab to display the Axes panel.

8. Select **Yes: Numbers** from the **Log Scale** pull-down list box for the vertical axis.

9. Select **Yes: Numbers** from the **Log Scale** pull-down list box for the horizontal axis.



10. Press the **Run** button to run the program and generate the following output.

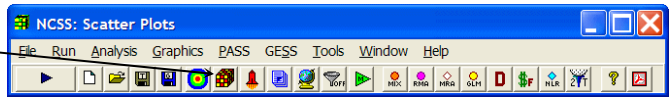


The final result is the plot at the right. Notice that the points now appear to be evenly spread across the plot. Also note that the points appear to fall along an upward-sloping straight line. This implies that a standard regression analysis should produce a reasonable model of this data.

Because of the visual results from using the logarithmic scale, our next task will be to create logarithmic versions of the two variables.

Create the Logarithmic Variables

1. Press the **Data Window** button on the toolbar to bring the **NCSS Data** window to the front of your screen.



This will bring up the **NCSS Data** window.

	Name	Body_Weight	Brain_Weight	Percent	SizeGroup
1	African Elephant	6654	5712	0.09	
2	Asian Elephant	2547	4603	0.18	
3	Giraffe	529	680	0.13	
4	Horse	521	655	0.13	
5	Cow	465	423	0.09	
6	Gorilla	207	406	0.20	
7	Pig	192	180	0.09	

Variable Info | Sheet1

This is the spreadsheet editor which lets you enter and edit your data.

2. Click on the **Variable Info** tab.

This will bring up the **Variable Info** screen.

LogBrain

	Name	Label	Transformation	Format	Data Type
2	Body_Weight				
3	Brain_Weight				
4	Percent		Brain_Weight/Body	0.00	
5	SizeGroup		(Body_Weight>=100)+1		
6	LogBody		Log(Body_Weight)		
7	LogBrain		Log(Brain_Weight)		
8	C8				

Variable Info | Sheet1

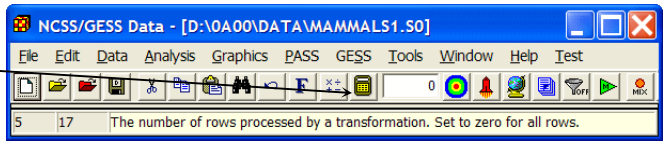
3. In the sixth row of the Transformation column enter **Log(Body_Weight)**.

4. In the seventh row of the Transformation column enter **Log(Brain_Weight)**.

5. In the sixth row of the Name column enter **LogBody**.

6. In the seventh row of the Name column enter **LogBrain**.

7. Click on the **Apply Transformations** button to create the transformed data.



8. Click on the **Sheet1** tab to return to your data. The datasheet will now appear as shown.

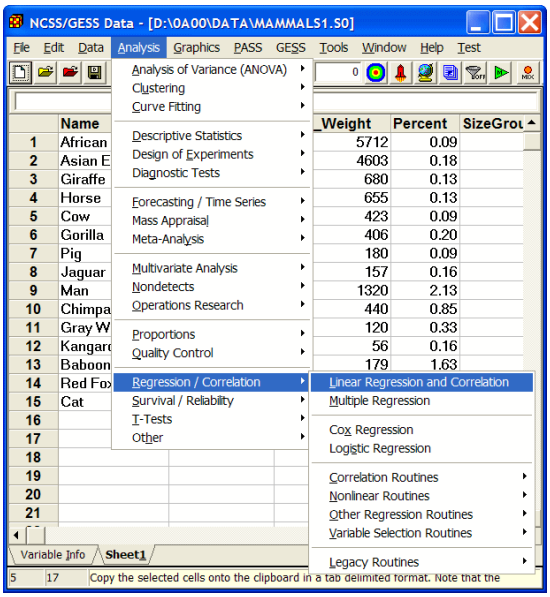
2.60852603357719

	Name	Body_Weight	Brain_Weight	Percent	SizeGroup	LogBody	LogBrain
1	African Elephant	6654	5712	0.09	2	3.8230828	3.7567882
2	Asian Elephant	2547	4603	0.18	2	3.40602894	3.66304097
3	Giraffe	529	680	0.13	2	2.72345567	2.83250891
4	Horse	521	655	0.13	2	2.71683772	2.8162413
5	Cow	465	423	0.09	2	2.66745295	2.62634037
6	Gorilla	207	406	0.20	2	2.31597035	2.60852603
7	Pig	192	180	0.09	2	2.28330123	2.25527251
8	Jaguar	100	157	0.16	2	2	2.19589965
9	Man	62	1320	2.13	1	1.79239169	3.12057393
10	Chimpanzee	52	440	0.85	1	1.71600334	2.64345268
11	Gray Wolf	36	120	0.33	1	1.5563025	2.07918125
12	Kangaroo	35	56	0.16	1	1.54406804	1.74818803
13	Baboon	11	179	1.63	1	1.04139269	2.25285303
14	Red Fox	4	50	1.25	1	0.60205999	1.69897
15	Cat	3	26	0.87	1	0.47712125	1.41497335
16							

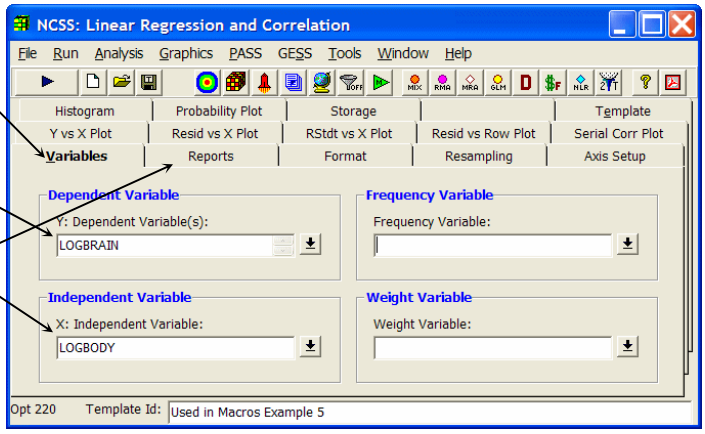
Variable Info | Sheet1

Run the Regression

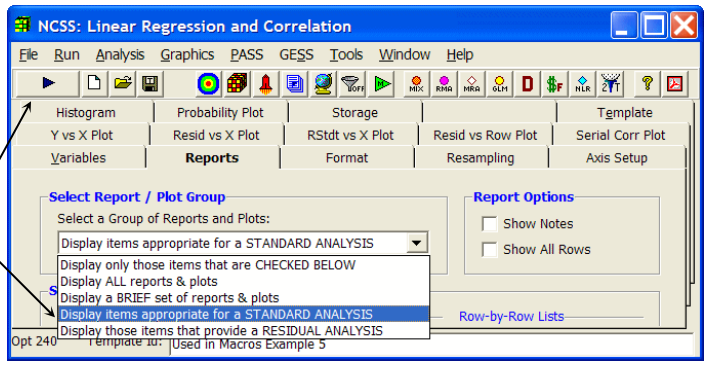
1. Select **Linear Regression and Correlation** from the Regression / Correlation submenu of the Analysis menu.



2. Click on the **Variables** tab.
3. Enter **LogBrain** for the **Y: Dependent Variable**.
4. Enter **LogBody** for the **X: Independent Variable**.
5. Click on the **Reports** tab.



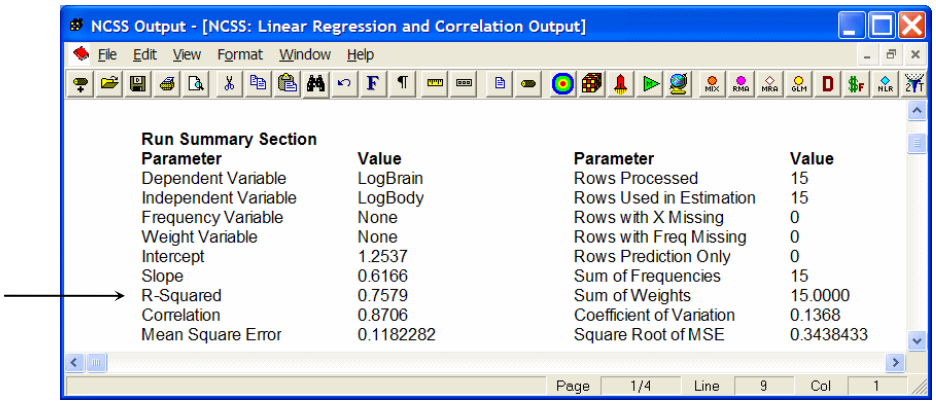
6. Under 'Select a Group of Reports and Plots', select **Display items appropriate for a STANDARD ANALYSIS**.
7. Click the **Run** button on the toolbar.



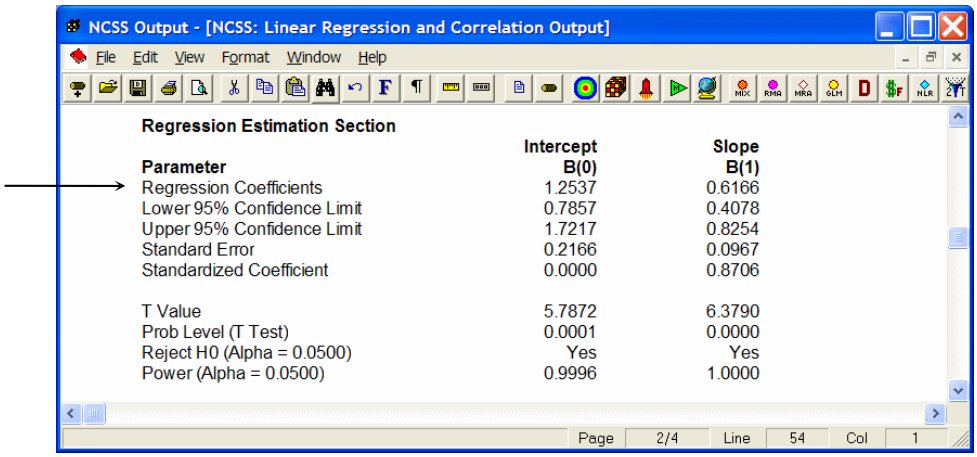
This will generate the output that follows.

6-6 Quick Start – Running a Regression Analysis

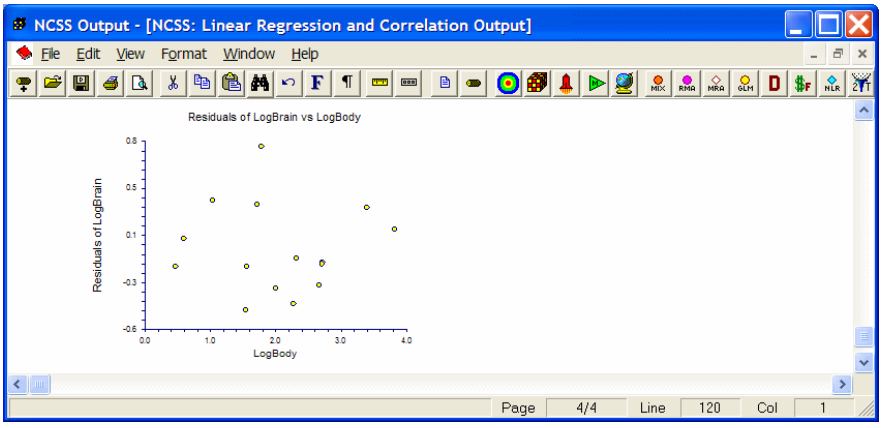
The main statistics of interest in a regression analysis are the R-Squared and the regression coefficients. The R-Squared is shown in the Run Summary Section:



The regression coefficients are shown in the Run Estimation Section.



The residual plot is found at the bottom of the output:



Of course, a complete regression analysis would require the studying of several reports and plots. A complete discussion of this is found in the regression chapters of the ***User's Guide***.

Chapter 7

Data Window

About This Chapter

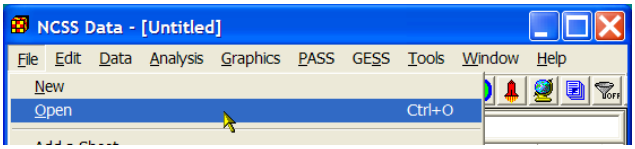
Data may be entered manually or imported from other files. The data are loaded in a spreadsheet from which they may be viewed, changed, stored, or printed. This chapter will show you how to manipulate your data using the spreadsheet.

Loading a Database

The tutorial in Chapter 2 explained the mechanics of entering, storing, and printing a database, so that material will not be repeated here. Instead, this chapter will focus on manipulating the data with the spreadsheet after it has been loaded. Our first task will be to load in a previously saved database.

If **NCSS** is not already running, start it now by selecting **NCSS** from the Windows Start menu (refer to the beginning of Chapter 2 for details). We will use the brain weight data that was entered in Chapter 2. These data are stored in the MAMMALS database in the \NCSS2007\DATA subdirectory. To begin this tutorial, take the following steps to load the MAMMALS database.

1. Select **Open** from the File menu of the Data window. The File Open window will appear.

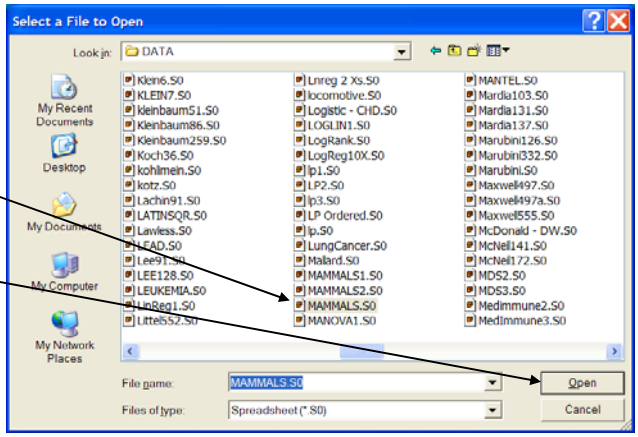


2. Double click the **Data** subdirectory to select it.

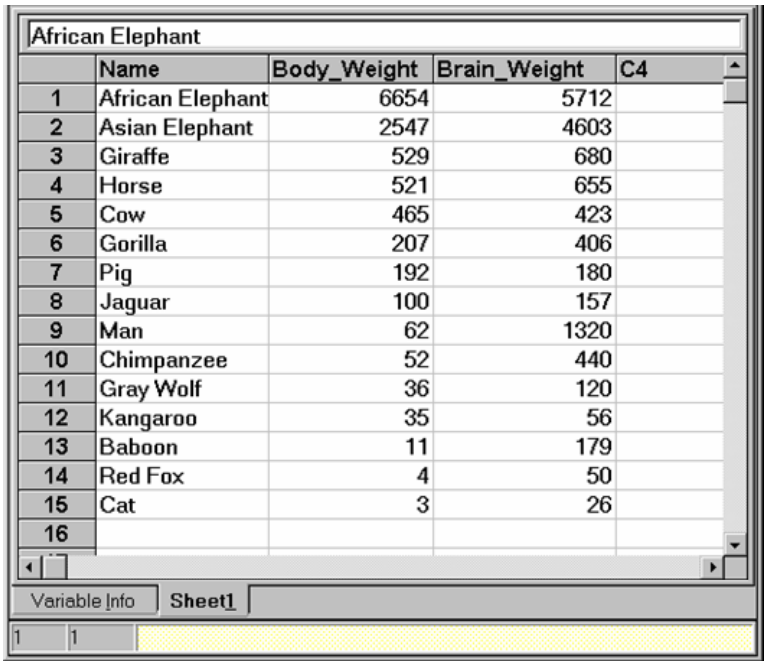
3. Double click **MAMMALS.S0** in the list of available files.

4. Click the ***Open*** button.

This will load the MAMMALS database into the Data window.



The Data window will appear as shown to the right.


African Elephant

	Name	Body_Weight	Brain_Weight	C4
1	African Elephant	6654	5712	
2	Asian Elephant	2547	4603	
3	Giraffe	529	680	
4	Horse	521	655	
5	Cow	465	423	
6	Gorilla	207	406	
7	Pig	192	180	
8	Jaguar	100	157	
9	Man	62	1320	
10	Chimpanzee	52	440	
11	Gray Wolf	36	120	
12	Kangaroo	35	56	
13	Baboon	11	179	
14	Red Fox	4	50	
15	Cat	3	26	
16				

Variable Info | Sheet1

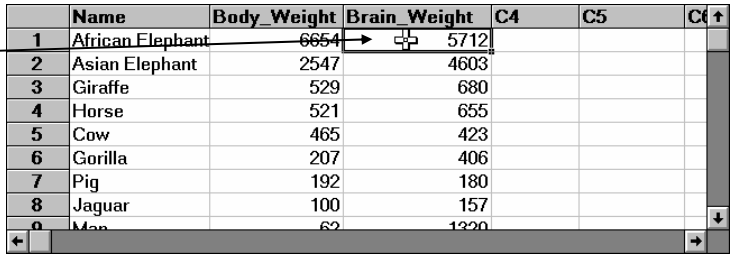
1 1

We next examine how to copy and paste data in the spreadsheet.

Copying and Pasting Data

We will now take you through the steps to copy and paste the data.

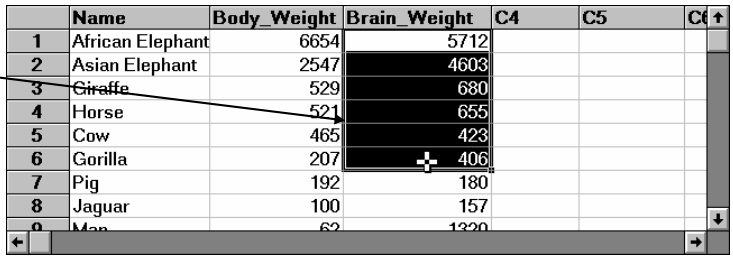
1. Position the cursor in row one column three (at the value ***5712***).



	Name	Body_Weight	Brain_Weight	C4	C5	C6
1	African Elephant	6654	5712			
2	Asian Elephant	2547	4603			
3	Giraffe	529	680			
4	Horse	521	655			
5	Cow	465	423			
6	Gorilla	207	406			
7	Pig	192	180			
8	Jaguar	100	157			

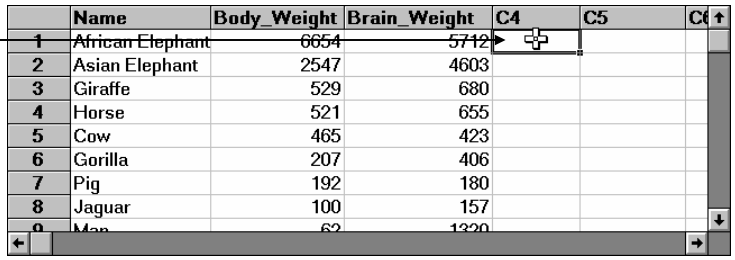
2. Drag the mouse down to row six. This will select the first six rows.

3. Press ***Ctrl-C***. This will copy the data to a temporary storage area called the *clipboard*.



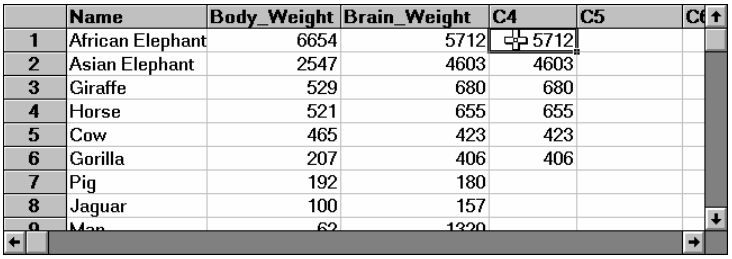
	Name	Body_Weight	Brain_Weight	C4	C5	C6
1	African Elephant	6654	5712			
2	Asian Elephant	2547	4603			
3	Giraffe	529	680			
4	Horse	521	655			
5	Cow	465	423			
6	Gorilla	207	406			
7	Pig	192	180			
8	Jaguar	100	157			

4. Position the cursor in the cell at row one and column four.



	Name	Body_Weight	Brain_Weight	C4	C5	C6
1	African Elephant	6654	5712			
2	Asian Elephant	2547	4603			
3	Giraffe	529	680			
4	Horse	521	655			
5	Cow	465	423			
6	Gorilla	207	406			
7	Pig	192	180			
8	Jaguar	100	157			
9	Man	62	1320			

5. Press ***Ctrl-V*** to paste the data from the clipboard. The resulting screen will appear as shown.

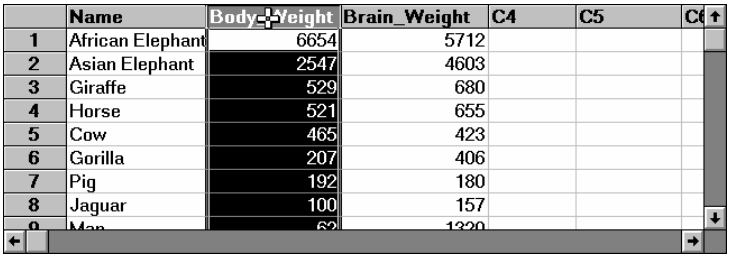


	Name	Body_Weight	Brain_Weight	C4	C5	C6
1	African Elephant	6654	5712	5712		
2	Asian Elephant	2547	4603	4603		
3	Giraffe	529	680	680		
4	Horse	521	655	655		
5	Cow	465	423	423		
6	Gorilla	207	406	406		
7	Pig	192	180			
8	Jaguar	100	157			
9	Man	62	1320			

Changing Column Widths

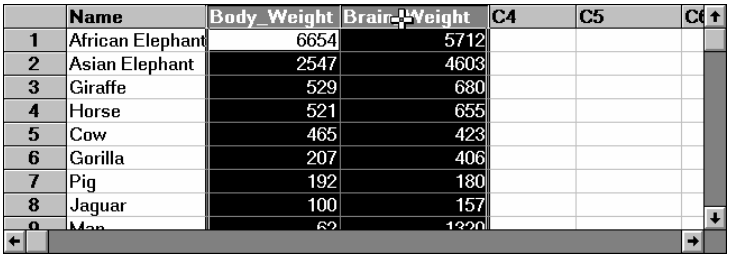
Occasionally, you will want to change the width of one or more columns. This section will show you how this is accomplished. We will resize the columns headed ***Body_Weight*** and ***Brain_Weight***.

1. Click on the column heading: ***Body_Weight***.



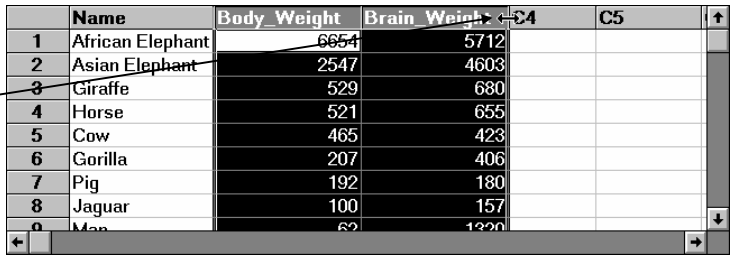
	Name	Body_Weight	Brain_Weight	C4	C5	C6
1	African Elephant	6654	5712			
2	Asian Elephant	2547	4603			
3	Giraffe	529	680			
4	Horse	521	655			
5	Cow	465	423			
6	Gorilla	207	406			
7	Pig	192	180			
8	Jaguar	100	157			
9	Man	62	1320			

2. Drag the mouse into the next column to the right and let go of the mouse button. This will select these two columns.



	Name	Body_Weight	Brain_Weight	C4	C5	C6
1	African Elephant	6654	5712			
2	Asian Elephant	2547	4603			
3	Giraffe	529	680			
4	Horse	521	655			
5	Cow	465	423			
6	Gorilla	207	406			
7	Pig	192	180			
8	Jaguar	100	157			
9	Man	62	1320			

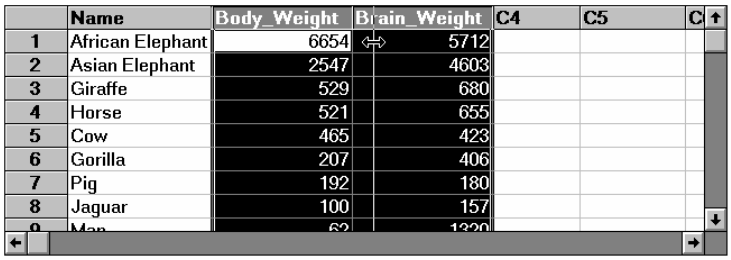
3. Move the cursor between the third and fourth columns. The cursor will change into a double-pointing arrow.



	Name	Body_Weight	Brain_Weight	C4	C5
1	African Elephant	6654	5712		
2	Asian Elephant	2547	4603		
3	Giraffe	529	680		
4	Horse	521	655		
5	Cow	465	423		
6	Gorilla	207	406		
7	Pig	192	180		
8	Jaguar	100	157		

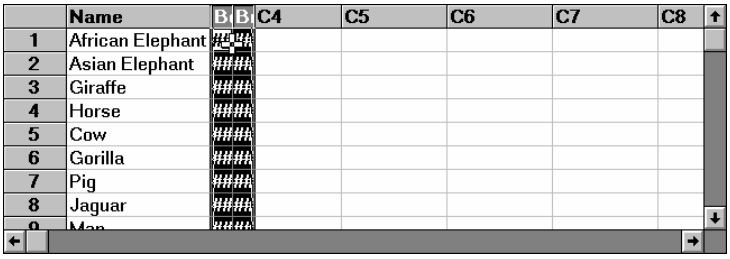
4. While holding down on the mouse button, drag it to the left until you are almost to the next cell border.

5. Let go of the mouse button.



	Name	Body_Weight	Brain_Weight	C4	C5	C
1	African Elephant	6654	5712			
2	Asian Elephant	2547	4603			
3	Giraffe	529	680			
4	Horse	521	655			
5	Cow	465	423			
6	Gorilla	207	406			
7	Pig	192	180			
8	Jaguar	100	157			

The resulting display will appear like this.



	Name	B	B	C4	C5	C6	C7	C8
1	African Elephant	###						
2	Asian Elephant	###						
3	Giraffe	###						
4	Horse	###						
5	Cow	###						
6	Gorilla	###						
7	Pig	###						
8	Jaguar	###						

6. Reverse this process to reset these columns to their original width.

Chapter 8

Procedure Window

About This Chapter

All **NCSS** procedures (e.g., T-Test, Multiple Regression, and Scatterplot) are controlled by a procedure window. The Procedure window contains all the settings, options, and parameters that control a particular procedure. These options are separated into groups called *panels*. A particular panel is viewed by pressing the corresponding *panel tab* that appears just below the toolbar near the top of the window.

The current values of all options available for a procedure are referred to as a *template*. By creating and saving template files, you can tailor each procedure to your own specific needs.

Below is a picture of the Descriptive Statistics procedure window. This chapter presents a brief tutorial of how to operate the Procedure window.

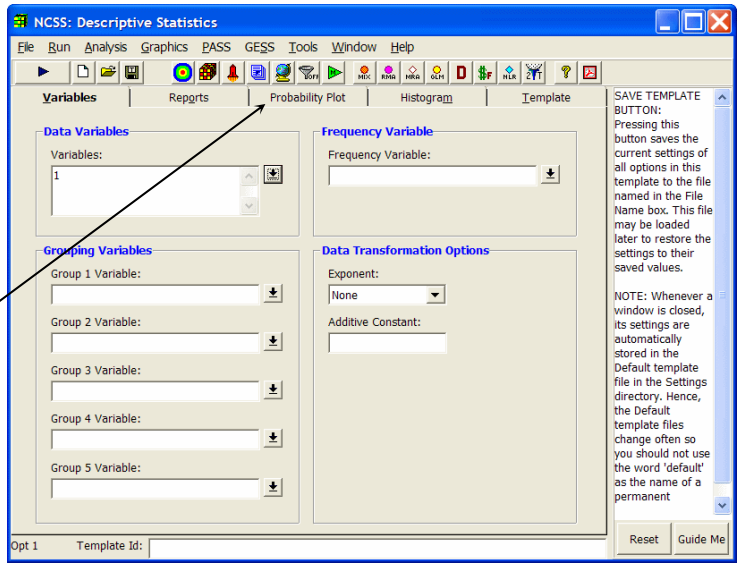
Navigating a Procedure Window

This section will show you how to move around a procedure window. The window is made up of two or more panels (in this example there are five panels: Variables, Reports, Probability Plot, Histogram, and Template). You control a procedure by changing the settings on each of these panels. Hence, navigating a procedure window simply means that you move from panel to panel.

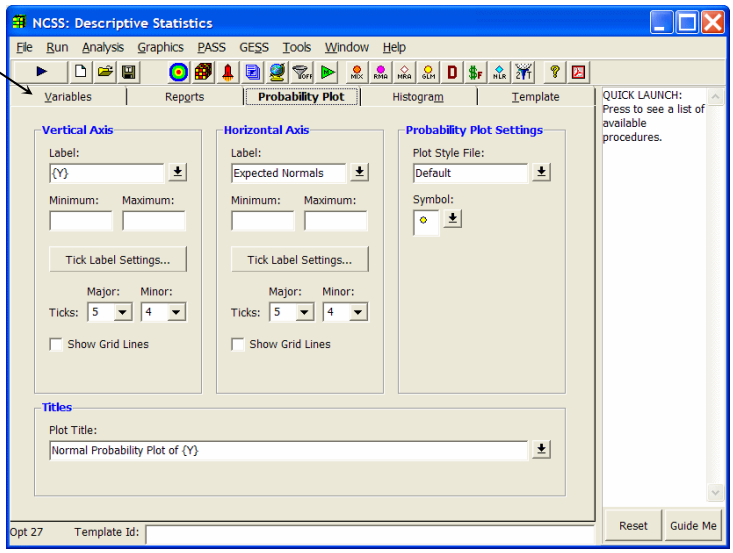
1. From the **Analysis** menu, select **Descriptive Statistics**, then **Descriptive Statistics**.

   The Descriptive Statistics procedure window will appear.

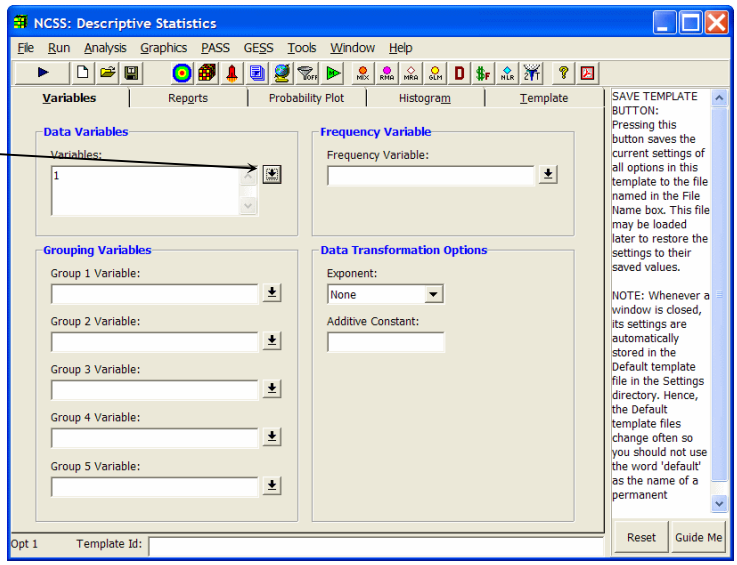
2. Press the **Probability Plot** tab to display the Probability Plot panel.



3. Press the ***Variables*** tab to redisplay the Variables tab.



Notice that many of the option boxes have small buttons on their right. These buttons may be used to activate a separate input window. For example, if you press the button to the right of the Variables box, the Variable Selection window will appear. This window will help you select the variables to be used.



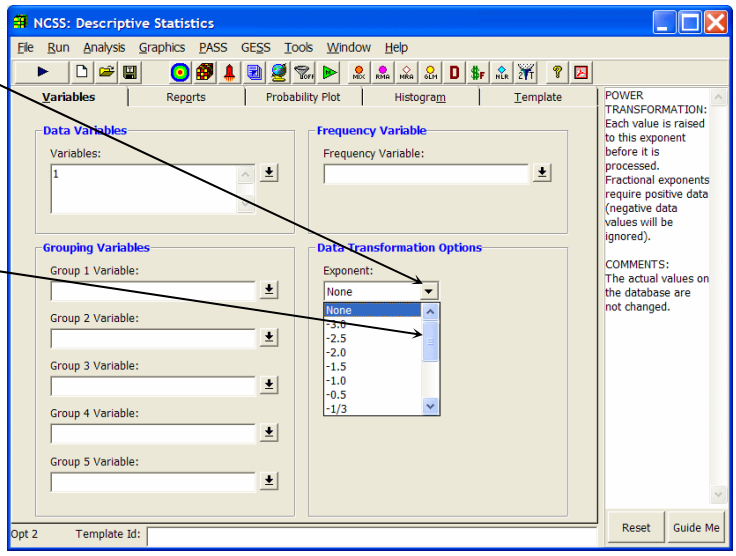
Changing an Option

Suppose you want to change the Exponent option from **None** to **3**.

1. Press the drop-down button on the right of the ***Exponent*** box.

This will activate the drop-down menu.

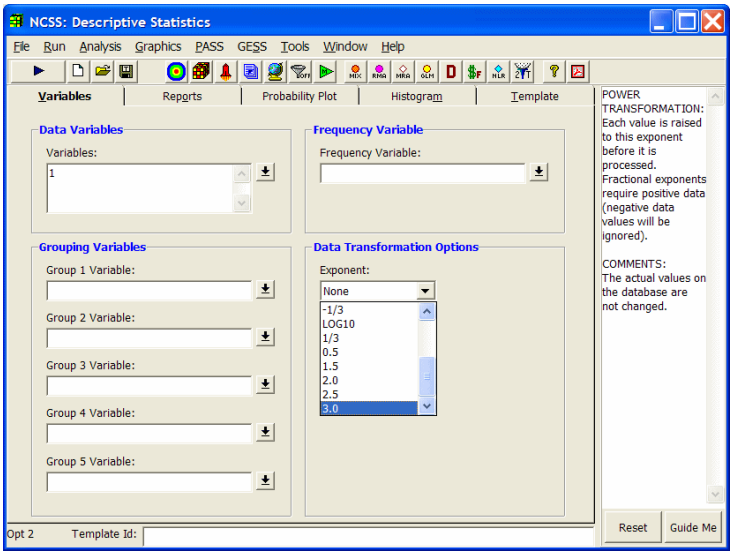
2. Move the scroll bar thumb down until the **3.0** appears.



3. Move the cursor down so that the **3.0** is highlighted.

4. Select the **3.0** by clicking it (or by pressing the Enter key while the 3.0 is highlighted).

Another way to change this option is to select it and press 3. The program searches through the options for the first item that begins with a 3.



Notes on Modifying Options

Many of the option boxes have alternative methods of entering data. For example, when you need to select a variable, you can type the variable name directly in the box or you can double click on the box to bring up a variable selection window.

Entering Text

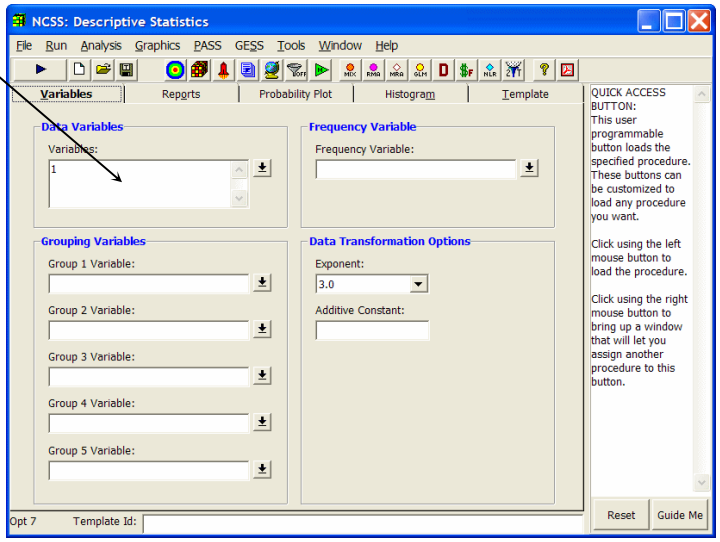
When an option needs text (such as the title of a graph), type the text directly into the box. Note that while you are typing, if you decide to revert back to the original text, you can hit the Escape (Esc) key.

Selecting Variables

When you need to specify variables, you can type their names directly into the box, you can enter their numbers directly into the box, or you can activate the variable selection window.

1. Double click in the **Variables** box.

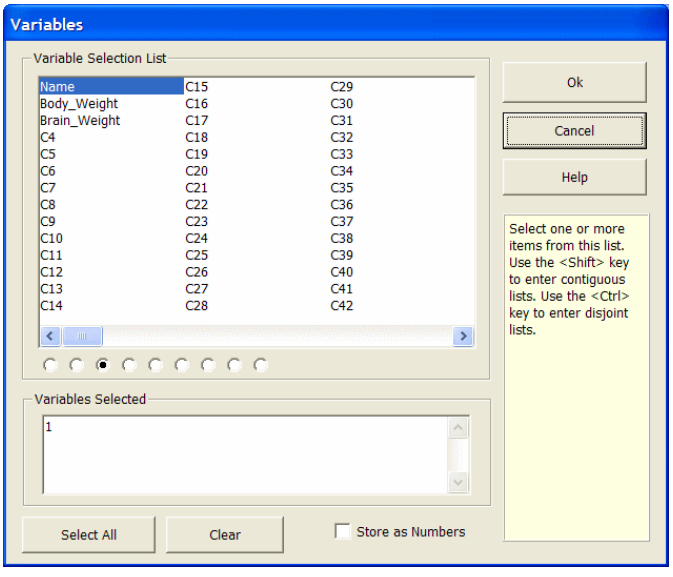
This will display the variable selection window. You can select the variables of interest and press the **Ok** button when you are finished.



Press the **Ctrl** key when you want to select several, noncontiguous, variables.

As you select variables in the Variable Selection List box, they will appear in the Variables Selected box at the bottom.

It may be convenient to specify variables by number rather than by name. For example, when you use numbers, you can use the same settings on several databases, even though the variables have different names.

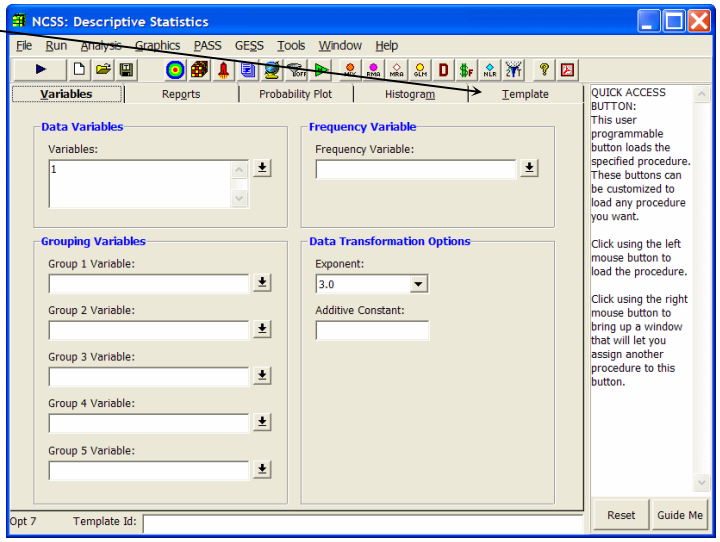


Saving a Template

Once you have filled out a procedure, you may want to save your choices so that you do not have to reset them again the next time you use the procedure. This is accomplished using the Template panel.

In this example, we will save the current settings to a file called TUTOR1.

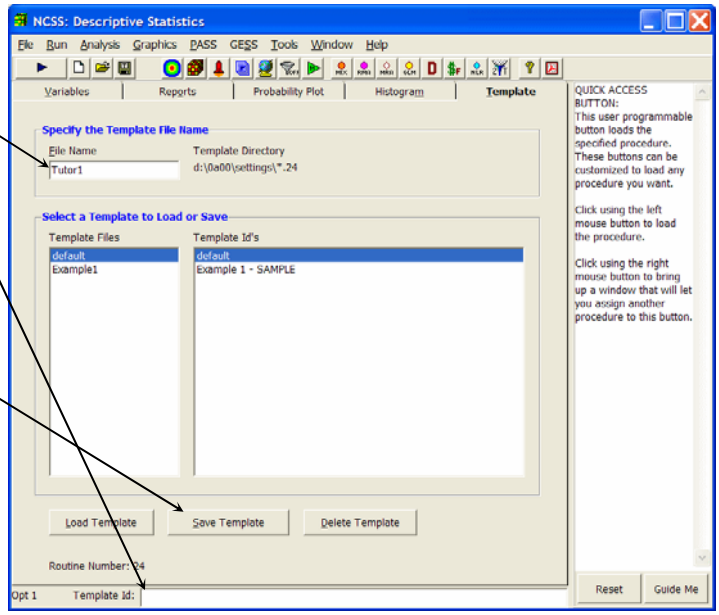
1. Press the **Template** tab to display the Template panel.



2. Enter **Tutor1** in the File Name box. This is the name where the template is stored.

3. Enter an identifying phrase in the Template Id box at the bottom of the screen.

4. Press the **Save Template** button to store the template file.



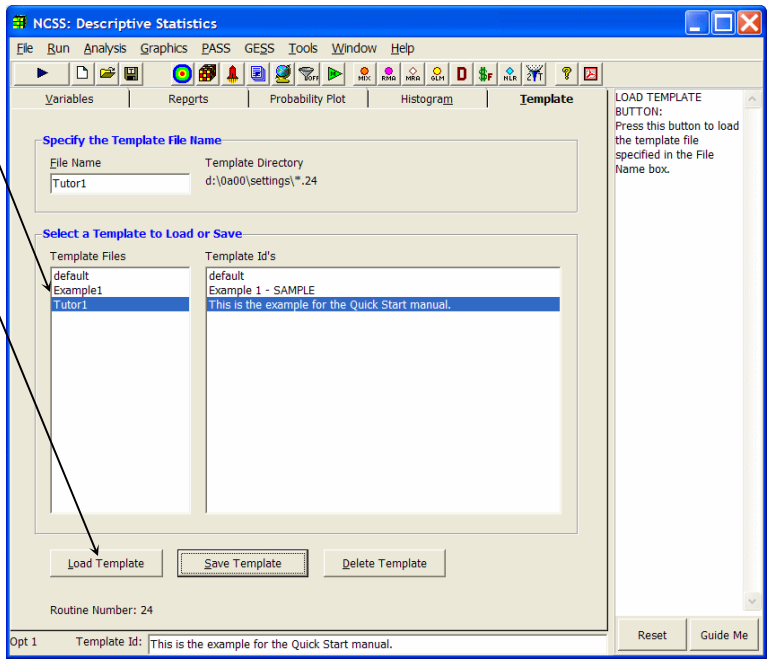
When you supply the template file name, you do not enter a three-character extension. **NCSS** adds the appropriate extension. This extension may be determined by looking at the Template Directory. In this example, the extension is the number 24. You can delete these files using your Windows Explorer program.

Loading a Template File

In this example, we will load the previously saved Tutor1 template file.

1. Select **TUTOR1** from the available template files.
2. Press the **Load Template** button.

   The settings are reset to how they were when Tutor1 was saved.



The Default Template

Whenever you close a procedure window, the current settings are saved in a template file named Default. When a procedure is loaded, **NCSS** checks to determine if the template file Default exists. If such a file exists, it is automatically loaded after the procedure window is loaded. Hence, the current settings of each procedure window are preserved between sessions. Because of this, you should avoid using Default as a template file name.

Chapter 9

Output Window

About This Chapter

NCSS sends all statistics and graphics output to its built-in word processor. In the word processor, the output can be viewed, edited, printed, or saved. Reports and graphs are saved in rich text format (RTF). Since RTF is a standard Windows document transfer format, these files may be loaded directly into your word processor for further processing. You can also cut data from the report and paste it into an **NCSS** datasheet for further analysis. This chapter covers the basics of our built-in word processor.

This chapter will continue the analysis of the brain weight data that was begun in Chapter 3. If you have not already done so, run the Descriptive Statistics reports as described in Chapter 4. Our analysis here will pick up where that chapter ended.

Viewing the Output

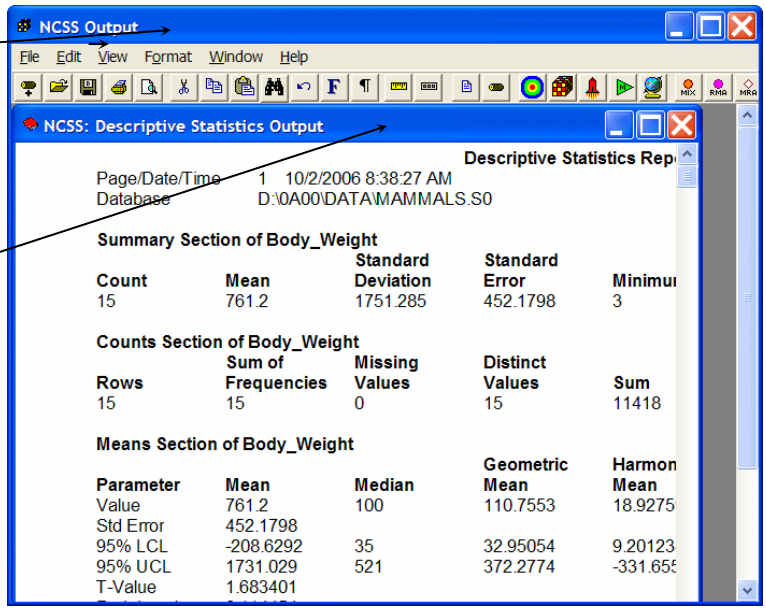
The output of the Descriptive Statistics program is shown below. Usually, you will find it useful to put the output window into full-screen mode.

1. Double click on the Output title bar.

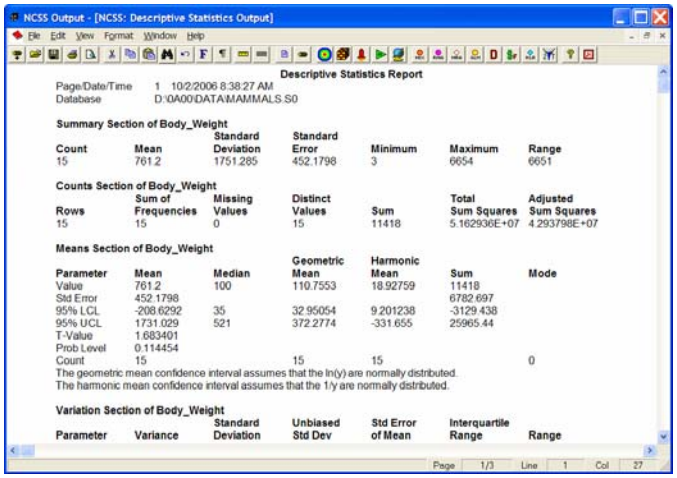
   This will put the word processor into full-screen mode.

2. Double click on the document title bar.

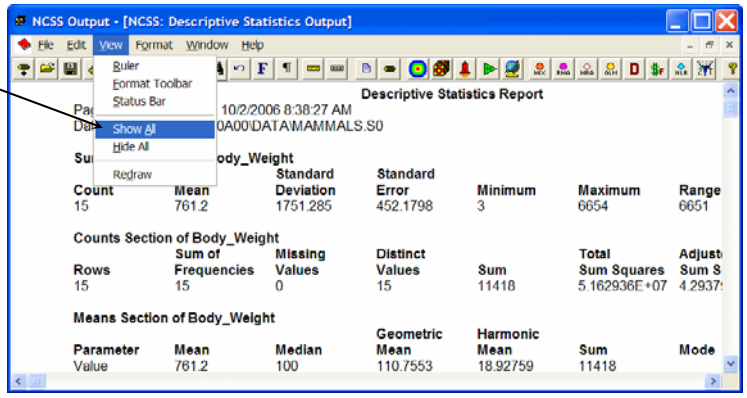
   This will put the document in full-screen mode also.



The screen will look similar to this. Note that the actual size of your screen depends on the resolution of your monitor, so it will vary.

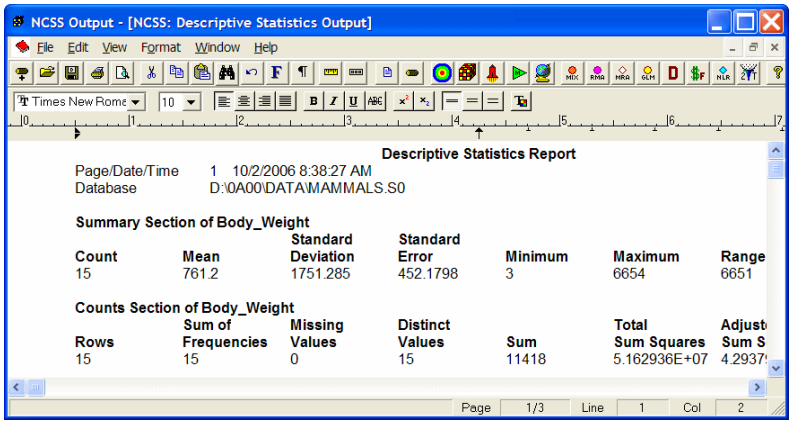


3. Select **Show All** from the View menu.



The screen will look similar to this.

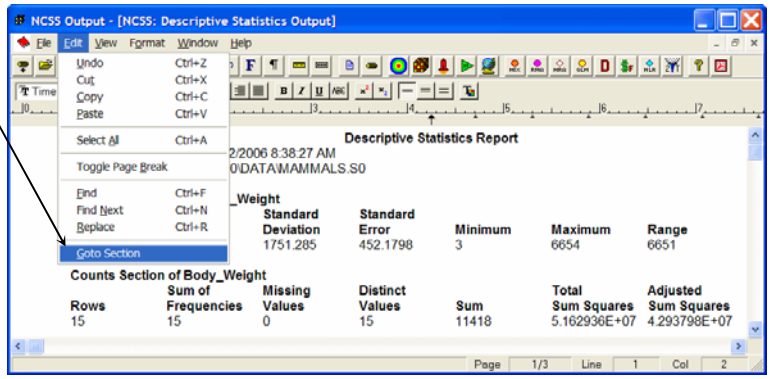
Notice the standard word processing ruler, tab bar, and button bar. These will aid you in editing your document.



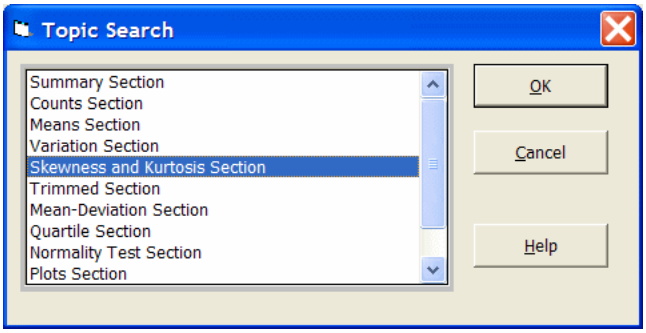
We will now show you a quick way to move about a lengthy document such as the current one.

1. Select **Goto Section** from the Edit menu.

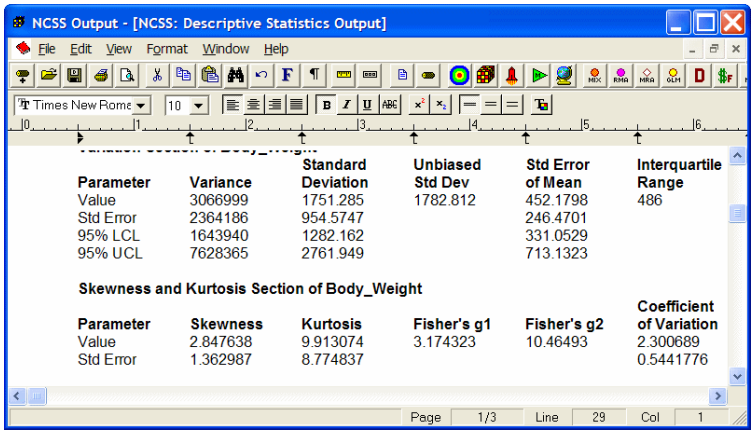
   This will load the Topic Search window.



2. Select **Skewness and Kurtosis Section**.
3. Press **OK.**



This will position the report so that the desired section title is showing.



At this point, you would scroll down through your output, perusing the results. Once you determine that you want to retain your results, you have four choices:

1. Print the document.
2. Save the document to a file.
3. Add the document to the log. (The log holds the output from several analyses in one file.)
4. Copy the report to a temporary holding area (the Windows clipboard) and paste it into another application.

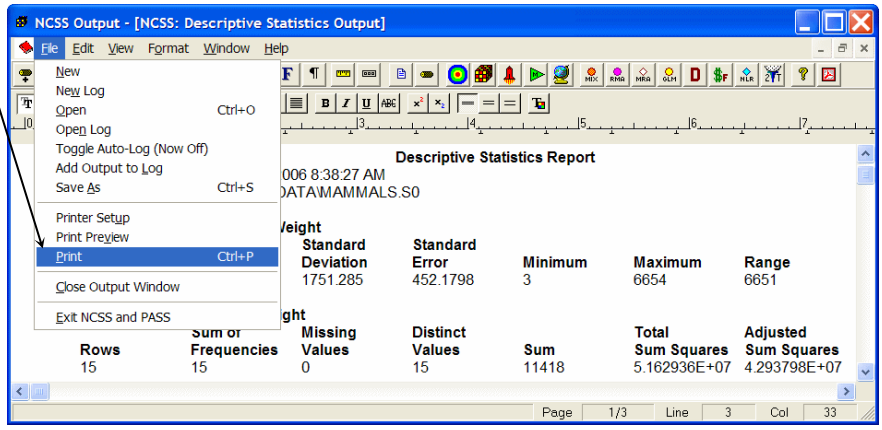
Note that you cannot just leave the output in the current window if you want to keep it because it will be replaced by the next analysis that is run.

Printing the Output

Before printing the report, you should scroll through it to determine if there are any portions that you want to delete before printing. To print the report, take the following steps.

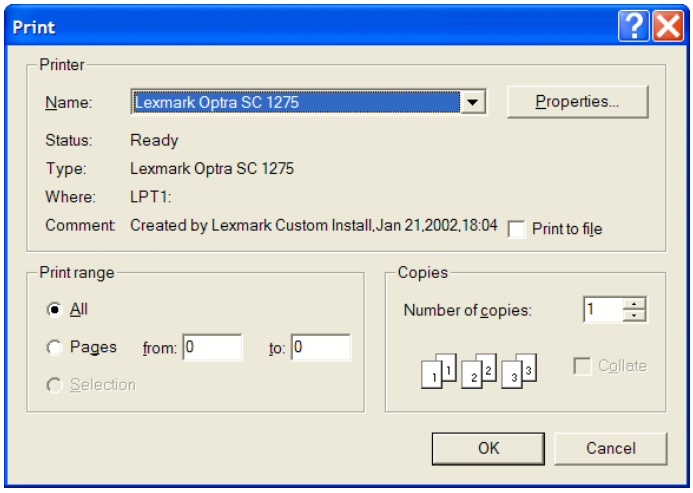
1. Select **Print** from the File menu.

This will bring up the Print Dialog box.



2. Click **OK** to begin printing your report.

This dialog box may appear different in different versions of Windows. However, the basic functionality will be the same.



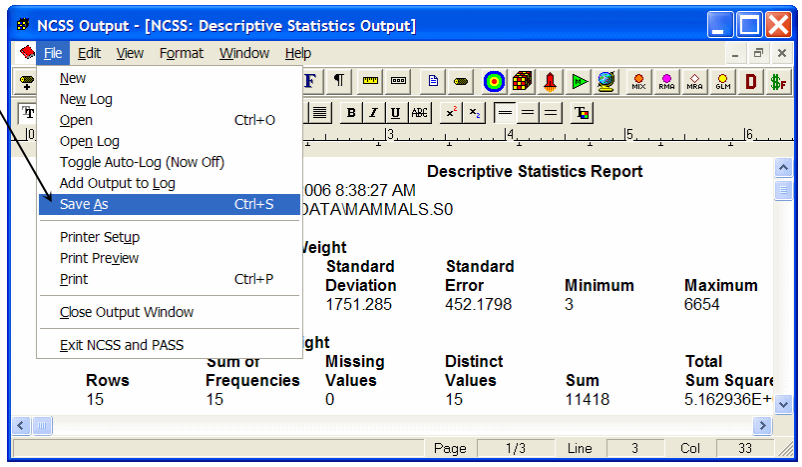
Saving the Output to a File

You can save the output to a file. The report is saved in rich text format (RTF) which is a standard document interchange format. This format may be read into commercial word processors such as Word and Word Perfect. This will allow you to export the reports to your favorite word processor.

Take the following steps to save the output to a file.

1. Select **Save As** from the File menu.

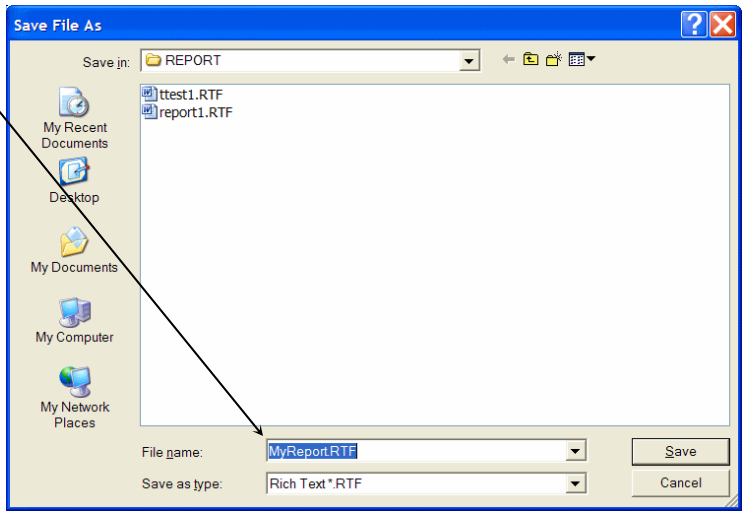
This will bring up the Save File As dialog box. Note that this dialog box may look different in Windows 95, but the basic functionality will be the same.



2. Enter a file name such as **MyReport.rtf**.

Note that the file name must end with the extension “rtf.”

3. Click **Save** to save your report.



Saving the Output to the Log

An analysis of a set of data usually requires the running of several statistical procedures. The *log* document provides a convenient way to store the output from several procedures together in one file. When you have a report or graph you want to keep, copy it from the output document to the log document.

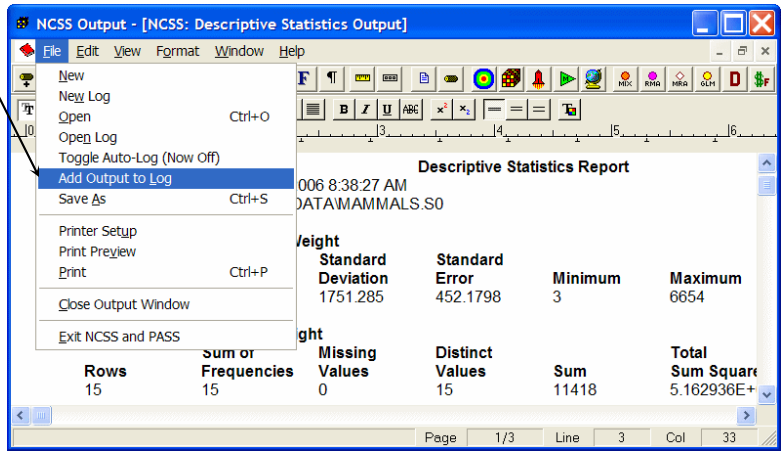
The log document provides four main word processing functions: loading, editing, printing, and saving. When you load a file into the log document, you can add new output to it. In this way, you can record your work on a project in a single file, even though your work on that project is may be spread out over a large time period.

Take the following steps to add the current output to the log document.

1. Select **Add Output to Log** from the File menu.

   This will copy the current document to the log file.

   To view the log document, select Log from the Window menu.



The log document resides in memory until you store it. To store the log document, take the following steps:

1. Select **Log** from the Window menu so that the log document is active.
2. Select **Save As** from the File menu and complete the Save File As dialog.

Warning: The log document is not automatically stored. You must store the contents of the log document to a file before exiting **NCSS**.

Chapter 10

Filters

About This Chapter

This chapter explains how to use *filters* to limit which rows (observations) are used by a particular procedure and which are skipped. For example, you might want to limit an analysis to those weighing over 200 pounds. You would use a filter to accomplish this.

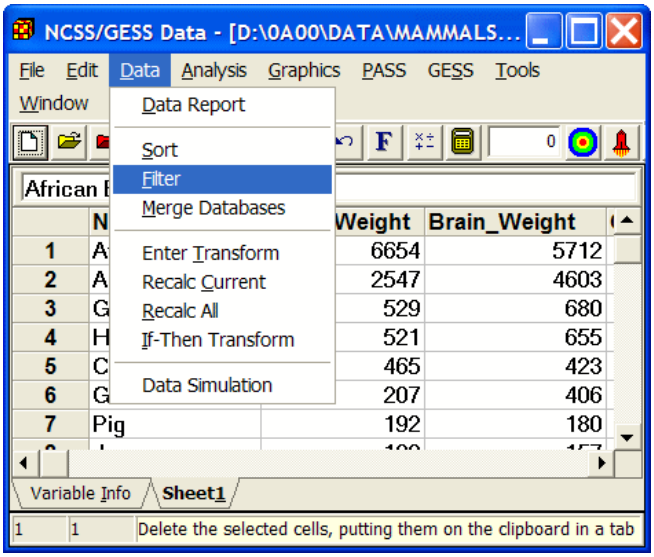
Setting Up a Filter

Using the MAMMALS database (see Chapter 3), we will setup up a filter so that only those animals with a body weight greater than 200 kilograms are used in the statistical calculations.

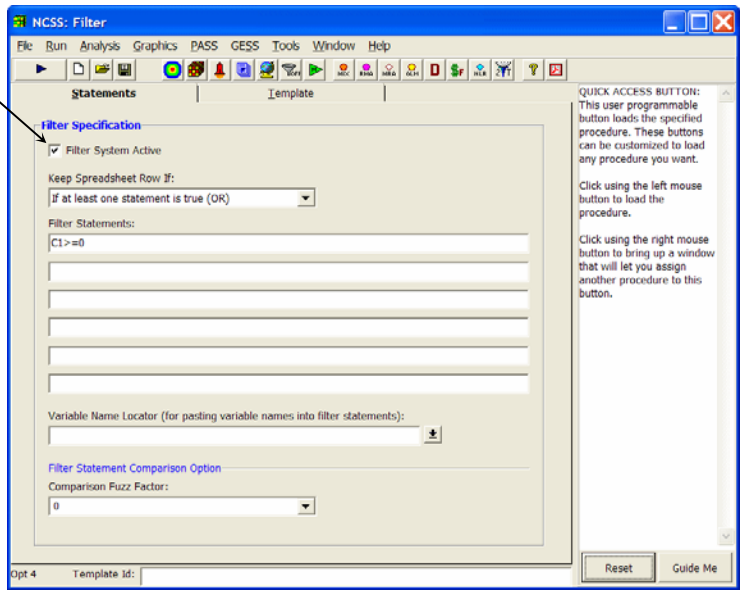
If the MAMMALS database is not currently loaded, select Open from the File menu, move to the \NCSS2007\DATA subdirectory, and double click on the file MAMMALS.S0. Your display should appear as follows.

1. Select **Filter** from the Data menu.

This brings up the Filter template.

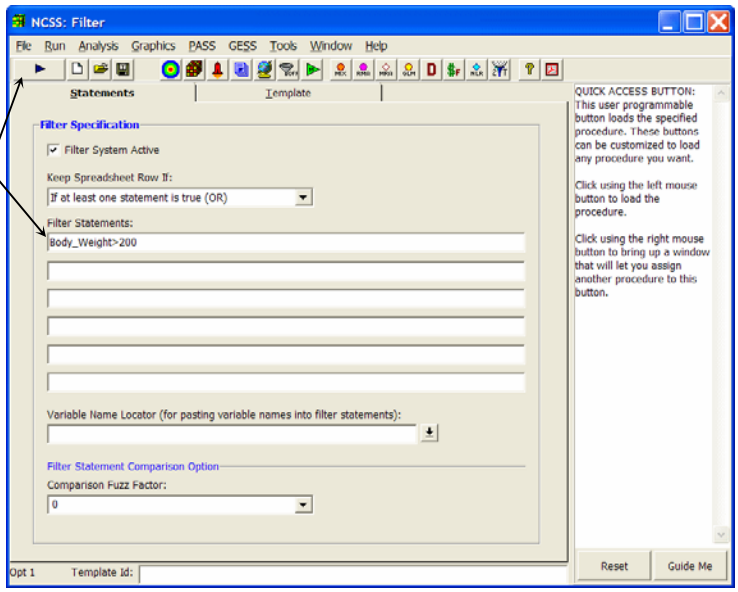


2. Check the **Filter System Active** box.

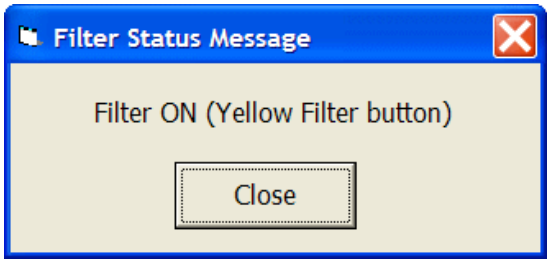


3. Enter the filter condition, **Body_Weight>200**, in the Filter Statements box.

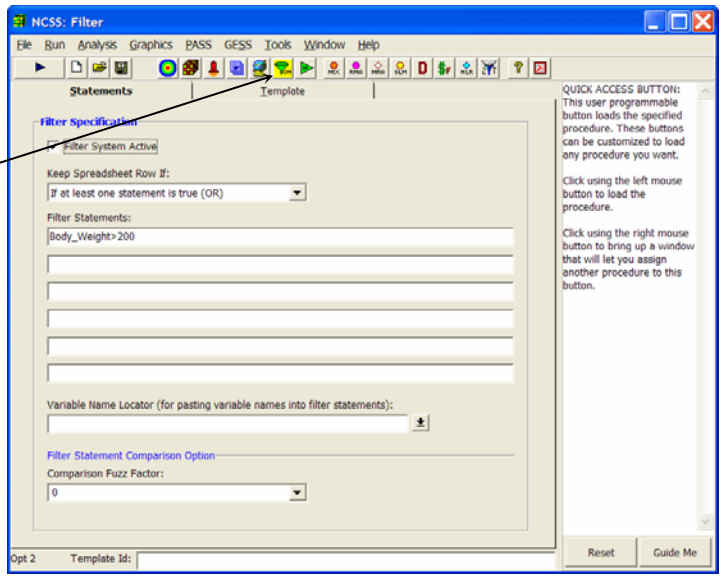
4. Press the **Run** button to activate the filter.



5. The Filter activated box will be displayed. You may press **Close**, or the message will automatically disappear.



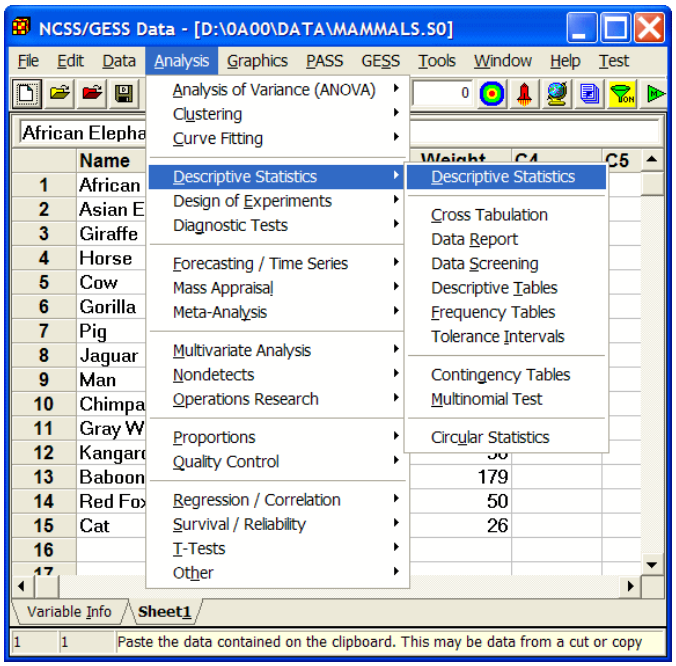
The filter is now setup. Notice that the Filter button on the both the Data and Filter toolbars has now changed to a green funnel with a yellow background and the word ON below it. This is a reminder that the filter system is active.



Using a Filter

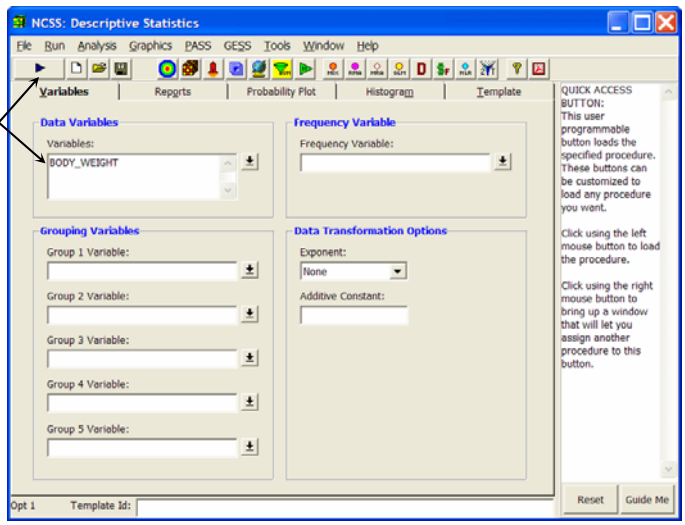
We will now show you how to use it in a procedure by obtaining the mean and standard deviation of the filtered database.

1. Open the **Descriptive Statistics** procedure by going to Analysis, Descriptive Statistics, Descriptive Statistics.



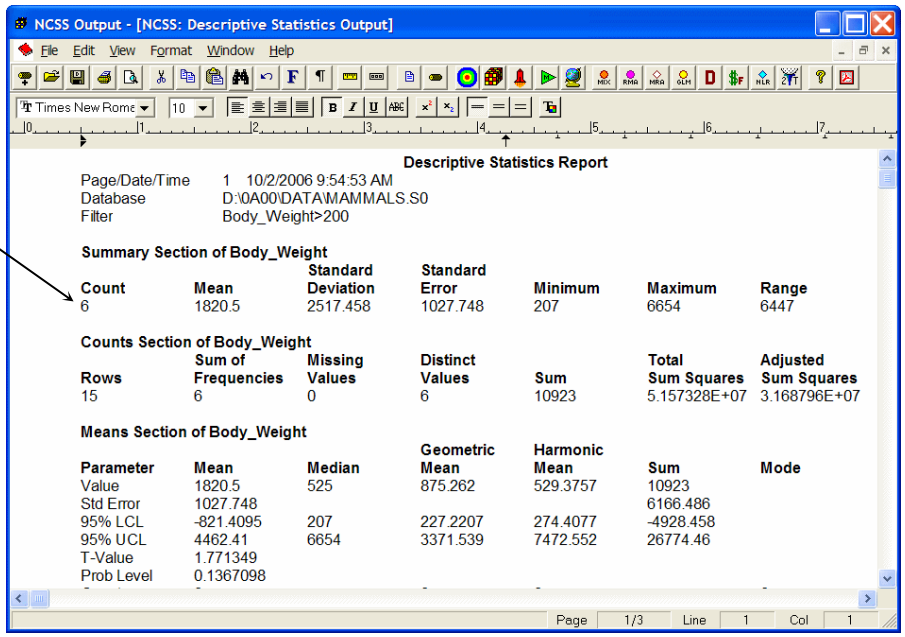
2. Enter **Body_Weight** in the Variables box.

3. Press the **Run** button to run the procedure.



4. Finally, view the output.

Notice that although fifteen rows were processed, only six rows were actually used in the computations.



Disabling the Filter

When you are finished using a filter, you can bring up the Filter procedure window, click the **Filter System Active** button so that it is not checked, and press the **Run** button to run the filter procedure. This will deactivate the filter.

Chapter 11

Writing Transformations

About This Chapter

The basics of entering transformations were covered in Chapter 3. This chapter gives examples of how to write more advanced transformations.

Recoding

Data *recoding* refers to replacing one set of values with another. For example, suppose you have each individual's age stored in a variable called AGE. Suppose that you want to create a new variable called AGEGROUP that classifies each individual into one of four age groups according to the following rule:

AGE Values	AGEGROUP Value
1 to 12	1
13 to 19	2
20 to 29	3
30 and above	4

Example of *Recode*

RECODE(Age; (1:12 = 1) (13:19 = 2) (20:29 = 3) (Else = 4))

Notice the basic syntax of this function. The variable being recoded is given first (here Age). Next, a set of statements that define the recoding are given.

Example Results

	Age	AgeGroup	C3	C4	C5
1	23	3			
2	15	2			
3	5	1			
4	33	4			
5	19	2			
6	46	4			
7	22	3			
8					

Basic Indicator

Indicator transformations are used in logic (if - then) situations. An indicator function evaluates to one if the condition is true, or to zero if the condition is false. The basic syntax is two arguments between parentheses separated by a logic operator. The possible logic operators are <, >, <=, >=, <>, and =.

Example of Indicator

(AGE > 20)

If AGE is greater than 20, the result will be a one. Otherwise, the result will be a zero.

Example Results

	Age	Indicator	C3	C4	C5
1	23	1			
2	15	0			
3	5	0			
4	33	1			
5	19	0			
6	46	1			
7	22	1			
8					

Compound Indicators

Since indicator functions evaluate to a numeric value (either 0 or 1), they may be combined with other functions, including other indicator functions. When combining several indicators, the logical AND is achieved by multiplying the indicators and the logical OR is achieved by adding.

Example of Compound Indicator

(AGE > 20)*(AGE<=40)

If age is greater than 20 *and* less than or equal to 40, the result will be a one. Otherwise, the result will be a zero.

Example Results

	Age	Indicator	C3	C4	C5
1	23	1			
2	15	0			
3	5	0			
4	33	1			
5	19	0			
6	46	0			
7	22	1			
8					

Using Indicators for If – Then Transformations

Indicator functions may be used in place of *if - then* statements. The following examples show how this is done.

Example 1 of If-Then Indicator

If Age is less than 20 set AdjIncome to 5000. Otherwise, set AdjIncome equal to Income.

(Age < 20)*5000+(Age>=20)*Income

Note that the indicator functions used here are opposites. When (Age<20) is 0, (Age>=20) will be equal to 1.

Example 1 Results

	Age	Income	AdjIncome	C4	C5
1	23	22000	22000		
2	15	5500	5000		
3	5	100	5000		
4	33	35400	35400		
5	19	9000	5000		
6	46	54000	54000		
7	22	6000	6000		
8					
9					

It may be helpful to look at how this expression works on the first two rows.

Calculation for the first row:

(23<20)*5000+(23>=20)*22000 = 0(5000) + 1(22000) = 22000

Calculation for the second row:

(15<20)*5000+(15>=20)*22000 = 1(5000) + 0(5500) = 5000

Example 2

If Age is less than 20 set AdjIncome equal to Income + 1000. Otherwise, set AdjIncome to Income + 2000.

(Age < 20)*(Income+1000)+(Age>=20)*(Income+2000)

Example 2 Results

	Age	Income	AdjIncome	C4	C5
1	23	22000	24000		
2	15	5500	6500		
3	5	100	1100		
4	33	35400	37400		
5	19	9000	10000		
6	46	54000	56000		
7	22	6000	8000		
8					
9					

It may be helpful to look at how this expression works on the first two rows.

Calculation for the first row:

(23<20)*(22000+1000)+(23>=20)*(22000+2000) = 0(23000) + 1(24000) = 24000

Calculation for the second row:

(15<20)*(5500+1000)+(15>=20)*(5500+2000) = 1(6500) + 0(7500) = 6500

Chapter 12

Importing Data

About This Chapter

This chapter presents an example of importing data from a comma delimited ASCII (text) file into **NCSS**.

The ASCII File

Following is a set of data contained in the file ASCII.TXT in your \DATA subdirectory. We will now go through the steps necessary to import the data from this file.

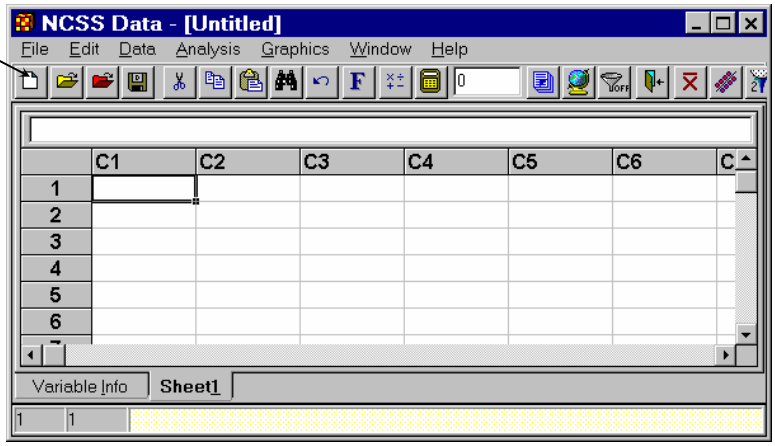
Bob,2,4,22,5
Judy,5,44,22,4
Sam,1,32,42,9
Mary,4,1,22,23
John,19,22,44,1
Linda,3,11,2,14

How to Import ASCII.TXT

1. Press the **New Database** button on the toolbar.

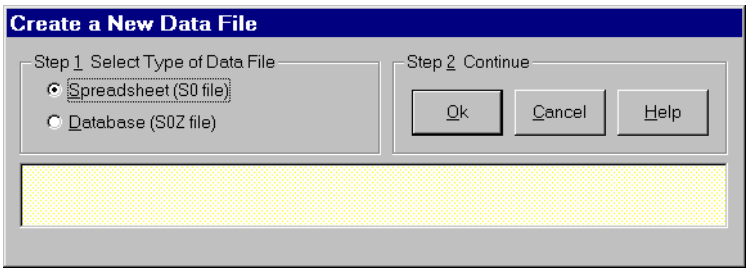
It is necessary to clear the previous database. Otherwise, the imported data would be added to it.

A dialog box, entitled **Create a New Data File**, will appear.

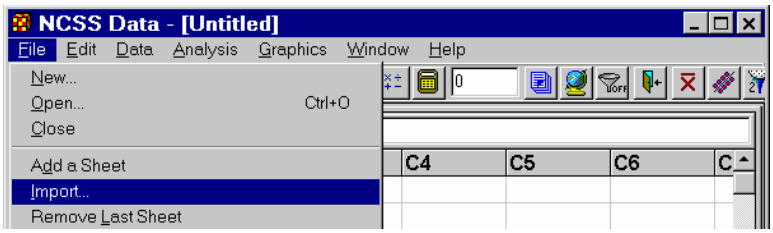


2. Indicate that you want a Spreadsheet-type data file since this is a small set of data.

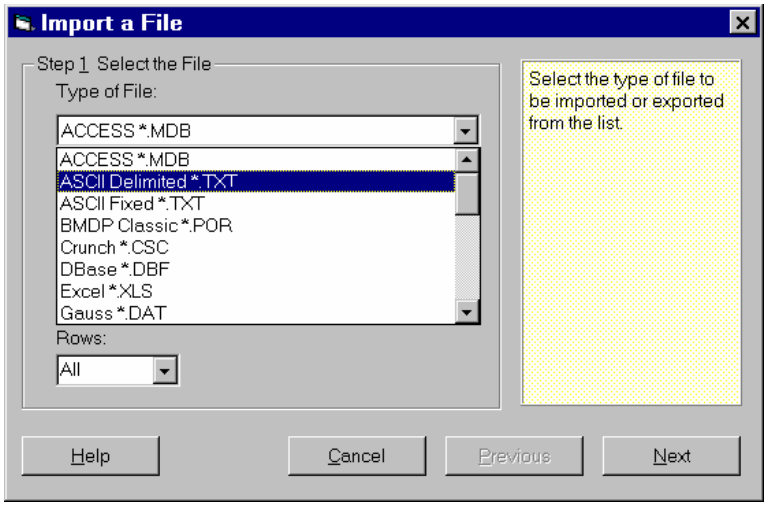
Since this is the default, just click **OK**.



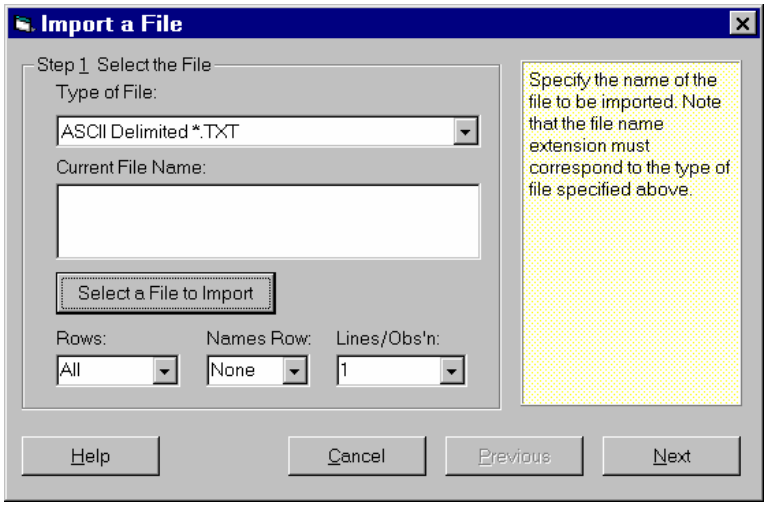
3. Select **Import** from the File menu.



4. Select **ASCII Delimited *.TXT** from the Select the File Type selection box.

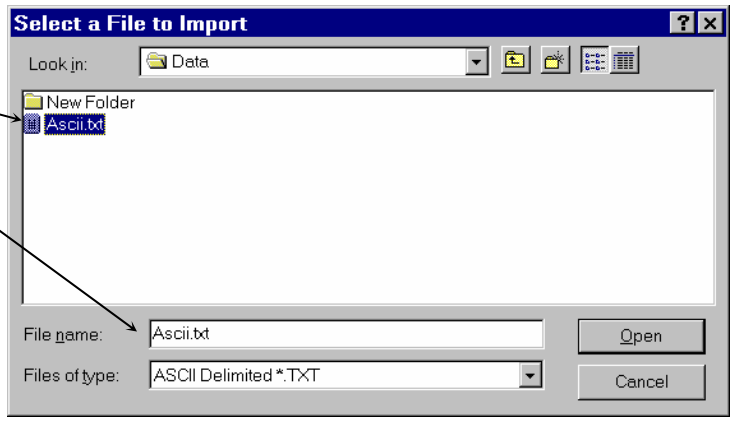


5. Press the **Select a File to Import** button to specify the file name.

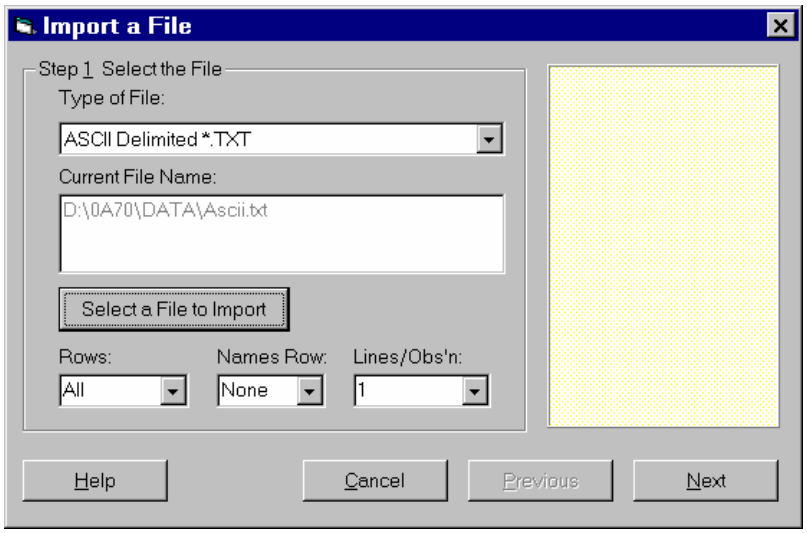


6. Click on the **Ascii.txt** in the Data directory to specify the desired file.

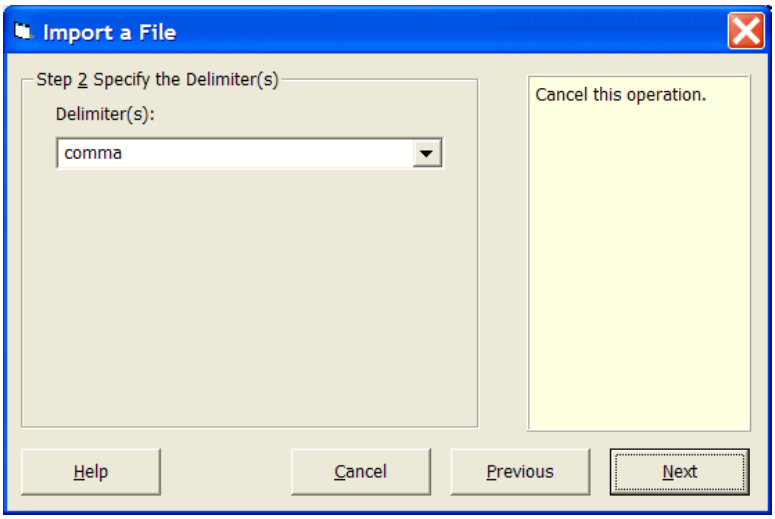
7. Press the **Open** button to finish selecting the file.



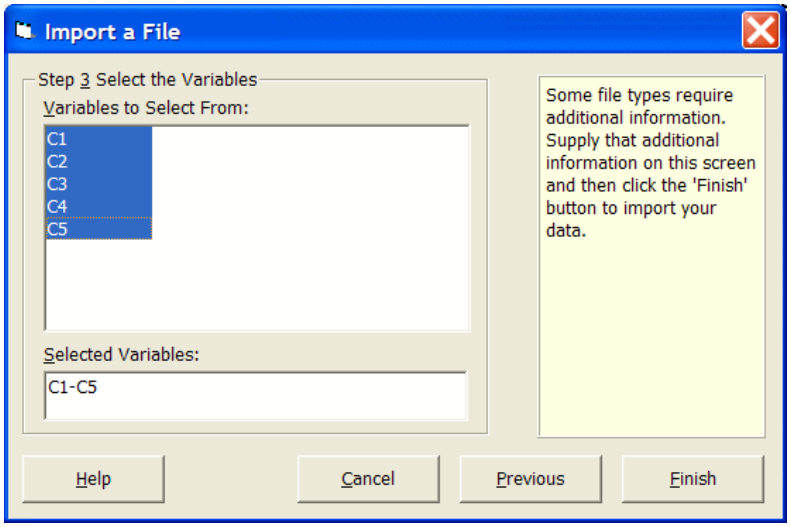
8. Press the **Next** button to move on to the next import screen.



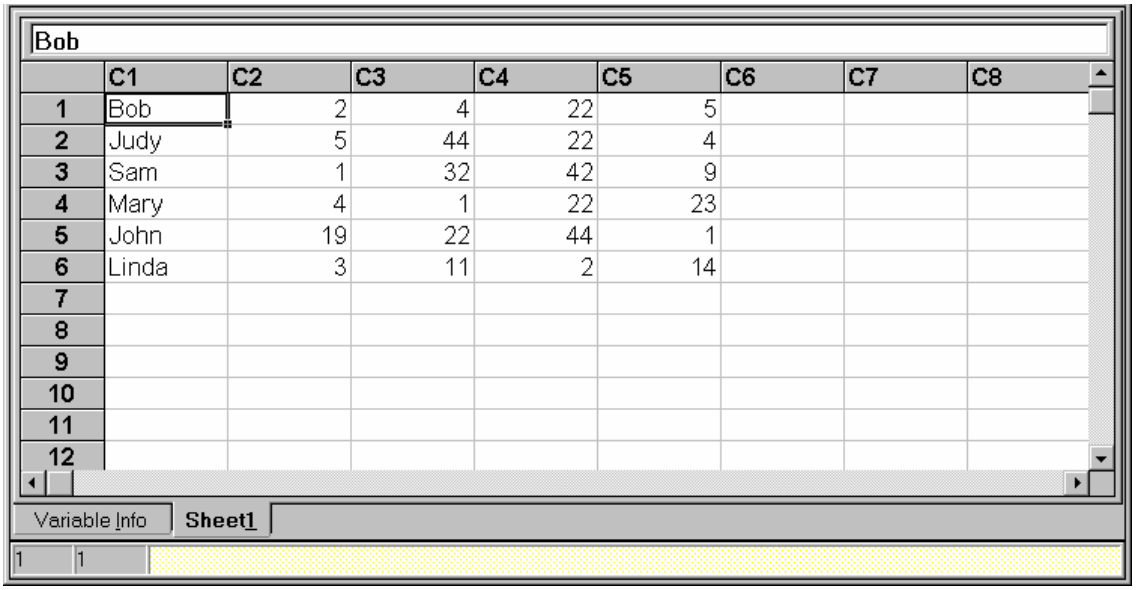
9. Since the correct delimiter (a comma) is specified, you are ready to continue. Press the **Next** button to move on to the next import screen.



10. Since we want to import all the data, we leave all the variables **C1 – C5** selected. Click the **Finish** button to begin the import.



The imported data will appear in the Data window.



Note that the imported database resides in your computer's memory, not on the hard disk. If you want to make a permanent copy of your data, you should select **Save As** from the File menu and save a copy of the imported data to your hard disk.

Chapter 13

Value Labels

About This Chapter

Value Labels provide a mechanism to attach labels to coded data. For example, in a questionnaire you might have questions whose responses fall along a Likert scale. Perhaps you have entered the data as numeric values from 1 to 5. Value labels may be attached to the responses so that 1 shows up on your printout as “Strongly Agree” and 5 is displayed as “Strongly Disagree.”

This chapter will provide you with a step by step outline of how to use value labels. The data for this example come from a four-item questionnaire that was given to twenty people as part of a political poll. The first three questions contain demographic information about the individual. The fourth question is their opinion about a hot political issue. You will find these data in the POLITIC database. The data were coded numerically for easy data entry as follows:

POLITIC Database

AgeGroup
- 1 = 25 and under
- 2 = 26 to 34
- 3 = 35 to 55
- 4 = 56 and above

State
- 1 = California
- 2 = Virginia
- 3 = Texas
- 4 = Other

Party
- 1 = Democrat
- 2 = Republican
- 3 = Other

Issue
- 1 = Strongly agree
- 2 = Agree
- 3 = Neutral
- 4 = Disagree
- 5 = Strongly disagree

	AgeGroup	State	Party	Issue	C5
1	1	1	3	4	
2	4	2	2	2	
3	3	2	1	3	
4	2	1	1	4	
5	2	4	2	5	
6	4	4	2	4	
7	2	3	1	1	
8	4	1	3	2	
9	2	2	2	1	
10	2	1	1	3	
11	1	4	3	2	
12	1	3	2	2	
13	3	3	2	1	
14	2	2	1	4	
15	3	1	2	5	
16	1	2	1	5	
17	4	2	3	2	
18	4	1	3	1	
19	2	4	1	1	
20	3	4	2	3	
21					

Adding the Value Labels

The next step is to add the value labels to the database. This is done by entering the values and corresponding labels in adjacent columns of the database. Leaving space for additional response variables, we put the value labels in columns 15 through 22. C15 contains the values of AgeGroup, C17 contains the values of State, and so on.

Note that we have resized the column widths to make the display easier to read (C15, C17, C19, and C21 are narrower than usual).

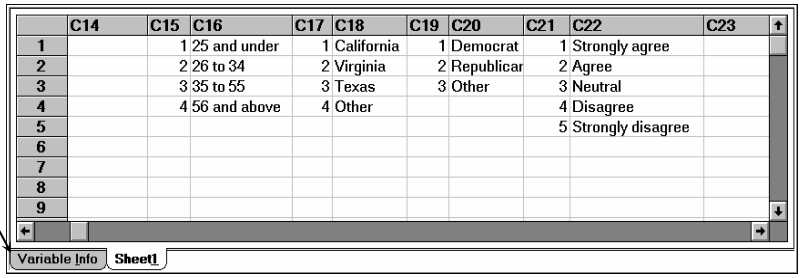
Although in this example we are constructing value labels for each variable, you do not have to do this. You can label as many or as few variables as you like.

	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23
1		1	25 and under	1	California	1	Democrat	1	Strongly agree	
2		2	26 to 34	2	Virginia	2	Republicar	2	Agree	
3		3	35 to 55	3	Texas	3	Other	3	Neutral	
4		4	56 and above	4	Other			4	Disagree	
5								5	Strongly disagree	
6										
7										

Attaching the Value Labels to the Variables

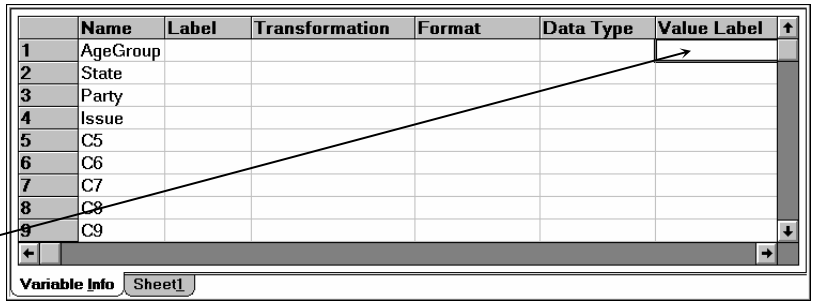
The final step is to attach the value-label columns to the appropriate variables. This is accomplished as follows:

1. Click the **Variable Info** tab.



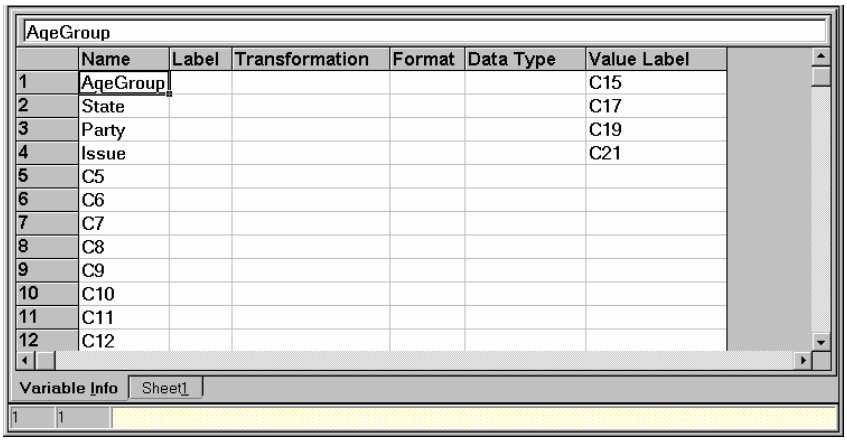
2. Use the **vertical scroll bar** or the **Page Up** key to reposition the view to the top of the Variable Info datasheet.

3. Click in the first cell under **Value Labels** to set the spreadsheet cursor there.



4. Type **C15.**
   Press **Enter.**
   Type **C17.**
   Press **Enter.**
   Type **C19.**
   Press **Enter.**
   Type **C21.**
   Press **Enter.**

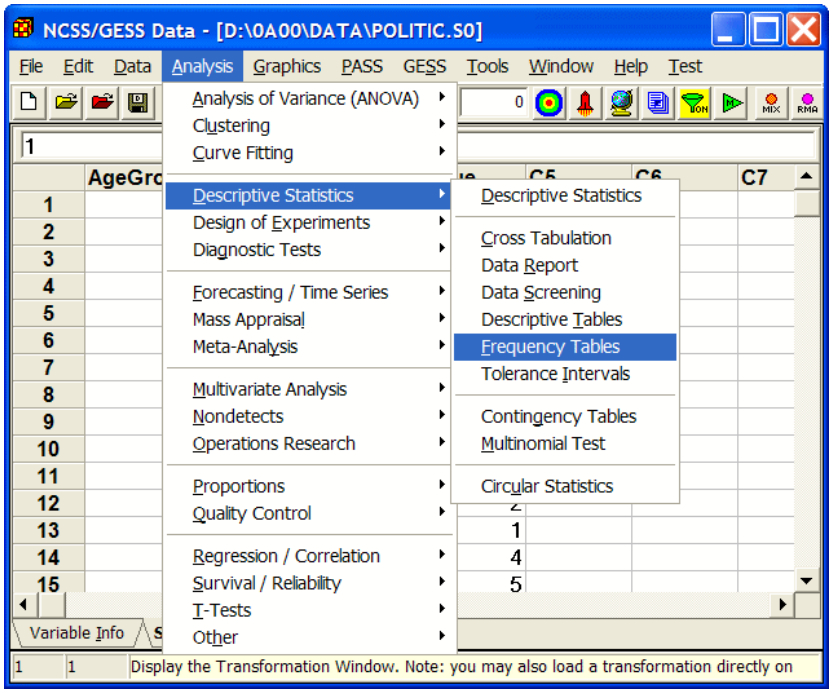
   This attaches each value label column to the appropriate variable. Note that you may use the same value label variable more than once.



Using the Value Labels in a Report

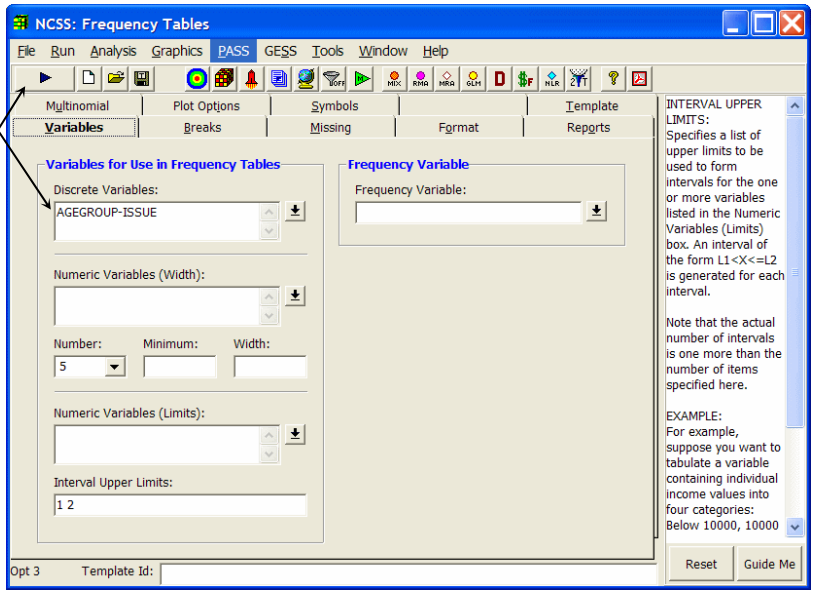
We will now show you how the value labels change the output of the Frequency Table procedure.

1. Select **Frequency Tables** from the Descriptive Statistics menu.



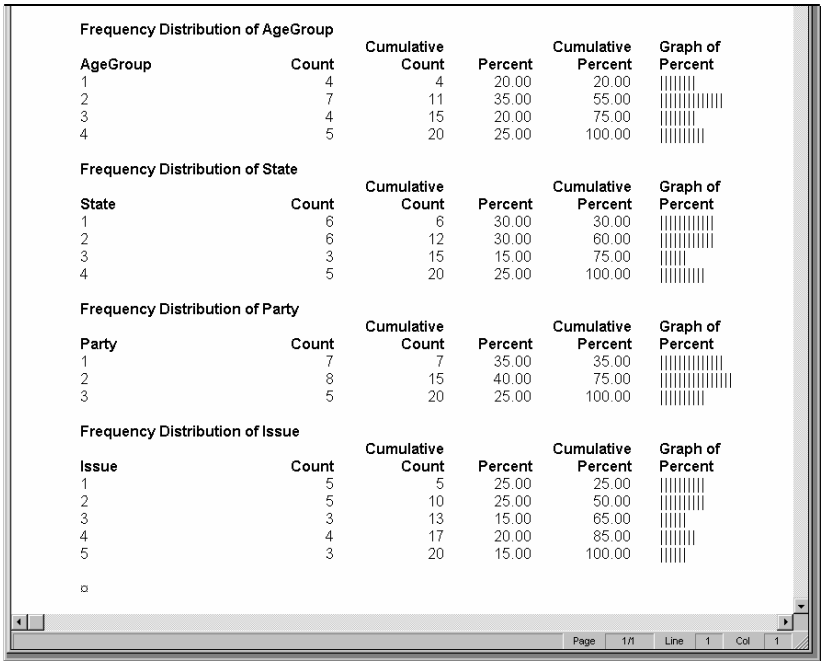
2. Enter **AgeGroup-Issue** as the Discrete Variables.

3. Run the procedure by pressing the **Run** button.



The output appears as shown.

Notice that the value labels have *not* been used.



Frequency Distribution of AgeGroup

AgeGroup	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
1	4	4	20.00	20.00	|||||||
2	7	11	35.00	55.00	||||||||||||
3	4	15	20.00	75.00	|||||||
4	5	20	25.00	100.00	|||||||||

Frequency Distribution of State

State	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
1	6	6	30.00	30.00	||||||||||
2	6	12	30.00	60.00	||||||||||
3	3	15	15.00	75.00	|||||
4	5	20	25.00	100.00	|||||||||

Frequency Distribution of Party

Party	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
1	7	7	35.00	35.00	||||||||||||
2	8	15	40.00	75.00	||||||||||||||
3	5	20	25.00	100.00	|||||||||

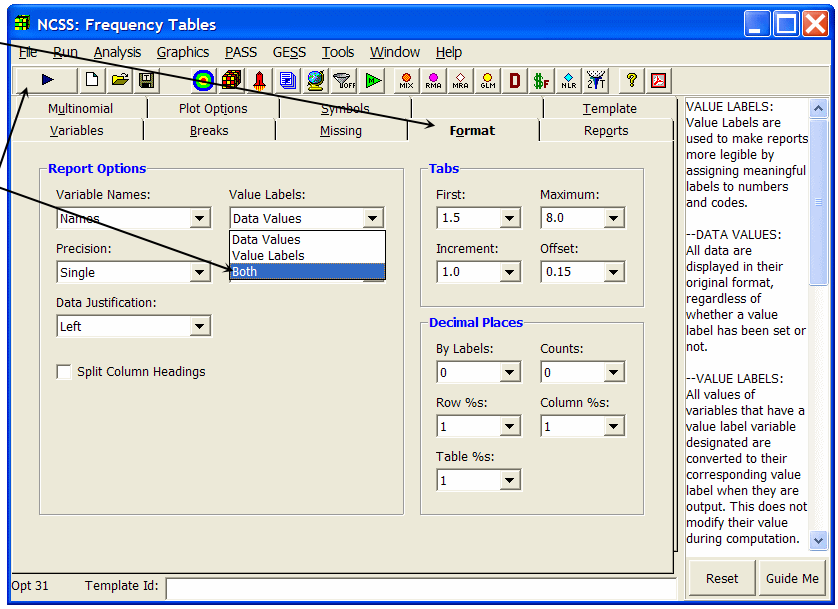
Frequency Distribution of Issue

Issue	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
1	5	5	25.00	25.00	|||||||||
2	5	10	25.00	50.00	|||||||||
3	3	13	15.00	65.00	||||||
4	4	17	20.00	85.00	||||||||
5	3	20	15.00	100.00	||||||

4. Select the **Format** tab to display the Format panel.

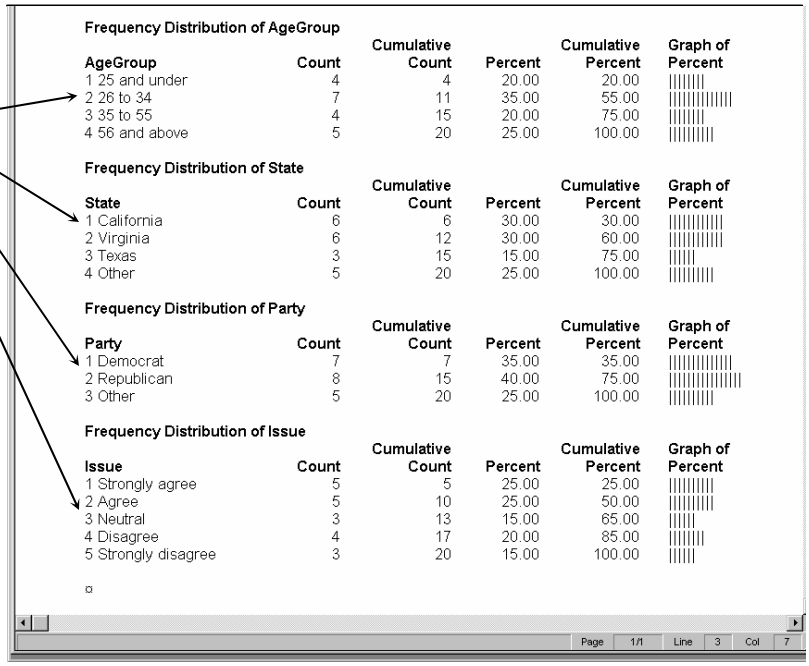
5. Select **Both** in the Value Labels box.

6. Run the analysis again by pressing the **Run** button.



The output window appears as shown.

Note that the value labels are now displayed.



Frequency Distribution of AgeGroup

AgeGroup	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
1 25 and under	4	4	20.00	20.00	|||||||
2 26 to 34	7	11	35.00	55.00	||||||||||||
3 35 to 55	4	15	20.00	75.00	|||||||
4 56 and above	5	20	25.00	100.00	|||||||||

Frequency Distribution of State

State	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
1 California	6	6	30.00	30.00	||||||||||
2 Virginia	6	12	30.00	60.00	||||||||||
3 Texas	3	15	15.00	75.00	|||||
4 Other	5	20	25.00	100.00	|||||||||

Frequency Distribution of Party

Party	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
1 Democrat	7	7	35.00	35.00	|||||||||||
2 Republican	8	15	40.00	75.00	|||||||||||||
3 Other	5	20	25.00	100.00	|||||||||

Frequency Distribution of Issue

Issue	Count	Cumulative Count	Percent	Cumulative Percent	Graph of Percent
1 Strongly agree	5	5	25.00	25.00	|||||||||
2 Agree	5	10	25.00	50.00	|||||||||
3 Neutral	3	13	15.00	65.00	|||||
4 Disagree	4	17	20.00	85.00	|||||||
5 Strongly disagree	3	20	15.00	100.00	|||||

Page 1/1 Line 3 Col 7

Chapter 14

Database Subsets

About This Chapter

It is often useful to store all of your data in one large database and then analyze various subsets of the database as necessary. This can often be accomplished using the Filter mechanism. Sometimes you will find it more convenient to create a subset of the original database that only contains those rows that you want to analyze.

This chapter will take you through the steps necessary to create a subset of the POLITIC database (described in Chapter 13) that contains only those individuals with AgeGroup equal to 2 (26 to 34).

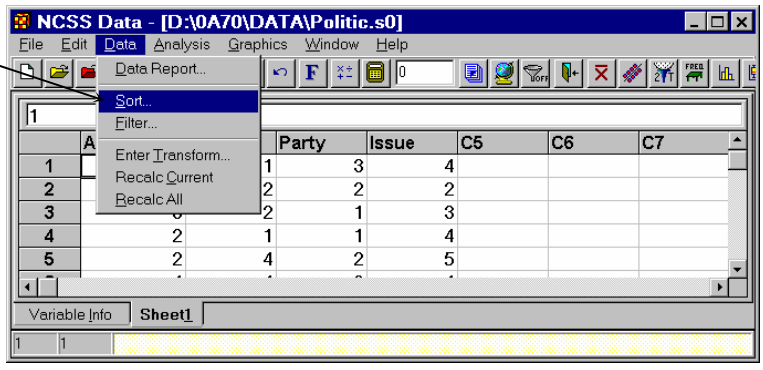
Creating a Database Subset

Use the following steps to create a database subset. If you have not already done so, open the POLITIC database now by selecting **Open** from the File menu of the Data window.

Step 1 – Sort the Database

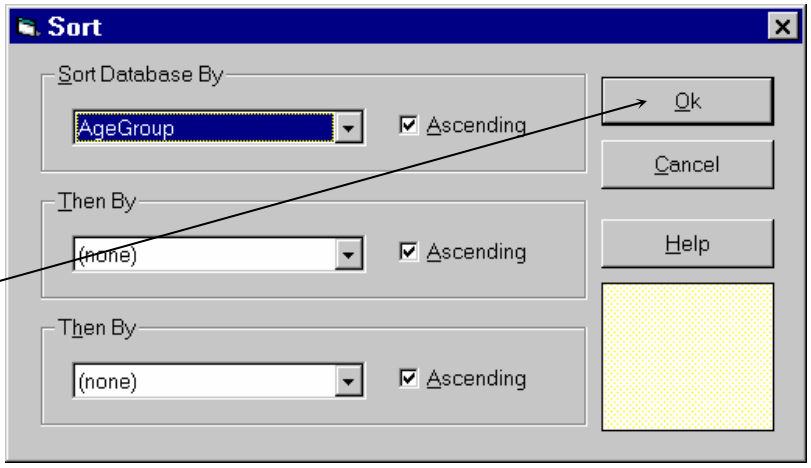
The first step is to sort the POLITIC database by the variable (or variables) that you want to subset on. This is done as follows.

1. Select **Sort** from the Data menu.



2. Select ***AgeGroup*** as the variable to sort the database by. This may be done by using the drop-down menu or by double clicking.

3. Click ***Ok*** to sort the database by the selected variable.

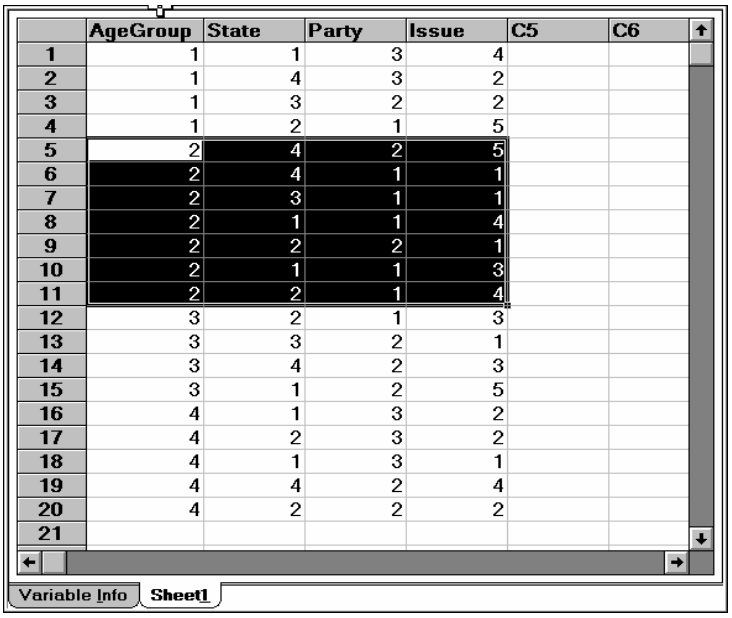


Step 2 – Copy Subset into New Database

The next step is copy the selected data from the POLITIC database to the new database (which will be named POLITIC2).

The database will be sorted by ***AgeGroup***.

1. Select the desired subset by dragging the mouse from row 5 of column 1 (AgeGroup) to row 11 of column 4 (Issue). Your selection should appear as shown.



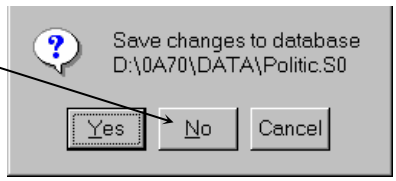
	AgeGroup	State	Party	Issue	C5	C6
1	1	1	3	4		
2	1	4	3	2		
3	1	3	2	2		
4	1	2	1	5		
5	2	4	2	5		
6	2	4	1	1		
7	2	3	1	1		
8	2	1	1	4		
9	2	2	2	1		
10	2	1	1	3		
11	2	2	1	4		
12	3	2	1	3		
13	3	3	2	1		
14	3	4	2	3		
15	3	1	2	5		
16	4	1	3	2		
17	4	2	3	2		
18	4	1	3	1		
19	4	4	2	4		
20	4	2	2	2		
21						

Variable Info | Sheet1

2. Press **Ctrl-C** to copy the selected data to the Windows clipboard (the clipboard is the name of temporary holding area used by Windows to store information that has been cut or copied).

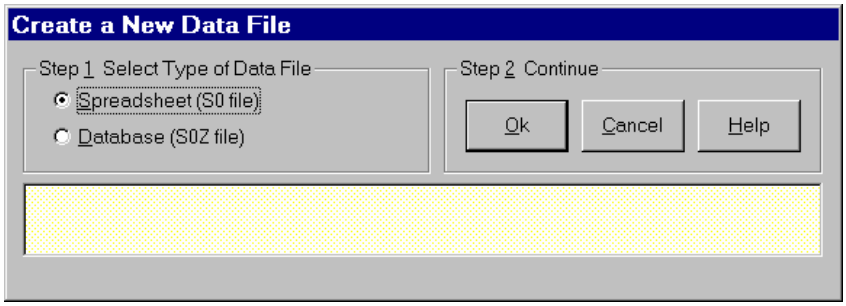
3. Select **New** from the File menu to create the subset database.

4. Select **No** from the message box that appears and asks if you want to save changes.



It is important not to save the sorted database because the value labels have also been sorted – something we do not want in this case.

5. Click **Ok** to create a spreadsheet type database.



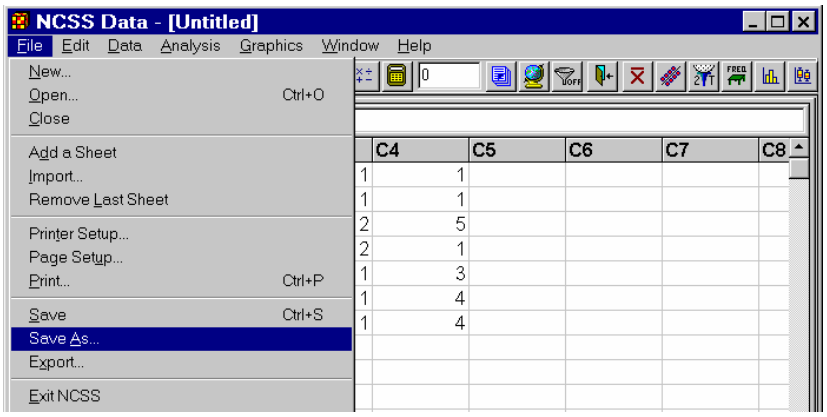
6. Position the cursor in the upper left cell of the new database by clicking in it.

	C1	C2	C3	C4	C5	C6
1						
2						
3						
4						
5						

7. Press **Ctrl-V** to paste the clipboard data into the new database.

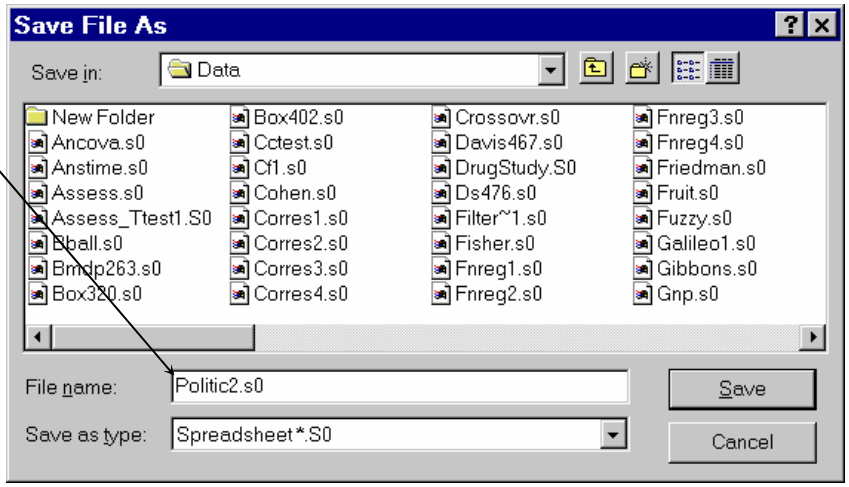
	C1	C2	C3	C4	C5	C6
1	2	4	2	5		
2	2	4	1	1		
3	2	3	1	1		
4	2	1	1	4		
5	2	2	2	1		
6	2	1	1	3		
7	2	2	1	4		
8						
9						

8. Select **Save As** from the File menu to name and save this new database.



9. Enter **Politic2.s0** as the name of the new database.

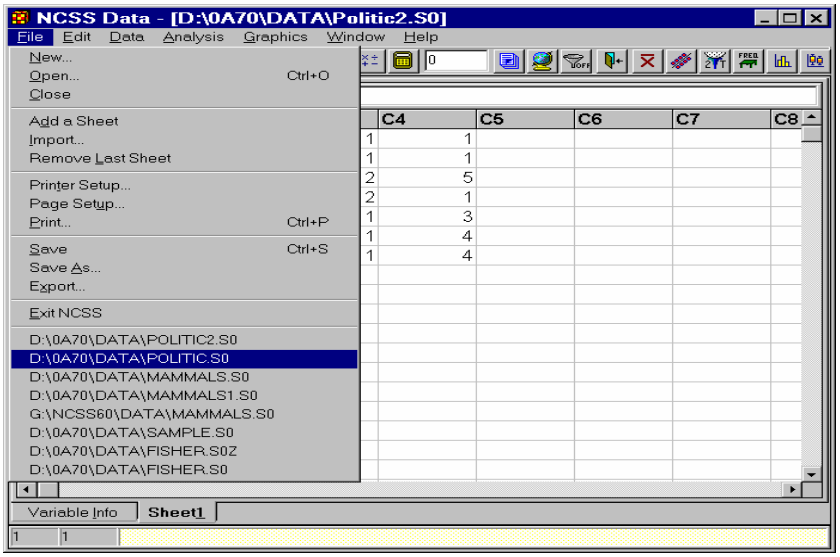
Click **Save**.



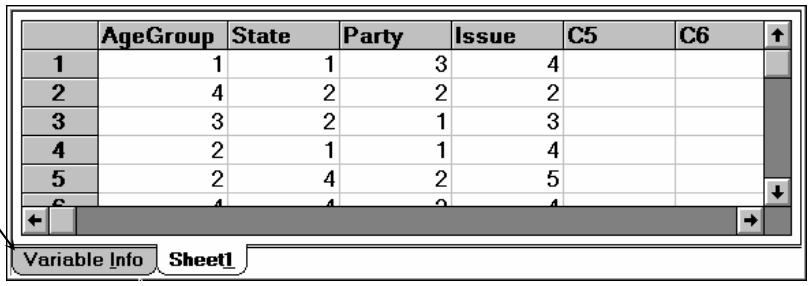
Step 3 – Copy Variable Info to New Database

The next step is to copy the variable information datasheet to the new database.

1. Open the **POLITIC** database by selecting it from the File menu.

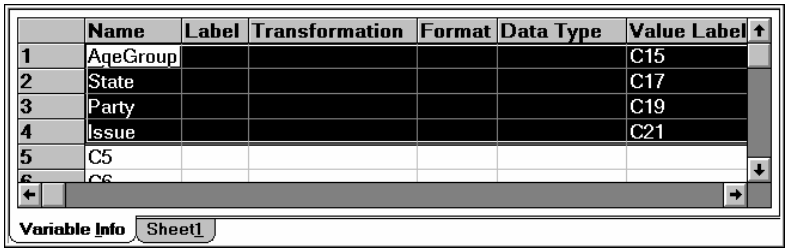


2. Click on the **Variable Info** tab to move to the Variable Info datasheet.

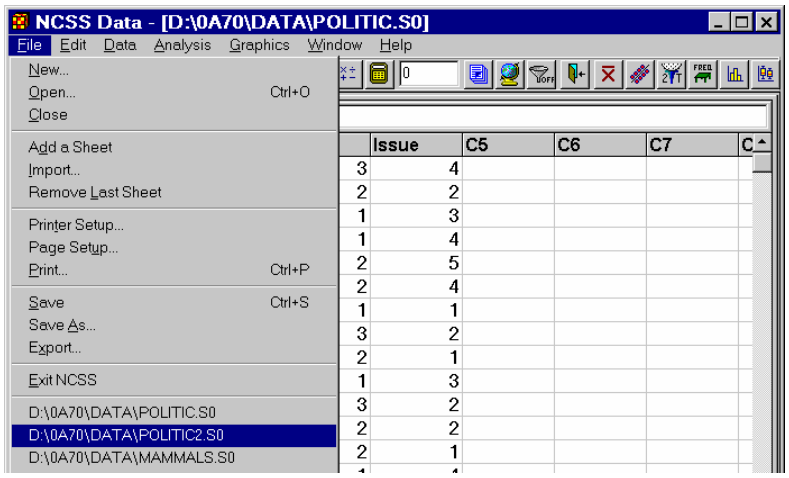


3. Select the information to be copied by dragging the mouse across it.

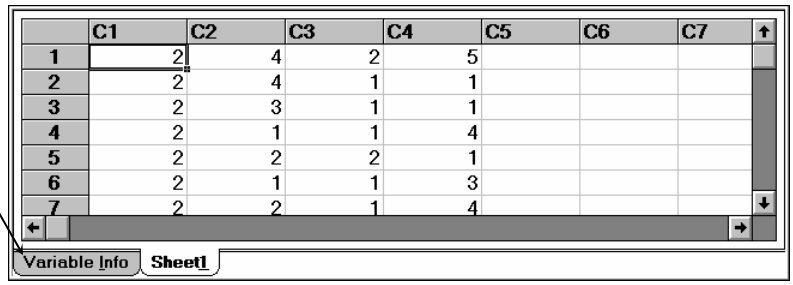
4. Press **Ctrl-C** to copy the information to the clipboard.



5. Open **POLITIC2.S0** by selecting it from the File menu.

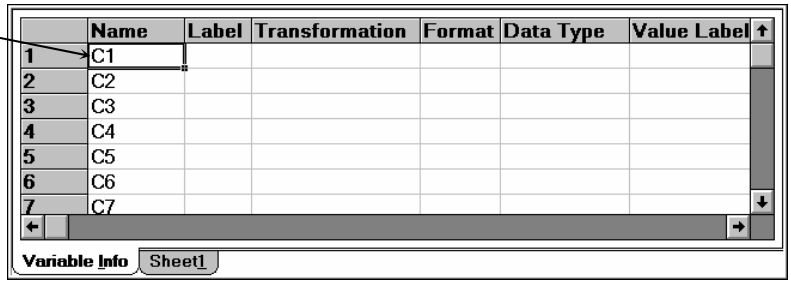


6. Move to the Variable Info datasheet by clicking the **Variable Info** tab.

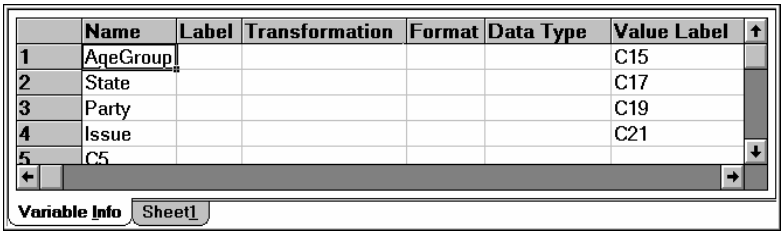


7. Position the cell cursor over the cell containing **C1**.

8. Press **Ctrl-V** to paste the label information into the subset database.



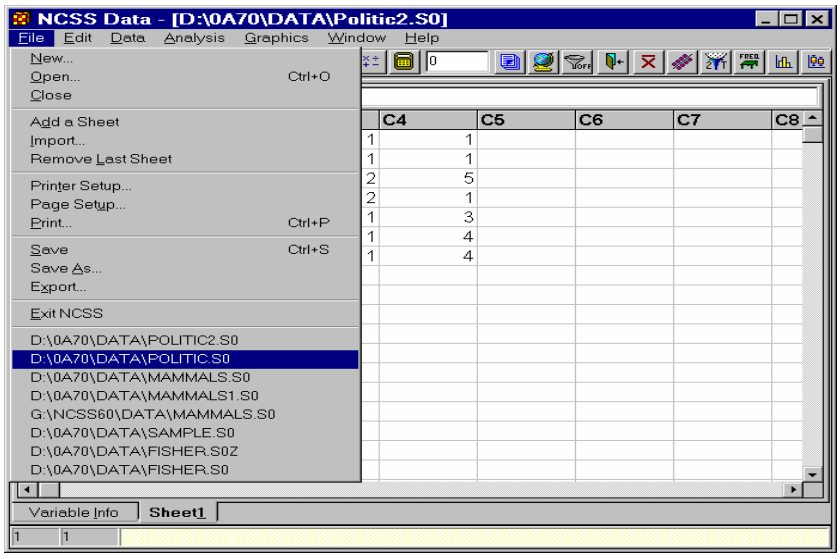
The result will appear as shown.



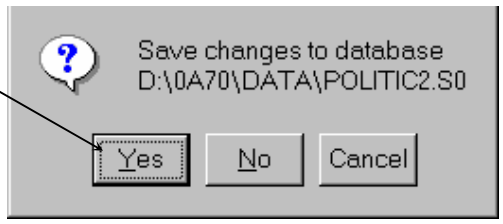
Step 4 – Copy Value Labels to the New Database

The final step is to copy the value labels from the old database to the subset database.

1. Open the **POLITIC.S0** database by selecting it from the File menu.



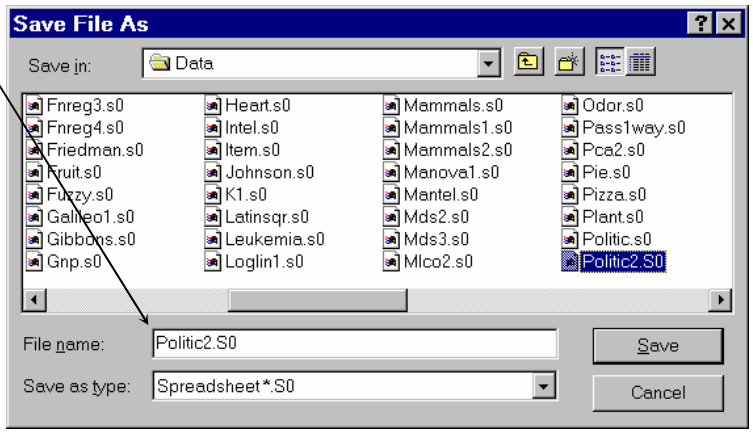
2. Select **Yes** to save the changes that you have just made to the POLITIC2 database.



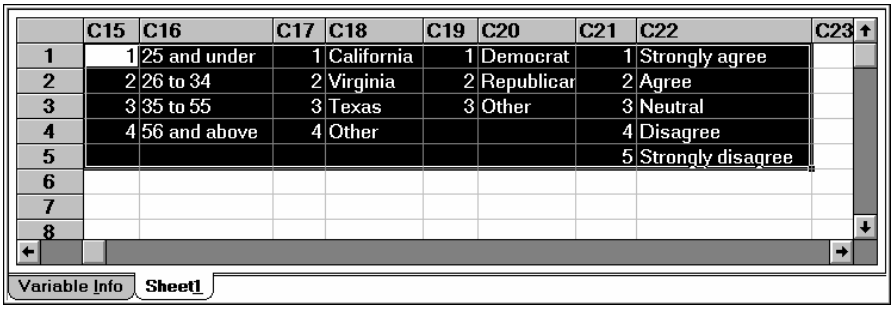
3. Enter **politic2.s0** in the File Name field of the Save File As dialog box.

Click the **Save** button.

When asked, indicate that you want to replace the previous version of this file.



4. Reposition the datasheet so that the first column is **C15**. Drag the mouse across and down so that all of the value labels are selected as shown below.

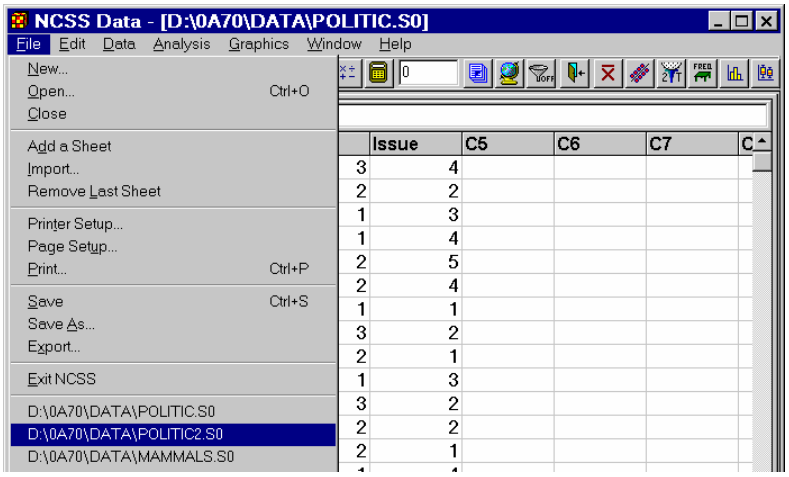


	C15	C16	C17	C18	C19	C20	C21	C22	C23
1	1	25 and under	1	California	1	Democrat	1	Strongly agree	
2	2	26 to 34	2	Virginia	2	Republicar	2	Agree	
3	3	35 to 55	3	Texas	3	Other	3	Neutral	
4	4	56 and above	4	Other			4	Disagree	
5							5	Strongly disagree	
6									
7									
8									

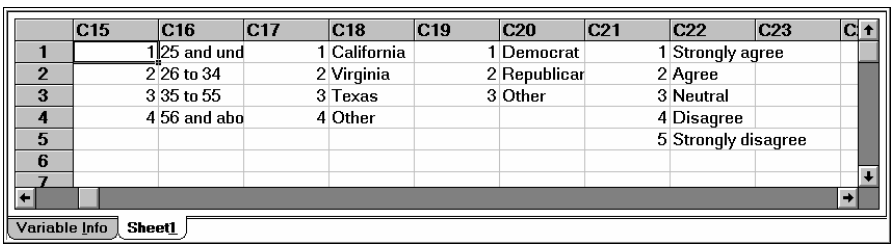
Variable Info | Sheet1

5. Press **Ctrl-C** to copy the selected information to the clipboard.

6. Open the **POLITIC2.S0** database by selecting it from the File menu.



7. Reposition the datasheet so that variable **C15** is at the left.



	C15	C16	C17	C18	C19	C20	C21	C22	C23
1	1	25 and und	1	California	1	Democrat	1	Strongly agree	
2	2	26 to 34	2	Virginia	2	Republicar	2	Agree	
3	3	35 to 55	3	Texas	3	Other	3	Neutral	
4	4	56 and abo	4	Other			4	Disagree	
5							5	Strongly disagree	
6									
7									

Variable Info | Sheet1

8. Click in the first row of variable **C15** so that this is the active cell.
9. Press **Ctrl-V** to copy the information. The final result should appear as below.
10. Select **Save** from the File menu to save the database.

Review

The following is a review of the steps for creating a database subset:

1. Sort the database by the variables on which you want to subset.
2. Copy the subset data to a new database.
3. Copy the variable info from the old database to the subset database.
4. Copy value label information (if it exists) from the old database to the subset database.

Chapter 15

Data Simulation

About This Chapter

There are many situations in which you want to generate data that follow a known distribution. For example, you may want to generate 100 uniform random numbers as an aid in selecting a random sample or you may want to generate five columns of normal random numbers to experiment with a particular statistical test. This chapter will show you how to use transformations to generate simulated data.

For transformations, **NCSS** directly generates two types of random numbers: uniform and normal. Other types of random numbers may be generated by using their inverse probability function on a set of uniform random numbers.

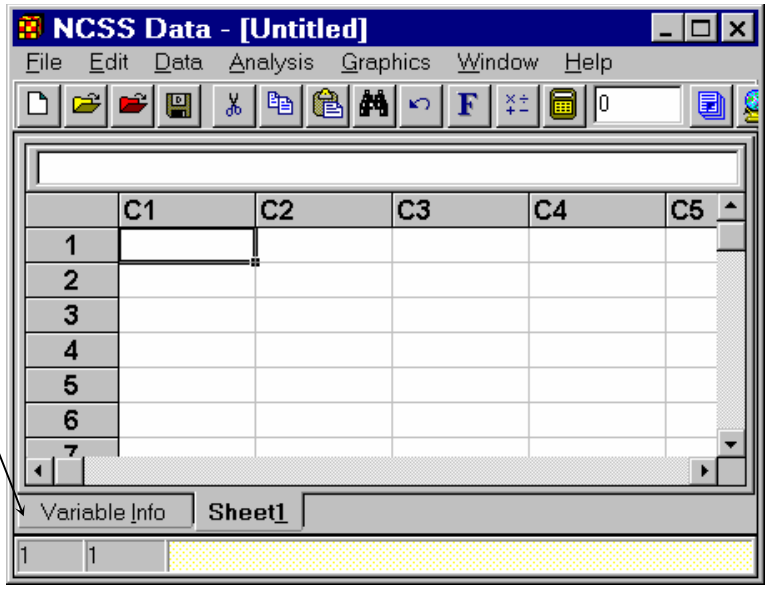
NCSS also has the Data Simulator procedure (Chapter 122), in which many more options for simulating data of various distributions, including mixture distributions, is discussed.

Generating Uniform Random Numbers

In this tutorial you will generate 100 uniform random numbers.

You should begin this tutorial with an empty database. If your database is not empty, select New from the File menu to clear it.

1. Move to the **Variable Info** datasheet by clicking the **Variable Info** tab.

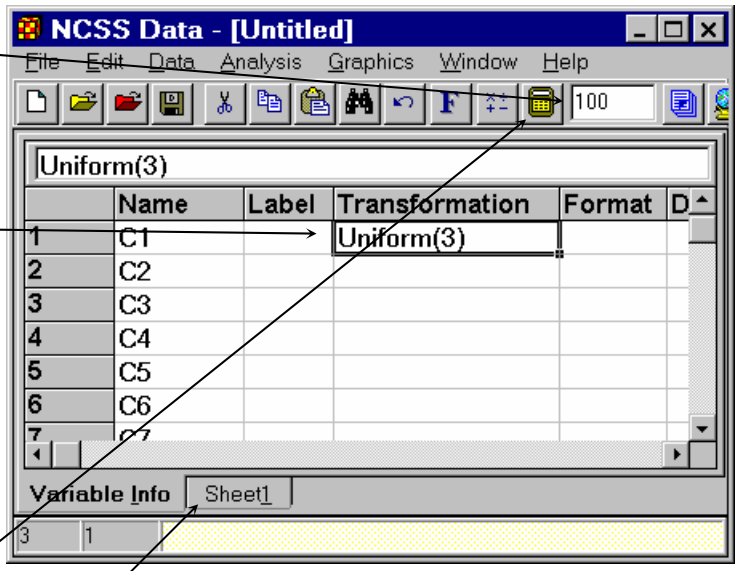


2. Enter **100** in the Number of Rows box.

This specifies the number of rows to be generated.

3. Enter **Uniform(3)** as the transformation for variable C1.

Note that the "3" in the parenthesis is ignored. The program generates a random "seed" so that a different set of random numbers will be used each time you recalculate the spreadsheet.

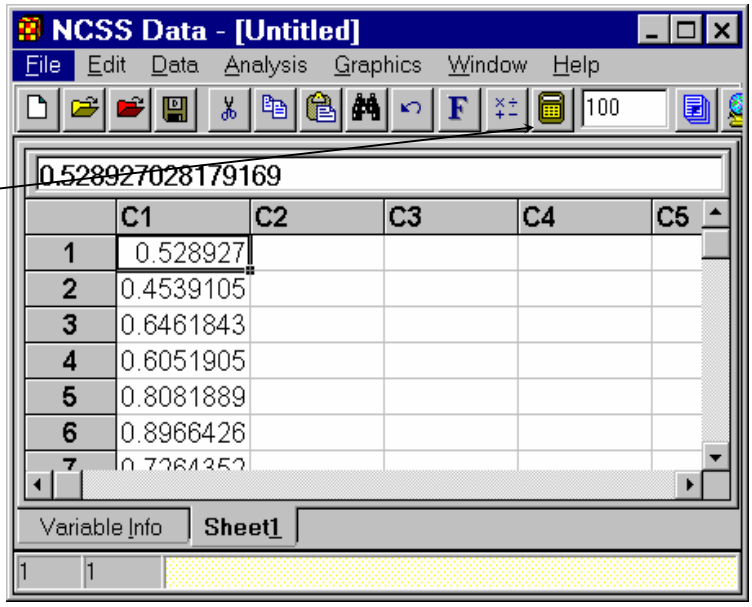


4. Press the **Apply Transformation** button to generate the random numbers.

5. Click the **Sheet1** tab to view the data.

The data will appear as shown. However, the numbers themselves will be different since each recalculation uses a different starting seed.

6. Press the **Apply Transformation** button a few more times to generate new sets of random numbers.



Simulating the T-Test with N = 5

We will now run a simulation in which we generate 100 one sample t-test values with a sample size of five. Four of the values will come from a normal distribution with mean 50 and standard deviation 2. The fifth value will come from a normal distribution with mean 50 and standard deviation 15. The t-test will test the null hypothesis that the population mean of the sample is 50.

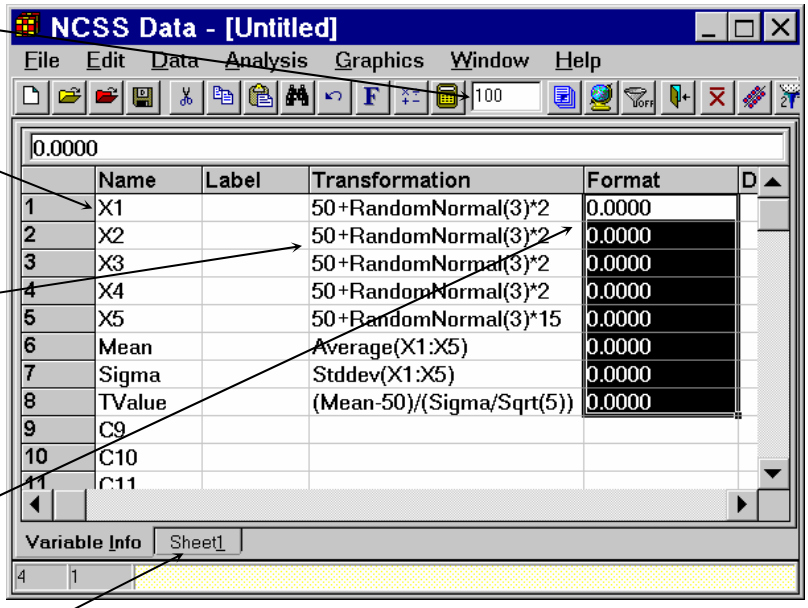
It will be interesting to study the distribution of these T-values since the T-Test makes the assumption that all five data values follow identical distributions. This simulation will allow us to study the distortion that occurs when this assumption is not met.

1. Enter **100** for the number of rows.

2. Enter the new variable names.

3. Enter the transformations. Notice that we multiply the random normal by the standard deviation (2 or 15) and then add the mean (50).

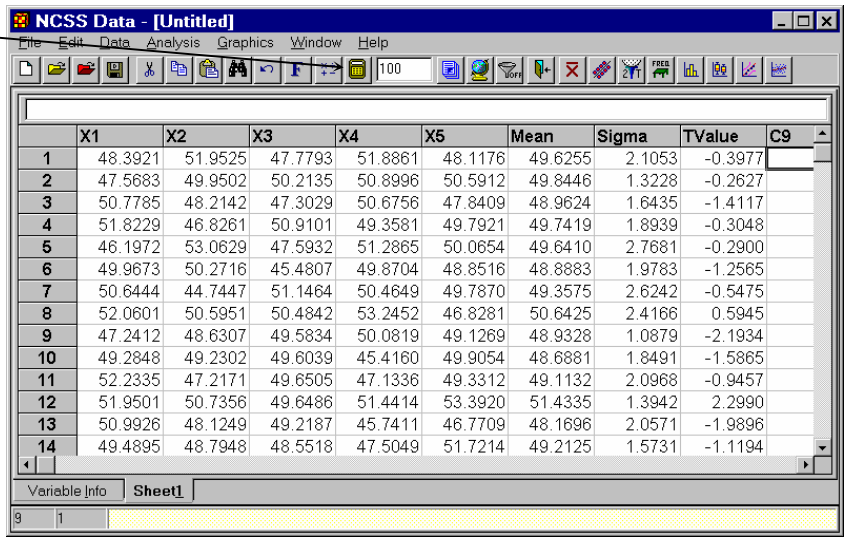
4. Enter **0.0000** as the format for each of the variables. This will make the data much easier to read.



5. Move to the empty spreadsheet by clicking **Sheet1.**

6. Click the **Apply Transformation** button to generate the simulated data. Your results will be similar to those shown here.

If you have made an error in entering the transformation formulas, you will have to go back to the Variable Info datasheet to make corrections.



	X1	X2	X3	X4	X5	Mean	Sigma	TValue	C9
1	48.3921	51.9525	47.7793	51.8861	48.1176	49.6255	2.1053	-0.3977	
2	47.5683	49.9502	50.2135	50.8996	50.5912	49.8446	1.3228	-0.2627	
3	50.7785	48.2142	47.3029	50.6756	47.8409	48.9624	1.6435	-1.4117	
4	51.8229	46.8261	50.9101	49.3581	49.7921	49.7419	1.8939	-0.3048	
5	46.1972	53.0629	47.5932	51.2865	50.0654	49.6410	2.7681	-0.2900	
6	49.9673	50.2716	45.4807	49.8704	48.8516	48.8883	1.9783	-1.2565	
7	50.6444	44.7447	51.1464	50.4649	49.7870	49.3575	2.6242	-0.5475	
8	52.0601	50.5951	50.4842	53.2452	46.8281	50.6425	2.4166	0.5945	
9	47.2412	48.6307	49.5834	50.0819	49.1269	48.9328	1.0879	-2.1934	
10	49.2848	49.2302	49.6039	45.4160	49.9054	48.6881	1.8491	-1.5865	
11	52.2335	47.2171	49.6505	47.1336	49.3312	49.1132	2.0968	-0.9457	
12	51.9501	50.7356	49.6486	51.4414	53.3920	51.4335	1.3942	2.2990	
13	50.9926	48.1249	49.2187	45.7411	46.7709	48.1696	2.0571	-1.9896	
14	49.4895	48.7948	48.5518	47.5049	51.7214	49.2125	1.5731	-1.1194	

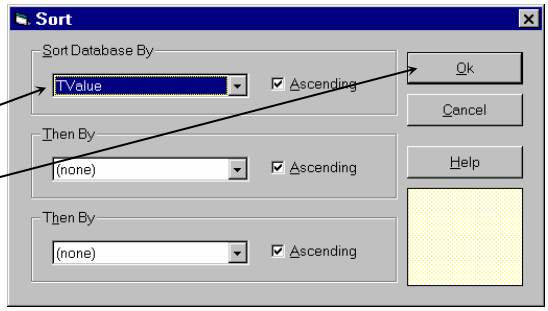
There are many ways to analyze the results. One of the easiest is to sort the TValue column and count the number of rows whose values are outside the theoretical bounds. If these data had come from a normal distribution with a mean of 50 and a standard deviation of 2, you could use the Probability Calculator to determine the theoretical cut off values. The two-tail critical value for a T distribution with 4 degrees of freedom and alpha = 0.05 is 2.78. Hence, you would expect that five of the one hundred values would be less than -2.78 or greater than 2.78.

Here's how to sort the data:

1. Select **Sort** from the Data menu. This will bring up the Sort window.

2. Select **TValue** as the sort variable.

3. Click **Ok** to perform the sort.



4. Scroll from through your data counting how many values are less than -2.78 or greater than 2.78.

	X5	Mean	Sigma	Tvalue	C9
1	49.3032	48.4274	1.3275	-2.6489	
2	29.2799	43.8124	8.2079	-1.6857	
3	40.6481	47.0399	3.9910	-1.6585	
4	43.0464	48.1429	3.0174	-1.3762	
5	40.2060	47.3166	4.4716	-1.3419	
6	30.3861	45.2287	8.3311	-1.2806	
7	24.2173	43.7391	11.3038	-1.2385	

	X5	Mean	Sigma	Tvalue	C9
94	54.8138	51.9395	2.6372	1.6445	
95	57.6396	52.3783	3.1958	1.6640	
96	57.1890	52.2512	3.0049	1.6752	
97	56.9590	52.1490	2.8435	1.6899	
98	56.6380	52.0277	2.6669	1.7002	
99	61.8765	53.8225	4.6030	1.8569	
100	51.1518	51.8507	1.1012	3.7580	
101					

In our case, only one row is outside the range. We repeated this simulation several times and never found more than three values outside the range, much less than the five values that the null hypothesis predicted.

Chapter 16

Cross Tabs on Summarized Data

About This Chapter

This chapter presents an example of how to enter and analyze a contingency table that has already been summarized.

Sample Data

The following data are the results of a study that tested the impact of three drugs on a certain disease.

	Drug		
Disease	Type 1	Type 2	Type 3
Yes	15	28	44
No	4	7	9

These data are entered into an **NCSS** database as follows.

	Drug	Disease	Count	C4
1	1	1	15	
2	1	0	4	
3	2	1	28	
4	2	0	7	
5	3	1	44	
6	3	0	9	
7				

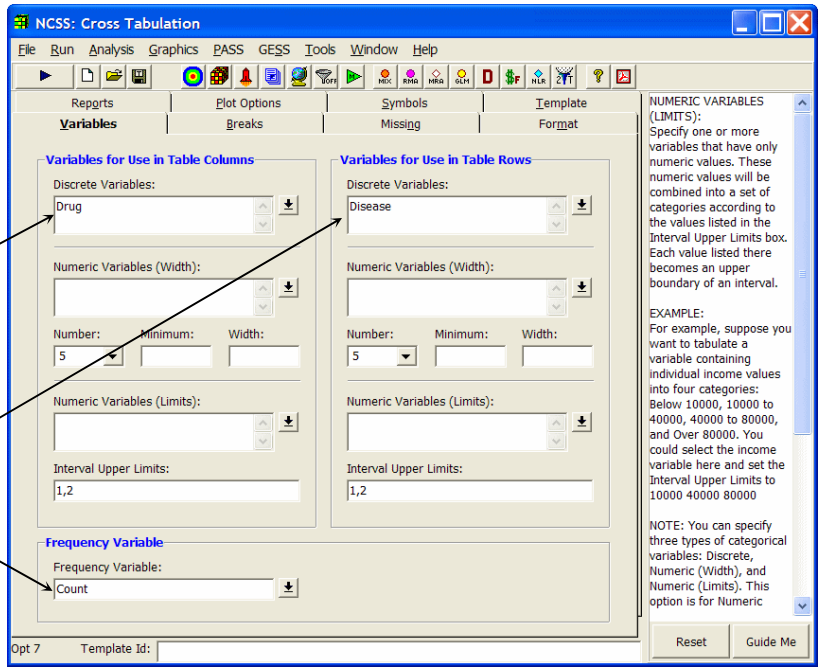
Notice that we have created three variables:

1. One containing the column identification number (**Drug**).
2. One containing the row identification number (**Disease**).
3. One containing the counts (**Count**).

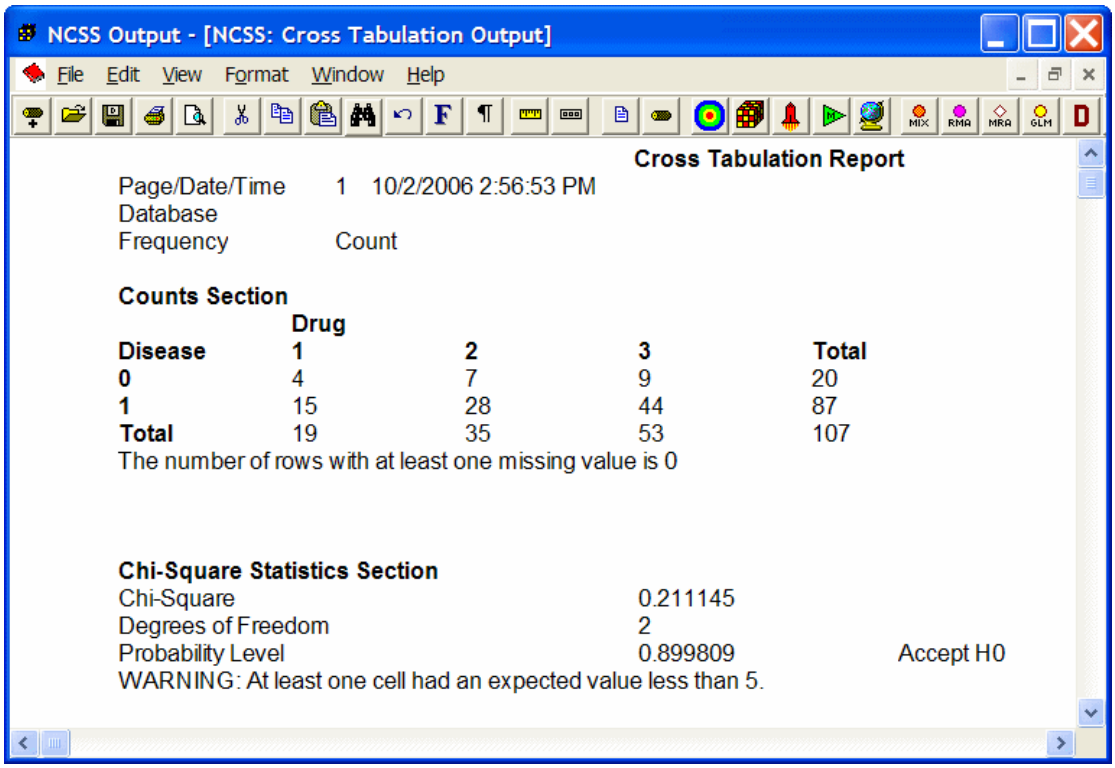
Filling Out the Cross Tabulation Window

The next step would be to fill out the Cross Tabulation window. This is done as follows.

1. Choose **Cross Tabulation** from the Descriptive Statistics submenu of the Analysis menu. This will load the Cross Tabulation window.
2. Enter **Drug** in the Discrete Variables box under Table Columns heading.
3. Enter **Disease** in the Discrete Variables box under Table Rows heading.
4. Enter **Count** in the Frequency Variable box.
5. Press the **Run** button to run the analysis.



The final result will appear as follows.



Quick Start Index

A

ASCII dataset, 12-1

B

Basics, 1-1
BRAIN WEIGHT dataset, 2-2

C

Chi-square test example, 16-1
Copying data, 7-2
Creating a database, 2-1
Cross tabulation
  summarized data, 16-1

D

Data
  entering, 2-1
  importing, 12-1
  printing, 2-7
  saving, 2-6
  simulation, 15-1
Data transformation, 3-1
Data window, 1-4, 7-1
Database
  clearing, 2-9
  creating, 2-1
  loading, 2-1, 2-10, 7-1
  printing, 2-7
  s0 and s1 files, 2-6
  S0-type, 2-9
  S0Z-type, 2-9
  subsets, 14-1
Dataset
  ASCII, 12-1
  BRAIN WEIGHT, 2-2
  MAMMALS, 3-1, 4-1, 10-1
  MAMMALS1, 5-1, 6-1
  POLITIC, 13-1, 14-1
Descriptive statistics, 4-1
Documentation, 1-11

F

Filter
  disabling, 10-4
Filters, 10-1

H

Help system, 1-10

I

Importing data, 12-1
Installation, 1-1
  folders, 1-1

L

Labeling variables, 2-4
Labels
  values, 13-1
Loading a database, 2-1, 2-10, 7-1
Log of output, 9-6

M

MAMMALS dataset, 3-1, 4-1, 10-1
MAMMALS1 dataset, 5-1, 6-1

O

Output
  log of, 9-6
  printing, 9-4
  saving, 9-5
Output window, 1-6, 9-1

P

Pasting data, 7-2

POLITIC dataset, 13-1, 14-1
Printing
  data, 2-7
  output, 9-4
  output reports, 4-5
Procedure window, 1-5, 8-1

R

Random numbers
  uniform, 15-1
Recode transformation, 3-4
Recoding, 11-1
Regression analysis, 6-1
Running a regression analysis, 6-1
Running a two-sample t-test, 5-1
Running descriptive statistics, 4-1

S

Saving
  data, 2-6
  output, 9-5
  template, 8-5
Selecting procedures, 1-7
Serial numbers, 1-3
Simulation
  data, 15-1
Starting NCSS, 1-2, 2-1
Subset of a database, 14-1
System requirements, 1-1

T

Template
  saving, 8-5
Transformation
  recoding, 3-4
Transformations, 3-1
  recoding, 11-1
  simulation, 15-1
Two-sample t-test, 5-1

V

Value labels, 13-1
Variable info tab, 2-4
Variable labeling, 2-4
Variable name, 2-4
Variable names
  rules for, 2-5

W

Window
  data, 7-1
  output, 9-1
Windows
  navigating, 1-4
Word processor, 9-1

Chapter 100

Installation

Installing NCSS

Insert the CD into the CD drive and wait for a small window to appear. Click the button that indicates that you want to install ***NCSS*** and follow the instructions on the screen. For help with installation, email support@ncss.com or call 801-546-0445.

Starting NCSS

After you have run ***NCSS*** Setup, you will see it listed on the Windows Start menu. You can start ***NCSS*** by selecting it from the Start menu.

Serial Numbers

When you start the software the first time, a window will ask for the serial number of each product. Enter the serial number(s) provided with the CD. Any programs without valid serial numbers will run in trial mode for 7 days.

System Tutorial

The first time you run ***NCSS***, we strongly suggest you go through the five-minute tutorial given in the next chapter. Once you have done this, you can enter your own data and proceed to the statistical procedure(s). Each procedure has its own brief tutorial. We recommend that you read through the tutorial before running each procedure.

Quick Start Files

The Quick Start files may be used to gain more familiarity with ***NCSS*** in an introductory setting. To access Quick Start files, go to 'Help' and click on 'NCSS Quick Start'.

Help System

To help you learn and use ***NCSS*** efficiently, the material in this manual is included in the Help system. The Help system is started from the Help Menu. ***NCSS*** updates, available for download at http://www.ncss.com, may contain adjustments or improvements of the ***NCSS*** Help system.

NCSS Support Services

If you have a question about ***NCSS***, you should first look to the printed documentation and the included Help system. If you cannot find the answer there, look for help on the web at www.ncss.com/support.html. If you are unable to find the answer to your question by these means, contact ***NCSS*** technical support for assistance by calling (801) 546-0445 between 8 a.m. and 5 p.m. (MST). You can contact us by email at support@ncss.com or by fax at (801) 546-3907. Our technical support staff will help you with your question.

If you encounter problems or errors while using ***NCSS***, please view our list of recent corrections before calling by going to www.ncss.com/release_notes.html to find out if you problem or error has been corrected by an update. You can download updates anytime by going to http://www.ncss.com/download.html. If updating your software does not correct the problem, contact us by phone or email.

To help us answer your questions more accurately, we may need to know about your computer system. Please have pertinent information about your computer and operating system available.

Chapter 101

Tutorial

Introduction

This chapter will introduce you to the **NCSS** system and acquaint you with our spreadsheet, procedures, and word processor. It will guide you through several basic tasks, such as creating a database, entering data, running a procedure, and printing your output.

We will not dwell on any one topic since more complete explanations are presented later. Instead, we will try to give you the big picture.

This tutorial will rely on a brief example. You should run the example on your computer, making each entry as instructed. In this way, we hope to show you the main concepts you will need to use **NCSS** quickly and effectively.

The basic steps to using **NCSS** are:

1. Open a database in the spreadsheet. This step may include entering or importing your data.
2. Select one or more analyses for the data. The various analyses are run using point-and-click procedures. No programming is required.
3. View your output in the word processor. From here you can edit, print, and save your output.

Introduction to Databases

An **NCSS** database consists of a collection of two or more spreadsheet files that we call *datasheets*. Each datasheet is displayed in table format with rows and columns. Statistics, such as averages and standard deviations, are calculated down a column (variable) of data and not across a row (observation).The first datasheet, called the *Variable Info* datasheet (with file extension S0), contains variable information such as names, labels, formats, and transformations.

Additional datasheets contain the actual data arranged as a rectangular table in which the columns are variables and the rows are observations. These have consecutive file extensions S1, S2, etc.

Variable Names

Variable names are used throughout the program to identify which columns of data to analyze. A variable name should begin with a letter (not a number); should contain only letters, numbers, and the underscore; and should not contain blanks. For correct formatting on reports, variable names should be less than twelve characters, although there is no maximum length.

Numeric Accuracy

NCSS maintains double-precision (sixteen digit) accuracy in all numeric calculations. When you apply a special format to a variable, this changes the display of the data only. The actual sixteen-digit number is maintained on the datasheet.

Starting NCSS

To start **NCSS**, you either click on the **NCSS** icon or select NCSS from the Windows Start menu. You must first install the program on your system. When **NCSS** begins, it will load and display an empty database.

Creating a New Database

When **NCSS** starts, it opens an empty database. You can return to a new database at any time by selecting 'New' from the File Menu.

Entering and Editing Data

Entering data is easy. Position the cursor in the desired cell and begin typing. Each time you complete an entry, hit the Enter key to move down to the next cell. Once you have put in all the data for one column, you move to the next. You can also import data from other files, but that is discussed elsewhere.

Try entering a few numbers until you feel comfortable with the data entry process. You will notice that **NCSS** behaves like most other spreadsheet programs.

Edit a data value by positioning the cell cursor on the item you wish to edit and typing the new value. When you have completed the replacement value, press the Enter key. The new value will replace the old.

Naming Variables

When a new database is created, default variable names are assigned. These are C1, C2, C3, and so on. You will usually want to change these names to something more easily remembered. To do this, you must open up the Variable Info datasheet. This is done by clicking on the *Variable Info* tab at the bottom of the datasheet.

You can change the names, labels, formats, and transformation formulas of each variable by positioning the cursor at the cell you want to change and typing the new entry.

Saving Your Data

As you enter data into a datasheet, it is stored in your computer's temporary memory but not on your hard disk. If you want to make a permanent copy of your database, you must select Save As from the File Menu. A dialog box will appear that will let you name and save your data as a database on your hard disk.

Remember that an **NCSS** database actually consists of two or more spreadsheet files. You will designate a name for the first datasheet, the one with the file extension S0. Additional datasheets will automatically be saved. For example, if you want to create the ABC database, you will name the file ABC.S0.

Opening an Existing Database

Take the following steps to open an existing **NCSS** database: the SAMPLE database. This database contains the data for many of the tutorials in this manual. To open this database, select Open from the File Menu. In the Data subdirectory, you will find a file named SAMPLE.S0. Select this file. The Sample database will be opened for your use.

Running a Procedure

We will now take you through the process of running an analysis. For our example, we will run a two-sample T-Test on the variables YldA and YldB contained on the SAMPLE database.

Step 1 – Open the SAMPLE Database

If you do not have the SAMPLE database open, open it now by selecting Open from the File menu. In the Open dialog box, move to the DATA directory and select SAMPLE.S0.

Step 2 – Open the T-Test Procedure Window

Once the desired database is open, you are ready to select your analysis. From the Analysis menu, select T-Tests. From the submenu, select T-Test - Two-Sample. The T-Test - Two-Sample procedure window will be displayed.

This window lets you set the options for the T-Test procedure. You will notice that this window is partitioned into various panels. Initially, you are positioned in the Variables panel. You can move from panel to panel by selecting the tabs at the top of the window.

As you move the mouse cursor over the various boxes on the panel, you will see that a help message appears in the large space at the right of the window. We call this the Immediate help.

Step 3 – Select the Variables to be Analyzed

Double click the Response Variable(s) option to display the Variable Select box. This dialog box lets you select the variables to be analyzed. Select YldA and YldB. Now click on the Ok button. The variable names YldA-YldB will be entered into the Response Variable(s) option. All other options can remain at their default values.

Step 4 – Run the Analysis

To run the analysis, press the arrow button (Run Procedure) on the left side of the toolbar.

The program will calculate for a few seconds, and then the word processor will be opened with the output displayed.

Step 5 – View the Output

Double-click on the title bar of the Output document to expand it to maximized position. You can now view, edit, save, or print your output. Our word processor will allow you to perform common editing tasks such as cut, copy, and paste. It will let you print your document or save it for later use. Note that when saved, the document may be opened in any of the commercial word processors that read RTF documents - including Microsoft Word.

As you run additional analyses, this output will be replaced by the new output. You can save all or part of your output to a log file by selecting the appropriate item from the File Menu. This log file serves as a permanent record of your work.

Quitting NCSS

Once you are through analyzing your data, you quit the **NCSS** program by selecting Exit NCSS from the File Menu of the spreadsheet, a procedure, or the word processor. **NCSS** will close all of its data files and remove itself from temporary memory.

Chapter 102

Databases

Introduction

NCSS analyzes data contained in a database. There are two types of **NCSS** databases: S0 and S0Z. We will now explain these two types of databases.

S0 (Spreadsheet) Database

An S0 database (sometimes referred to as a worksheet) is made up of a Variable-Info sheet and one or more datasheets. Variable names, labels, formats, and transformations are contained on the Variable-Info sheet. Each datasheet contains 256 variables (columns) and room for up to 16,384 observations, although we recommend that you use this type for databases with less than 1000 observations. You can add as many datasheets to a database as you like (thereby increasing the number of variables), but you cannot increase the number of observations.

When we refer to a variable on an **NCSS** database, we are actually referring to a specific column of a datasheet. All procedures analyze one or more variables from these datasheets.

NCSS accepts both numeric and text data. Numeric data may contain up to 16 digits (double-precision). Text data may contain up to 1000 characters per cell.

Physically, an S0-type database is made up of two or more files with appropriate extensions. The Variable-Info file has the extension S0. The datasheets have the extensions S1, S2, etc. Hence, a database called “ABC” with 512 variables (two datasheets) would appear on your hard drive as three files: ABC.S0, ABC.S1, and ABC.S2. This is important to remember when backing up or copying an S0-type database.

Each of these files is actually a *Microsoft Excel 4.0* compatible spreadsheet file. This is where the row and column limits come from since an Excel 4.0 spreadsheet can contain up to 256 columns and 16,384 rows. We have used this format because it is popular, transportable, and because it allows us to provide a familiar, spreadsheet-style interface complete with formatting and formulas.

S0Z (Zipped) Database

The S0Z-type database is disk-based: only small portions of the database reside in your computer’s memory at any one time. Because this type of database is not memory resident, it can be much larger: up to about 5,000 variables and over 100,000 rows. Because the complete database is not stored in your computer’s RAM memory, it is slower to access and the “Undo” feature is not available. On the positive side, because all data is not loaded into memory at one time, it does not require as much RAM memory to run medium (100 - 500 rows) to large (10000+ rows) databases.

Unlike some commercial databases, variables (fields) in S0Z-type databases can contain a mixture of numeric and text values. You do not have to specify the field type in advance. Numeric data may contain up to 16 digits (double-precision). Text data may contain as many characters as you have made space for on a data record. You define the record size when the database is initialized.

Technical Details of S0Z Databases

An understanding of how an S0Z database is constructed will help you make intelligent choices when you create one. The S0Z database may be viewed as one long row of data which is made up of chunks called “records.” Each record contains a fixed number of characters (bytes), which cannot be changed (without copying the existing data to a new database with a larger record length).

Record 1

The first record contains database-size information such as the number of variables, number of rows, and record length. Its length is 512 bytes.

Records 2 through M+1

The next M records (where M represents the number of variables) contain information about the variables. Each record contains information about the corresponding variable, such as its name, label, transformation, format, and data type.

Records M+2 to M+N+1

The next N records (where N represents the number of observations or rows) contain the data values. Each record holds one row of data. The record is made up of M blocks of k bytes each followed by T blocks of m bytes each.

The first M blocks hold the regular data values, one per block. Numeric values (including date) use the first 8 bytes of a block. Text values use the last k-4 bytes of a block. Hence, the default, 10-byte, block can hold a text value of up to 6 bytes.

The last T blocks provide additional storage for extra long text values. When a text value will not fit in the regular data block, it is stored in the first available text block at the end of the record. If all of these overflow blocks are full, the text is truncated to k-4 bytes.

When you create a database, you must pay particular attention to how much space you want to provide for text data. You have two alternatives. First, you can increase the size of k so that most of your text values will be stored in the first M blocks. This method increases the overall size of the database, but provides the fastest access. Second, you can increase T so that there are more overflow blocks. This method provides slightly slower access to your data, but makes for a smaller database.

Note that if you are planning to use value labels, you need to include enough room to hold them.

Calculating Record Size

The record size is calculated as

$$L = Mk + Tm$$

where

L	Record size in bytes
M	Number of variables
k	Length of a variable field (default is 10 bytes)
T	Number of extra text fields
m	Length of an extra text field (default is 25 bytes)

For example, suppose you want to create a database that will have space for 50 numeric variables and 10 text variables. Suppose the largest text field is 30 characters long. The record size would be

$$50(10)+10(30) = 800 \text{ bytes.}$$

As explained next, you should be liberal in your assignment of the maximum number of variables. Make sure you include enough space for new and transformed variables. Also, be sure you include enough text variables to hold all value labels that you might use.

Zipped Database

To provide flexibility in what may be stored, we suggest that you always add a little extra to the number of bytes you think will be required. Because this will obviously “waste” some disk space, we have added one final operation: your database is permanently stored in a compressed (zipped) format. Because of this, you do not have to be so careful about how much extra space you add to each record. When the file is compressed, it is often reduced to one-tenth the size. For example, the size of our SAMPLE database (50 columns by 75 rows) is about 100,000 bytes. When compressed, its size is only 12,000 bytes!

When you load a S0Z database, the actual file is not “loaded.” Instead, an uncompressed copy of the database (with the extension “S0N”) is created in your temporary NCSS directory. As you read and write to and from the database, you are actually using the temporary database, not the original.

When you save the database, the program compresses the temporary database and replaces the existing S0Z file with the new compressed version. This means that you can always “undo” operations that have resulted in your data being unintentionally modified. You simply re-open your database without saving the modifications. When the program asks if you want to save changes, just say no! The modified temporary file will be replaced by another expanded copy of your compressed database.

Comparison of S0 and S0Z Databases

The following table presents a brief comparison of these two types of databases. It will help you determine which type to use in a particular situation.

Criterion	S0 Database	S0Z Database
Rows	Recommend < 1000	Unlimited
Columns	Recommend < 50	Up to 16,000
Transformations	Yes	Yes
Cut/Copy/Paste	Yes	Yes
Insert/Delete	Yes	Yes
Undo	Yes	No
Speed	Fast	Medium
File Size	Regular	Small (Compressed)
Text Data	Yes	Yes
Numeric Data	Yes	Yes
Flexibility	Yes	Must stay within limits set when created.

Accuracy

NCSS maintains double-precision (sixteen digit) accuracy in the data values. When you apply a special format to a variable, only the display of the data is modified. The actual sixteen-digit number is maintained on the datasheet.

All numbers are stored in the IEEE floating-point format. Such numbers range from 4.94E-324 to 1.798E308 for positive values, 0, and from -1.798E308 to -4.940E-324 for negative values.

Because of the rounding that has to occur to fit into the IEEE format, it is impossible to express all numbers exactly. For example, the common fraction of one-third cannot be expressed exactly, but must be rounded at the last digit. This rounding may cause strange results for certain decimal numbers. For example, you might enter 453.4537 and have it displayed as 453.45369999999999. There is nothing we can do to change this oddity. You can rescale your data by multiplying the variable by 10 or 100. However, this problem occurs rarely in practice and will not change the accuracy of your results.

The “E” notation is used for expressing large and small numbers in scientific notation. The number after the E is the exponent (base 10) that is applied to the base number. Thus, 3.238E-03 means 0.003238. Other examples are:

E-Notation	Decimal Equivalent
-3.238E-06	-.000003238
-3.238E-02	-.03238
3.238E04	32380.0
3.238E06	3238000.0

Missing Values

Missing values are represented by empty cells. In procedures that use only numeric values, such as the average, text values are also treated as missing values. Unless otherwise specified, **NCSS** uses *row-wise deletion* of observations with missing values. This means that if any of the active variables in a row has a missing value, the whole observation is omitted from the calculations.

Variable Info

Each column of a datasheet is called a *variable*. Each S0 datasheet contains space for 256 variables. Hence, if you have a survey with 700 questions, you will need three datasheets in your database to hold the data, or you could use one S0Z database.

You should note that even though a datasheet has space for 256 variables, only cells that actually contain data are stored when you save the database. You will not be wasting a lot of disk space when you use only 5 or 10 of the 256 variables that are available on the datasheet.

All aspects of a variable (except for the actual data) are modified on the Variable Info sheet. You can view the Variable Info sheet by clicking the Variable Info tab at the bottom of the spreadsheet.

The Variable Info may be printed out using the Print option of the File Menu.

Note that you can modify the information on the Variable Info datasheet without actually viewing it. This may be accomplished by selecting Variable Info from the Edit menu. The Edit menu may be activated using the standard menu at the top of the screen or by clicking the right-mouse button while the mouse pointer is over a specific variable on the datasheet.

We will now explain each column of the Variable Info sheet. Note that each row of this sheet refers to a specific variable on the database.

Number

The first column of the Variable Info sheet gives the variable numbers. A variable's number is determined by its position on the database. Although you refer to a variable by its name when you write transformations and make variable selections, your selection may be stored in the Procedure Template by number. This is all handled internally, so you will not have to worry about it except when you move variables around within a database. When you do this, you will have to check stored procedure templates to make sure they still refer to the correct variables.

Name

The second column of the Variable Info sheet gives the variable names. Each variable is given a default name when the database is created. Throughout the program, you refer to the variable by its name. The default names are C1, C2, C3, etc.

The variable's name may be changed at any time by editing the Name column of the Variable Info sheet. The syntax for the naming of variables follows these rules:

1. Variable names must begin with a letter.
2. Variable names can contain only letters, numbers, and the underscore.

3. Variable names may not include spaces.
4. Variable names may not include mathematical symbols.
5. Upper- and lower case letters are the same. Names are case insensitive.

Label

The third column of the Variable Info sheet gives the variable labels. Each variable may have a label associated with it. This label is of arbitrary length. No attempt is made to trim this label if it is too long for the space provided to display it in a particular report. Each Procedure Template has an option in which you designate whether to include these labels on the output.

Transformation

The fourth column of the Variable Info sheet holds a variable's transformation, if one has been assigned. Variable transformations are discussed in detail in another section.

Format

The fifth column of the Variable Info sheet gives the variable format. When this value is left blank, the *General* format is assumed. You can enter a separate format statement for each variable. The format controls both the display and color of the data.

A special Format window may be used to modify a variable's format. This window may be viewed by double clicking in the Format column or by selecting Edit, then Variable Info, then Format from the menus. Note that the Edit menu may also be activated using the right-mouse button.

Several built-in formats are available and it is easy to write your own custom format. Each custom format can have as many as four sections, separated by semicolons:

1. One for positive numbers
2. One for negative numbers
3. One for zeros
4. One for text

Each section is optional, but if you have more than one, you define the format section by the placement of extra semicolons. If you only use one format section, it defines the format for all numbers, both positive and negative. Following is an example of a typical, four-section format:

0.000;(0.000);0; "Numeric Only"

The following table lists the format symbols that can be used in a custom format.

Format Symbol	Description
General	Display a number in general format.
0	This is a digit placeholder. If the number contains fewer digits than the format contains placeholders, the number is padded with 0's. If there are more digits to the right of the decimal than there are placeholders, the decimal portion is rounded to the number of places specified by the placeholders. If there are more digits to the left of the decimal than there are placeholders, the extra digits are retained.
#	Digit placeholder. This placeholder functions the same as the 0 placeholder except that the number is not padded with 0's if the number contains fewer digits than the format contains placeholders.
?	Digit placeholder. This placeholder functions the same as the 0 placeholder except that spaces are used to pad the digits.
. (period)	Decimal point. Determines how many digits (0's or #'s) are displayed on either side of the decimal point. If the format contains only #'s left of the decimal point, numbers less than 1 begin with a decimal point. If the format contains 0's left of the decimal point, numbers less than 1 begin with a 0 left of the decimal point.
%	Display the number as a percentage. The number is multiplied by 100 and the % character is appended.
, (comma)	This is the thousands separator. If the format contains commas separated by #'s or 0's, the number is displayed with commas separating thousands. A comma following a placeholder scales the number by a thousand. For example, the format 0, scales the number by 1000 (e.g., 10,000 would be displayed as 10).
E- E+ e- e+	Displays the number in scientific notation. If the format contains a scientific notation symbol to the left of a 0 or # placeholder, the number is displayed in scientific notation and an E or an e is added. The number of 0 and # placeholders to the right of the decimal determines the number of digits in the exponent. E- and e- place a minus sign by negative exponents. E+ and e+ place a minus sign by negative exponents and a plus sign by positive exponents.
$-+/(): space	Displays that character. To display a character other than those listed, precede the character with a back slash "\" or enclose the character in double quotation marks. You can also use the slash "/" for fraction formats.
\	Display the next character. The backslash is not displayed. You can also display a character or string of characters by surrounding the characters with double quotation marks. The backslash is inserted automatically for the following characters: ! ^ & ' ' `~ { } = < >
* (asterisk)	Repeats the next character until the width of the column is filled. You cannot have more than one asterisk in a format section.

Format Symbol	Description
_ (underline)	Skips the width of the next character. For example, to make negative numbers surrounded by parentheses align with positive numbers, you can include the format _) for positive numbers to skip the width of a parenthesis.
“text”	Displays the text inside the quotation marks.
@	Text placeholder. If there is text in the cell, the text replaces the @.
m	Month number. Displays the month as digits without leading zeros (e.g., 1-12). Can also represent minutes when used with h or hh formats.
mm	Month number. Displays the month as digits with leading zeros (e.g., 01-12). Can also represent minutes when used with the h or hh formats.
mmm	Month abbreviation. Displays the month as an abbreviation (e.g., Jan-Dec).
mmmm	Month name. Displays the month as a full name (e.g., January-December).
d	Day number. Displays the day as digits with no leading zero (e.g., 1-31).
dd	Day number. Displays the day as digits with leading zeros (e.g., 01-31).
ddd	Day number. Displays the day as an abbreviation (e.g., Sun-Sat).
dddd	Day number. Displays the day as a full name (e.g., Sunday-Saturday).
yy	Year Number. Displays the year as a two-digit number (e.g., 00-99).
yyyy	Year Number. Displays the year as a four-digit number (e.g., 1900-2078)
h	Hour number. Displays the hour as a number without leading zeros (e.g., 1-23). If the format contains one of the AM or PM formats, the hours are based on a 12-hour clock. Otherwise, they are based on a 24-hour clock.
hh	Hour number. Displays the hour as a number with leading zeros (e.g., 01-23). If the format contains one of the AM or PM formats, the hours are based on a 12-hour clock. Otherwise, they are based on a 24-hour clock.
m	Minute number. Displays the minute as a number without leading zeros (e.g., 0-59).
mm	Minute number. Displays the minute as a number with leading zeros (e.g., 00-59). The mm format must appear immediately after the h or hh symbol. Otherwise, it is interpreted as a month number.
s	Second number. Displays the second as a number without leading zeros (e.g., 0-59).
ss	Second number. Displays the second as a number with leading zeros (e.g., 00-59).
AM/PM, am/pm, A/P, a/p	12-hour time. Displays time using a 12-hour clock. Displays AM, am, A, or a for times between midnight and noon; displays PM, pm, P, or p for times from noon until midnight.
[*color*]	Displays the output in the specified color where *color* is BLACK, BLUE, CYAN, GREEN, MAGENTA, RED, WHITE, or YELLOW.

Format Symbol	Description
[COLOR n]	Displays the text using the corresponding color in the color palette. n is a color in the color palette.
[conditional value]	Under normal circumstances, the four sections of a custom format are for positive numbers, negative numbers, zeros, and text. Use brackets to indicate a different condition for each section. For example, you might want numbers less than three to be black and those greater than three to be red. You would use: [>3] [RED][General] ; [<-3][RED][General];[BLACK][General]

The following table shows some examples of how these formats are displayed.

Format	Cell Data	Display
#.##	654.429	654.43
#.##	0.429	.43
#.0#	654	654.0
0.00	654.429	654.43
0.00	654.4	654.40
0.00	.429	0.43
#,##0“CR”;#,##0“DR”;0	654.429	654CR
#,##0“CR”;#,##0“DR”;0	-654.429	654DR
#,	10000	10
“Resid=”0.0	123.45	Resid=123.5
000-00-0000	123456789	123-45-6789
m-d-yy	2/3/94	2-3-94
mm dd yy	2/3/94	02 03 94
mmm d, yy	2/3/94	Feb 3, 94
mmm d, yyyy	2/3/94	Feb 3, 1994

Data Type

The sixth column of the Variable Info sheet gives the variable's data type. Currently, there are five data types:

0. Text - case used. (default with Data Type left blank)
1. Text - case ignored
2. Numeric
3. Month (sorts alphabetic month values in January through December)
4. Fixed (not rearranged by sorting or insert/delete; usually used for value label information)

The Data Type of a variable is only used when the variable is employed as a grouping or break variable, as in cross tabulation. In these cases, the data type specifies how the group values are sorted. For example, suppose a variable contains the numeric values: 11, 6, 10, 1, 22, 7. When treated as text, these numbers are sorted in the order: 1, 10, 11, 22, 6, 7. If you designated the Data Type as Numeric, these values would be sorted in the usual numeric order: 1, 6, 7, 10, 11, 22.

The Data Type is ignored when the variable is used in numeric calculations.

Value Label

The seventh column of the Variable Info sheet specifies the value labels. Value labels are labels that are displayed in the place of the original value on the database. For example, you might be tabulating a survey in which a Yes was coded as a 1 and a No was coded as a 0. The final report will be much more interpretable if the Yes and No are displayed rather than the 1 and 0. One approach is to generate a new variable using a Find/Replace operation or the Recode transformation to replace the 1's and 0's with Yes's and No's. A simpler approach is to use Value Labels.

Creating Value Labels

Two methods are available for creating value labels. The original method was to store a list of possible values and their labels directly on the spreadsheet. A new method which was recently added is to store the values and their labels in a text file. We recommend using the second method. We will now explain each approach.

Creating Value Label Files (Recommended)

Values label files are easily constructed using Windows Notepad or similar text editor. The files contain a list of values and their corresponding labels, one set per row. The value and the label must be separated by a tab (not a blank since the labels may contain blanks). Note that this is the format of data that is copied and pasted into a text file from a spreadsheet such as **Excel** or **NCSS**.

Here is an example of a value label file called Likert.txt:

```
1	Strongly Disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly Agree
```

Attaching the Value-Label File to Variables

Once you have created a value-label file, you must associate it with those variables whose values you want to label. This is accomplished by entering the name of value label file (Likert.txt in this example) on the *Variable Info* sheet in the *Value Label* column. Note that several variables can share the same value-label file.

Step-by-Step Instructions for Using Value Labels

1. Click on the *Variable Info* tab at the bottom of the spreadsheet.
2. Double-click in the *Value Label* column on the row corresponding to the variable you want to label.
3. Click the button *Create/Edit a V.L. File* to run the Notepad program.
4. Enter the values and their labels, one per line. The syntax is *Value <tab> Label*. The order of the values does not matter. Note that intermediate blanks in the values and the labels are retained and treated like any other characters.
5. After completing the value label file, save it into the same directory in which the database resides. Note that NCSS first searches for a value label file in the directory in which the database resides and then in the NCSS Data directory. Next, exit the Notepad program. We suggest that you use '.txt' as the extension for value label files. This will make them easy to find.
6. Press the *Select a V.L. File* button, select the appropriate file, and press the *Open* button to complete the selection.
7. Back on the *Specify the Value Labels* window, click the *Ok* button to complete the selection.
8. This completes the specification of the value label file for this variable. Note that if you knew the name of a previously created file, you could have just typed it in directly into the space provided.
9. To use the value labels in a specific procedure, you must set the Value Labels option (usually under the Format or Reports tabs) to *Value Labels*.

Creating Value Label Variable in the Database (Not Recommended)

Values labels are constructed from two contiguous variables. The variable on the left contains the original value. The variable on the right contains the corresponding label. For example, you might have several variables whose values range from 1 to 5. Suppose you want to label these values as shown below. Further suppose that variables C11 and C12 are empty columns on the spreadsheet. You would enter the five possible values in variable C11 and the corresponding labels in C12, as shown below.

C11	C12
1	Strongly Disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly Agree

Attaching the Value-Label Variables to Other Variables

Next, you must associate this set of value labels with those variables that contain this type of data. This is accomplished by entering the value label variable (C11 in this example) on the *Variable Info* sheet in the *Value Label* column. For example, suppose you wanted to attach this set of value labels to variable C4. You would specify C11 in the Value Label column of the Variable Info sheet in the row corresponding to C4. Now, when you use C4 in reports that display individual values (like crosstabs), the value labels in C12 will be displayed.

Note that the values of the original variable need not be numeric. You can associate value labels with a text values as well numeric. Also note that several variables can use the same value labels.

Copying Value Labels Among Databases

Value labels variables must be included on each database. You cannot read the value labels from another database. If you have a set of value label variables that you wish to use on more than one database, you will have to Copy and Paste them from one database to another. This is done by copying the value-label variables to the clipboard, opening the second database, and pasting the clipboard information onto it.

Rows / Observations

The rows of the database correspond to the observations or cases. Normally, you begin adding data at the top of the datasheet and work your way down.

Currently, there is a limit of 16,382 rows per S0 database. S0Z databases may have as many rows as your hard disk will hold.

Chapter 103

Spreadsheets

Introduction

This chapter discusses the operation of the *Spreadsheet*, one of the three main windows of the **NCSS** system. The other two windows, the Output window and Procedure window will be discussed in other chapters. The Spreadsheet is the window that lets you enter, view, and modify your data. It is the first screen that you are presented with when you start the program.

The operation of the **NCSS** spreadsheet is similar to the operation of other spreadsheets with which you are familiar. In fact, it has most of the operational features of Microsoft Excel. Since the operation of these spreadsheets is so common, we will not spend a lot of space teaching them to you. We will now go over the details.

First, we will discuss the menu bar which appears at the top of the spreadsheet window.

Spreadsheet Menus

You should be familiar with the operation of pulldown menus. We will discuss the various options that are on these menus.

File Menu

The File Menu controls the opening and closing of databases. Note that some of the basic File Menu operations are also provided on the Toolbar.

We will now discuss each of these options.

- **New**

  This option closes the current database (if any) and creates a new database. A dialog box appears that lets you select the type of database: S0 or S0Z. See the *Introduction* to the *Database* chapter for details on the differences between these two types of databases.

  When you create a new S0Z database, an expanded dialog box will appear that will allow you to specify the name of the database, the number of variables (columns), and the maximum length of the numeric and text values. You can increase the number of variables and the length of the record using the *Insert-Columns* operation.

- **Open**

  The Open option lets you open existing **NCSS** databases. It will cause the Open Dialog box to appear from which you can select a file. If you want to open a S0Z database, change the type of files that are scanned for by the dialog box. These files will have the extension “S0Z.”

  When selecting an S0 database, note that you select the file with the *S0* extension. Attempting to open **NCSS** files with other extensions (such as *S1*, *S2*, etc.) will produce unpredictable results.

Note that the database is copied into memory. Once you open a database, you actually have two copies of it—one in memory and one on your disk. No automatic relationship is maintained between the loaded database and the disk database. Changes made to the copy in memory will not automatically change your disk database files unless you save them!

- **Add a Sheet**

  Each datasheet contains 256 variables. This option lets you add an additional datasheet to your database. When selected, an additional Datasheet Tab will appear at the bottom of the spreadsheet and additional variables are added to the Variable Info sheet.

- **Import**

  This option lets you import data from various other spreadsheets, databases, and statistical systems into **NCSS**. There is a whole chapter devoted to this topic, so refer to that chapter for further details.

- **Remove Last Sheet**

  When your database (spreadsheet-type only) contains two or more datasheets, this option will remove the last one from the database. Note that the datasheet must be blank before it can be removed. If the last datasheet contains data, select the data and cut it to remove it. This will clear the datasheet so that it can be removed.

- **Printer Setup**

  This option brings up a window that lets you set parameters of your printer(s).

- **Page Setup**

  This option lets you specify the format of your datasheet printout. You can specify headers, footers, margins, page order (across or down), and scale (size of the print). It is often used to enable the printing of the row and column labels.

  Headers and footers can contain text and special formatting codes. The following table lists the special formatting codes. Header and footer codes can be entered in upper or lower case.

Format Code	Description
&L	Left-aligns the characters that follow.
&C	Centers the characters that follow.
&R	Right-aligns the characters that follow.
&D	Prints the current date.
&T	Printers the current time.
&F	Prints the worksheet name (this is an internal name that may not be useful).
&P	Prints the page number.
&P+*number*	Prints the page number plus number.
&P-*number*	Prints the page number minus number.
&&	Prints an ampersand.
&N	Prints the total number of pages in the document.

The following font codes must appear before other codes and text or they are ignored. The alignment codes (e.g., &L, &C, and &R) restart each section; new font codes can be specified after an alignment code.

Format Code	Description
&B	Use bold font.
&I	Use an italic font.
&U	Underline the header.
&S	Strikeout the header.
&"fontname"	Use the specified font.
&nn	Use the specified font size - must be a two digit number.

- **Print**

  This option prints the entire datasheet or a portion that you designate. Note that you must print each datasheet of a database separately. If the row and column labels do not show in your printout, select Page Setup from the File menu and check the appropriate selection boxes.

- **Save**

  This option saves the current database. Remember that a database consists of several files. All of those files will be replaced.

- **Save As**

  This option saves the current database to a database with a different name. For example, if you are working on a database called "XYZ" and will be making changes to it, you might want to save a copy of it as "XYZ1" so that any mistakes you might have made will not destroy your original data.

- **Export**

  This option lets you export the current database to another format for use in other spreadsheets, databases, etc. There is a chapter devoted to this topic, so we refer you to that chapter for further details.

- **Exit NCSS**

  This option quits the **NCSS** system.

- **Previously Open Files**

  A list of previously open databases is presented. You may select any of them to revert to that database directly.

Edit Menu

The Edit Menu controls the editing of databases. Note that some of the basic Edit Menu operations are provided on the Toolbar.

- **Undo**

  Undo allows you to undo the last edit operation made. Note that only the most recent edit operation may be undone. Also not that the undo only works for S0 databases. It does not work for S0Z databases.

  When you make wholesale changes to your database (by cutting or pasting, for example), the Undo system requires a lot of memory to store additional information needed for an undo operation. If you see system resources getting low, make an additional change to a single cell. This will reset the undo system and free up system memory.

- **Cut**

  The Cut option copies the currently selected (highlighted) data to the Windows clipboard and clears those cells. This data may be pasted at another location within **NCSS** or to another Windows program. We will discuss the process of selecting cells later in this chapter.

- **Copy**

  The Copy option copies the currently selected (highlighted) data to the Windows clipboard. The selected data is untouched. The copied data may be pasted at another location within **NCSS** or to another Windows program. We will discuss the process of selecting cells later in this chapter.

- **Paste**

  The Paste option copies data from the clipboard to the current datasheet at the currently selected location. The contents of the clipboard may have come from a previous Cut or Copy operation within **NCSS** or from another Windows program.

  For example, an easy way to analyze the means from an analysis of variance is to copy them from the Output screen and paste them in a datasheet. Furthermore, a quick way to import data from an *Excel* spreadsheet is to copy the data in *Excel* and paste it into an **NCSS** datasheet.

  The Paste option behaves differently depending on whether part of the datasheet is selected (highlighted). The data on the clipboard acts like a rectangular array of data (it has rows and columns). If there is no selection on the datasheet, the paste operation will place the data on your spreadsheet just as it appeared when it was copied to the clipboard. However, if the spreadsheet has a selected area, the paste operation will do two things. First, it will insert the data inside the selected area only (extra data will be omitted). Second, it will repeat either all or part of the clipboard so that the selected area is filled.

- **Paste Rotated**

  This option works like the Paste option (above), except that the data are rotated ninety degrees so that the rows become the columns and the columns become the rows. The following example shows the result of using this option:

  Data that was Copied

  1 2
  3 4
  5 6
  7 8

  Result of Paste Rotated Operation

  1 3 5 7
  2 4 6 8

- **Clear**

  The Clear option erases the data that is selected. Note that unlike the Cut option, the Clear option does not put the data on the clipboard.

- **Insert**

  The Insert option inserts rows or columns into your datasheet at the current position of the cursor. When you select Insert, a dialog box appears that allows you to indicate whether you want to insert rows or columns (variables). The number of rows (or columns) inserted is determined by the number of rows (or columns) selected.

  Hence, the steps to insert columns are as follows:

  1. Select the number of columns you want to insert, beginning your selection at the column where you want them added.
  2. Select Insert from the Edit Menu.
  3. Select Columns from the dialog box.

  A datasheet on an S0 database contains exactly 256 variables. When you insert new variables, the current variables are shifted to the right. The variables at the right of the datasheet are "pushed off" the datasheet and lost. For this reason, **NCSS** first checks to make sure that there are enough empty variables at the edge of the datasheet to accommodate the inserted variables. If you find that you don't have room to insert variables that you need, simply add a new datasheet, cut the last several variables at the right of the datasheet, and paste them to the next datasheet. This will make room for inserting variables.

  When you insert columns in a S0Z database, a new database is created with the new column count. The record size is increased by ten bytes for each new variables (assuming that you are using the default ten bytes per number).

- **Delete**

  The Delete option removes the currently selected rows or columns from your datasheet. When you select Delete, a dialog box appears that lets you indicate whether to delete rows (or columns). The number of rows (or columns) deleted is determined by the number of rows (columns) selected.

- **Fill**

  This option brings up the Fill window which fills the current variable (or the currently selected block of cells) with the value specified. The value may be incremented so that special patterns such as 1,2,3,4 may be easily generated.

- **Find**

  This option searches through your data for a designated value. Once you have started a find operation, use the Find Next button continues your search. You can search for a single digit or for the complete number.

- **Replace**

  The Replace option allows you to quickly replace data throughout your datasheet. You can replace only those cells that match a certain pattern, or you can replace individual letters and digits.

- **Variable Info**

  This main brings up a submenu that lets you modify information about the variable such as its name, label, format, data type, or transformation. Each of the items on the submenu call up dialog windows that let you modify the corresponding information.

- **Font**

  This option allows you to change the font of cells in the database.

- **Options**

  This brings up a window that allows you to change various options that control the way the program functions.

- **Serial Numbers**

  This brings up a window displaying the serial numbers being used and the product(s) licensed.

- **Enter File Name F7**

  This option launches the Windows dialog used to find a file on your computer. The file name is entered into the active cell on the spreadsheet. This is commonly used when entering file names in **GESS** procedures.

Data Menu

The Data Menu lets you display or modify the database. It also lets you reduce the number of rows that are processed using the Filter procedure.

- **Data Report**

  This option brings up the Data Report procedure which lets you create a nicely formatted printout of all or part of your data.

- **Sort**

  The Sort option lets you sort a datasheet by up to three variables. The sort may be either ascending or descending on a sorted variable. To sort your data, simply select the variables you want to sort on and click the Ok button.

Note that the sort procedure rearranges only the datasheet on which the variables reside. Other sheets are untouched.

- **Filter**

  This section explains how to use *filters* to limit which rows (observations) are used by the other procedures and which are skipped. For example, you might want to limit an analysis to those weighing over 200 pounds. You would use a filter to accomplish this.

  The filter tab is used to enter the desired filter statements, as well as all filter options.

Filter Specification

Filter System Active

This statement must be checked for the Filter to be activated.

Note: You must RUN this screen to activate (or de-activate) the Filter System.

Keep Spreadsheet Row If:

You can specify several filter expressions. This option specifies how these expressions are combined.

- **OR**

  If you select the 'OR' option, the condition on only one of the expressions must be met to retain the row in the analysis.

- **AND**

  If you select the 'And' option, the conditions in all of the expressions must be met in order to retain the row in the analysis.

Filter Statements

These boxes contain the filter statements. Each box may contain several filter expressions separated by semicolons.

Note that text must be enclosed in double quotes.

Example: C1<5; C2=4; C3=1,2,3; C4<C5; C6<C7+C8; C1<>Missing

SYNTAX:

The basic syntax of a filter statement is

VARIABLE LOGIC OPERATOR VALUE

where VALUE is an expression the yields a text value or a number constructed from variable names, numbers, and the symbols +, -, *, and / (add, subtract, multiply, and divide). Note that parentheses are not allowed. A list of values may be used.

Examples of valid VALUE expressions are

C1

Height

1.0

C1+5

Height/100

1,2,3,4,5

X+Y/100+4

Missing (may be used to indicate a missing value)

“John” (must be enclosed in DOUBLE quotes)

LOGIC OPERATOR

The Logic Operator is one of the following operators:

=	Equal to
<>	Not equal
<	Less than
<=	Less than or equal
>	Greater than
>=	Greater than or equal

Note that only one operator may be specified in an expression.

Variable Name Locator (for pasting variable names into filter statements)

The selected variable name may be copied to the clipboard and pasted into the filter statement.

This field provides no active options. It is here to let you have easy access to a list of all variable names on the current database.

Filter Specification - Filter Statement Comparison Option

Comparison Fuzz Factor

When you make a comparison, you may want to allow for a certain amount difference between two numbers that may occur because of rounding error, etc. For example, you may want the statement .3333=.3334 to evaluate to true instead of false. If the fuzz factor is set to .000001, this expression will be false. However, if the fuzz factor is set to .0001, then this expression will be true.

- **Merge Databases**

  This option brings Merging Two Databases procedure.

- **Enter Transform**

  This option brings up the Transformation window. A discussion of transformations is the subject of another chapter.

- **Recalc Current**

  The variable at which the cursor is located is recalculated using the transformation formula stored in the Variable Info sheet. Data that are currently in this column are replaced with the transformed values.

- **Recalc All**

  The whole database is recalculated, proceeding from left to right. Note that data that are currently stored in the database are replaced by the transformed values. If a variable does not have a transformation formula, its values are untouched.

- **If-Then Transform**

  This brings up the If-Then Transformations window. A discussion of if-then transformations is the subject of a later chapter.

- **Data Simulation**

  This option brings up the Data Simulator.

Analysis and Graphics Menus

These menus load the corresponding procedure windows. For example, select Descriptive Statistics, then Data Screening will load the Data Screening procedure window. This window controls the running of the Data Screening procedure.

Tools Menu

From this menu you can load the Macro Command Center, the Data Simulator window, or the Merging Two Databases window.

Window Menu

This menu lets you transfer to one of the other **NCSS** menus such as the Output window or one of the currently open procedure windows.

Help Menu

From this menu you can launch the **NCSS** Help System and view PDF documentation, tutorials, and references. From this menu you can also view serial numbers and licensing information.

Spreadsheet Toolbar

The toolbar is provided for single-click access to the most commonly used menu options. You will find that each of the options on the toolbar can also be found in the menus. The toolbar has a feature called a "tool tip." This means that when you hold the mouse pointer over a certain square for at least a second, a small help box will appear that explains what this particular toolbar button is for. Most of the buttons on the toolbar follow Windows standards, so you will recognize them right away.

The last eight buttons on the toolbar represent procedures that you may transfer to. These are completely customizable. You can designate which eight procedures you want to be able to transfer to by right clicking on them in the Navigator window.

Cell Reference

Two boxes at the bottom of the screen give the *Cell Reference*. The first box gives the variable (column) number of the current location of the cell cursor. The second box gives the row number of the current location of the cell cursor. The cell cursor is the active cell. You can recognize the current cell because it will have an extra dark border.

Cell Edit

The Cell Edit box provides an alternate place to edit data. As you move around the spreadsheet, the contents of the active cell are copied to this Cell Edit box. Occasionally, your data will be longer than can easily be displayed in the cell. Although you could reset the column width, you usually find it easier to edit the data in the Cell Edit box.

Any changes you make will not be entered into the datasheet until you hit the ENTER key or position the mouse on another cell. After making changes to data in the Cell Edit box, you can press the ESC key to withdraw the changes.

Datasheet

This section of the screen shows the data. We will now describe how to use the spreadsheet to modify the data contained in a datasheet. You should know how to select cells, ranges, rows, and columns. Your work will go much faster if you learn how to quickly enter, modify, and delete data. These will all be described in this section.

Navigating the Datasheet

This section describes how to move around the datasheet using the keyboard and the mouse. In addition to moving around the datasheet, we will also describe how to make selections, copy data, and move data.

Active Cell

The datasheet cursor is always located on a single cell, even when a range of cells is selected. The cell on which the datasheet cursor is located is called the *Active Cell*. Any typing that is done will only affect the active cell. The contents of the active cell are displayed in the Cell Edit box. The address of the active cell is displayed in the Cell Reference boxes.

Keyboard Commands

The following commands are used mainly for data entry.

Key	Description
ENTER	Accepts the current entry and moves the active cell down one cell. When a range of cells is selected, accepts the current entry and moves down to the next selected cell. When the bottom of the selection is reached, the active cell moves to the top of the next selected column to the right.
SHIFT-ENTER	Acts like the ENTER key, except that cell-to-cell movement is upward and to the left instead of downward and to the right.
TAB	Accepts the current data entry and moves the cursor one cell to the right. When a range of cells is selected, the tab moves the cursor to the right to the next cell in the selection.
SHIFT-TAB	Acts just like the TAB key, except that cell-to-cell movement is to the left instead of to the right.
F2	Enters edit mode. Pressing F2 a second time brings up a cell-text data entry box.
DEL	Clears the current entry or selection.
ESC	Cancels the current data entry.

The following commands are used mainly for moving about the datasheet.

Key	Description
UP ARROW	Moves the active cell up one row.
DOWN ARROW	Moves the active cell down one row.
LEFT ARROW	Moves the active cell left one column.
RIGHT ARROW	Moves the active cell right one column.
CTRL UP / DOWN / LEFT / RIGHT ARROW	Moves to the next range of cells containing data. If there is no additional data in any of the cells in that direction, the active cell is moved to the edge of the datasheet.
PAGE UP	Moves up one screen.
PAGE DOWN	Moves down one screen.
CTRL PAGE UP	Moves left one screen.
CTRL PAGE DOWN	Moves right one screen.
HOME	Moves to the first column in the current row.
END	Moves to the last column in the current row that contains data.
CTRL HOME	Moves to the upper-left corner of the datasheet (cell 1,1).
CTRL END	Moves to the last row and column that contains data.
SCROLL LOCK	Modifies the action of the above movement keys. This key causes the datasheet window to scroll without changing the current selection. It works with all movement keys except HOME, END, CTRL HOME, and CTRL END.
SHIFT + any movement key	Extends the current selection in the direction indicated.

Mouse Actions

The mouse is used mainly for positioning the active cell and making selections. You can also use the mouse to move a block of cells around the datasheet.

Action	Description
Left Click	Accepts the current entry and moves the active cell to the position of the mouse.
Right Click	Does nothing.
Left Click in a Row or Column Heading	Selects the entire row or column.
Left Double Click	In-cell editing is invoked.
Right Double Click	Does nothing.
Left Click and Drag	Selects a range of cells. If other ranges were selected, they are unselected.
CTRL + Left Click and Drag	Selects a range of cells. If other ranges were selected, they remain selected. Note that edit commands, such as cut and copy, will only work on a single range.
SHIFT + Left Click and Drag	Extends the current selection in the direction indicated.
Dragging a Selection's Copy Handle	Copies the selection to a new location. The copy handle is the small plus sign at the lower-right corner of a selection.
Dragging a Selection's Border	Moves the selection to a new location. Note that when you move data around the datasheet in this fashion, no attempt is made to update the variable names. You will have to do that manually. For this reason, it is best to do this kind of wholesale editing before you attach names and transformations to the variables. Note also that Cell Transformation formulas are updated.

Selecting Cells

Many operations require one or more cells to be selected. There are three types of selections: a single cell, a single rectangular range of cells, and multiple ranges of non-adjacent cells. Cells may be selected either with the mouse or with the keyboard.

Selecting Cells with the Mouse

To select a range of cells with the mouse, click and hold the left mouse button down on the upper-left cell of the range you want to select. Drag the mouse cursor to the lower-right cell, while continuing to hold the left mouse button down. When the desired cells are selected, release the mouse button.

To select multiple ranges with the mouse, press the CTRL key while making each additional selection. Note that multiple selections are only useful for controlling cell cursor movement during data entry. They are not used by any of the edit functions.

To select an entire row or column, click on the row or column heading.

Once a range is selected, you can move the active cell within the selection using the Enter, SHIFT+ENTER, TAB, and SHIFT+TAB keys without destroying the selection.

Selecting Cells with the Keyboard

To select a range of cells with the keyboard, position the active cell at the upper-left corner of your desired selection. While holding down the SHIFT key, use the cursor movement keys (such as the arrow keys) to move to the lower-right corner of the selection.

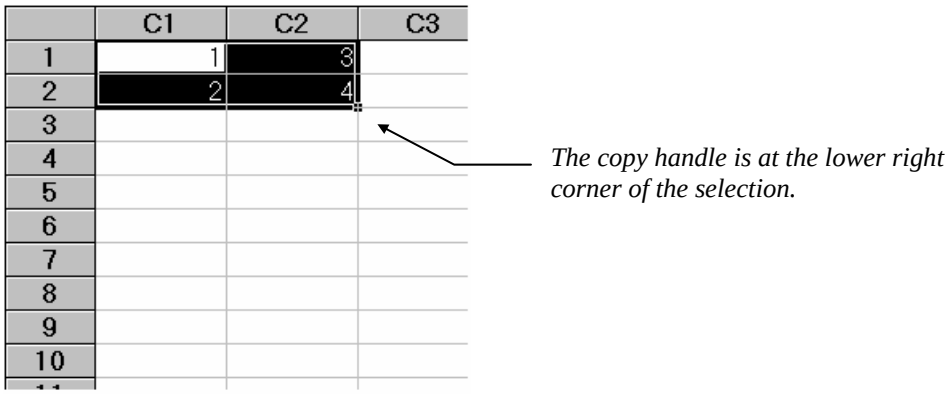
Editing Datasheets Interactively

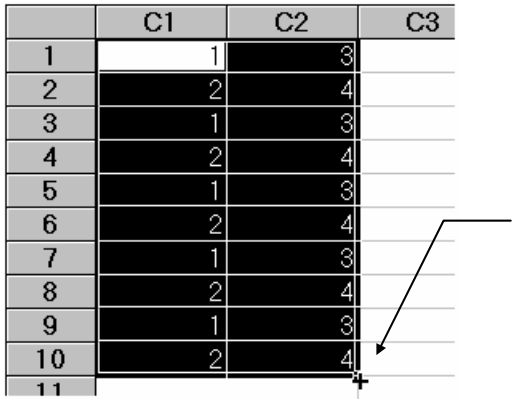
Data can be entered into a datasheet in many different ways. In this section, we will explain how you can quickly move and copy ranges of cells by clicking and dragging the copy handle of a selection.

Copying Data Interactively

You can copy a range of cells quickly by using only a few clicks of your mouse. The steps for doing this are enumerated next followed by figures that illustrate the action.

1. Select a range of one or more cells.
2. Select the *copy handle* (the small crosshair that appears at the lower-right corner of a selection) with your mouse cursor by positioning the mouse cursor over the copy handle and pressing the left mouse button.
3. Drag the copy handle through the range of cells that are to receive the copied data.
4. Release the left mouse button.





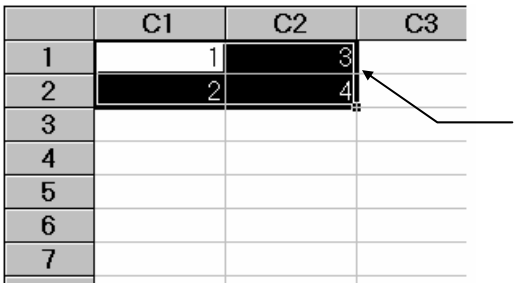
The cursor changes to a crosshair as the copy handle is dragged down.

Note that the copied selection is repeated so that the whole copied area has data.

Moving Data Interactively

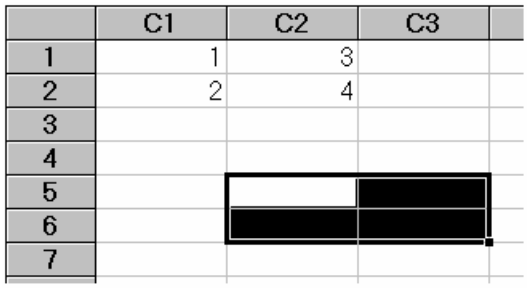
You can move a range of cells quickly by using only a few clicks of your mouse. The steps for doing this are enumerated next followed by figures that illustrate the action.

1. Select a range of one or more cells.
2. Position the mouse cursor on the border of the selection. When positioned on the border, the pointer changes to an arrow.
3. Drag the selection to the new location.
4. Release the left mouse button.

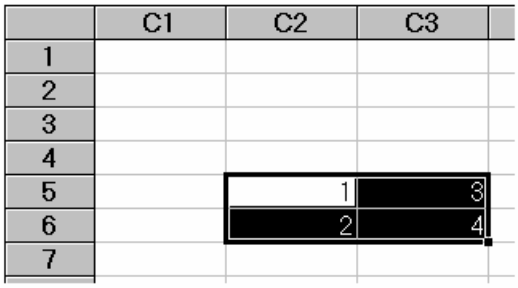


First, select the cells you wish to move.

Next, position the mouse pointer at the border of the selected area. The mouse pointer will change to an arrow. Once the pointer has changed, do not release the mouse button.



Move the selected cells to their new location.



Release the left mouse button. The contents of the selected cells will move to the new location.

Caution: the variable names have not changed. You will have to manually adjust variable names and transformations when you move data using this method.

Changing Row Heights and Column Widths

You can interactively resize the height of a row or the width of a column using the mouse. Position the pointer on the right edge of a column heading or the bottom edge of a row heading. The pointer will change shape. Simply drag the pointer to resize the row or column.

If multiple rows are selected when you resize a row, all selected rows are resized as you drag a row border. Multiple columns can be resized in like manner.

You can also set the size of a selected group of columns or rows to equal the size of another row or column. First, select the group of columns or rows you want to resize, including the column or row whose size you want to match. Next, click the right border of the column header or the bottom border of the row whose size you want to match. The columns or rows will all be resized.

Datasheet Tabs

The Datasheet Tabs at the bottom of the spreadsheet let you move quickly from one datasheet to another. These tabs are especially useful for allowing you to move to the Variable Info sheet to make changes to variable names, transformation, and formats.

Chapter 104

Merging Two Databases

Introduction

Occasionally, it is useful to merge two databases according to the value of one or more common (index) variables. This module allows you to merge two databases, or, alternatively, update one database with the contents of another.

For example, consider the following dataset, called COUNTY, which contains county-level information for two states.

COUNTY dataset

State	County	Pop	Age	Income
TX	1	72	34	65
TX	2	33	42	45
TX	5	25	23	46
TX	6	54	36	65
TX	7	11	42	53
TX	8	28	25	62
TX	9	82	35	66
TX	10	5	40	75
TX	11	61	27	22
MD	2	5	23	69
MD	4	98	25	73
MD	3	64	29	75
MD	2	36	24	65
MD	1	24	25	66
MD	5	34	31	78
MD	6	89	22	81
MD	8	21	25	73
MD	7	21	30	62

A second dataset, called STATE, contains similar information at the state level.

STATE dataset

State	Pop	Age	Income	Education
TX	23543	32	54	10.2
MD	10343	29	69	10.3
IN	5231	41	35	10.1
CA	29587	35	67	10.4
NY	18142	34	78	10.2

Suppose that we wish to update the county dataset with the corresponding information from the state dataset. The resulting dataset, called COUNTYSTATE, might appear as follows.

COUNTYSTATE dataset

State	County	Pop	Age	Income	St Pop	St Age	St Income	St Education
TX	1	72	34	65	23543	32	54	10.2
TX	2	33	42	45	23543	32	54	10.2
TX	5	25	23	46	23543	32	54	10.2
TX	6	54	36	65	23543	32	54	10.2
TX	7	11	42	53	23543	32	54	10.2
TX	8	28	25	62	23543	32	54	10.2
TX	9	82	35	66	23543	32	54	10.2
TX	10	5	40	75	23543	32	54	10.2
TX	11	61	27	22	23543	32	54	10.2
MD	2	5	23	69	10343	29	69	10.3
MD	4	98	25	73	10343	29	69	10.3
MD	3	64	29	75	10343	29	69	10.3
MD	2	36	24	65	10343	29	69	10.3
MD	1	24	25	66	10343	29	69	10.3
MD	5	34	31	78	10343	29	69	10.3
MD	6	89	22	81	10343	29	69	10.3
MD	8	21	25	73	10343	29	69	10.3
MD	7	21	30	62	10343	29	69	10.3

Note that only those states from the States dataset that were included on the County dataset are transferred to the resulting CountyState database.

Missing Values

The basic principle governing the treatment of missing values is that they cannot be matches. That is, even though values for corresponding by variables are both blank (missing), they will not be considered as a match. Only matches of non-missing values are recognized the program.

Procedure Options

This section describes the options available in this procedure.

Merge Tab

This panel specifies the datasets and variables to be merged.

Datasets

Merge (A)

This is the fully-qualified name of the first dataset. The contents of the second dataset will be merged with this one according to the values of the corresponding By Variables.

With (B)

This is the fully-qualified name of the second dataset. The contents of this dataset are merged with dataset A. Only rows in this dataset that have matching values in the corresponding 'By Variables' will be kept.

To Make

This is the fully-qualified name of the dataset produced by the merge operation. '*Fully-qualified*' means that the drive and folder containing the dataset are included in the entry.

The type of database created depends on the extension that you use. Specify a regular spreadsheet by using a name that ends with '.S0'. Specify a database by using a name that ends with '.S0Z'. You must use the '.S0Z' format if the dataset will have more than 16000 rows. If you specify an .S0Z database, the parameters of the new database are specified under the Options tab.

Existing data in this dataset will be replaced. So do not select a dataset that contains data you need to keep.

Merge By Matching Values from the Following Variable Pairs

Match this Variable from Dataset A

Specify a single variable from dataset A whose values will be used by comparing them with those of the corresponding variable from dataset B.

Note that blanks in both values are not considered to be a match!

With this Variable from Dataset B

Specify a single variable from dataset B whose values will be used by comparing them with those of the corresponding variable from dataset A.

Note that blanks in both values are not considered to be a match!

Options for Datasets A and B

Copy these Variables from A (or B)

Select the variables from dataset A (or B) that are to be retained in the resulting dataset. Note that this does not include the match variables, since they are included automatically.

Prepend to Names (Datasets A & B)

Specify a few letters to be added to the beginning of each variable name that was kept from the dataset corresponding dataset (A or B). This allows you to rename variables when there are names that are common to both datasets that might cause confusion in the resulting database.

If you do not want to change the variable names of a dataset, leave this option blank.

For example, if the variable names that were kept were 'TIME' and 'AMOUNT', and 'AA_' is specified here, the resulting variable names are 'AA_TIME' and 'AA_AMOUNT'.

Note the only letters, integers, and the underscore may be added.

Append to Names (Datasets A & B)

Specify a few letters to be added to the end of each variable name that was kept from the dataset corresponding dataset (A or B). This allows you to rename variables when there are names that are common to both datasets that might cause confusion in the resulting database.

If you do not want to change the variable names of a dataset, leave this option blank.

For example, if the variable names that were kept were 'TIME' and 'AMOUNT', and '_TOTAL' was specified here, the resulting variable names would be 'TIME_TOTAL' and 'AMOUNT_TOTAL'.

Note the only letters, integers, and the underscore may be added.

Keep All Rows in this Dataset (A)

Check this box if you want to keep all rows from dataset A in the resulting database, even if they do not have a match in dataset B.

If this box is not checked, rows with missing values in the 'By Variable' and rows that do not have a match in dataset B will be omitted from the resulting dataset.

Keep All Rows in this Dataset (B)

Check this box if you want to keep all rows from dataset B in the resulting database.

If this box is not checked, rows with missing values in the 'By Variable' and rows that do not have a match in dataset A will be omitted from the resulting dataset.

Options Tab

These options control the resulting S0Z database when it is used.

Special Options Used When the New Dataset is a Database (Extension=S0Z)

Text Variables

Additional space is added to the end of each record for this many extra long (greater than six characters) text variables.

Field Length

This is the length of a data-cell's field. This must be at least ten. You can make this larger for longer text values. Each text value requires four additional bytes of storage.

Text Length

The length of the extra text fields that are stored at the end of each record of a database.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Merging Two Datasets

This section presents an example of how to merge the two datasets County.s0 and State.s0 shown in the example above. Note that the folders shown here may be different from those on your machine. You will have to make appropriate changes to the folder names.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Merging Two Databases window.

1 Open the Merge Two Databases window.

- On the menus, select **Data**, then **Merge Two Databases**. The procedure window will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the datasets.

- Specify the **Merge (A) Dataset** as **C:\Program Files\NCSS2007\DATA\County.s0**.
- Specify the **With (B) Dataset** as **C:\Program Files\NCSS2007\DATA\State.s0**.
- Specify the **To Make Dataset** as **C:\Program Files\NCSS2007\DATA\CountyState.s0**.
- Set **Match this Variable from Dataset A** as **State**.
- Set **With this Variable from Dataset B** as **State**.
- Set **Copy these Variables from A** as **County-Income**.
- Set **Copy these Variables from B** as **Pop-Education**.
- Check the **Keep All Rows in Dataset for Dataset A** option.
- Set **Append to Names for dataset B** as "**_Total**" (no quotes).

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

This procedure does not produce any output. Instead, the resulting dataset is opened in the spreadsheet.

Chapter 105

Procedures

Introduction

Each procedure has its own Procedure window. The Procedure window contains all the settings, options, and parameters that control a particular procedure. These options are separated into groups called *panels*. A particular panel is viewed by pressing the corresponding *panel tab* that appears just below the toolbar near the top of the window.

The current values of all options available for a procedure are referred to as a *template*. The template may be stored for future use in a *template file*. By creating and saving template files (often referred to as *templates*), you can tailor each procedure to your own specific needs. For example, the multiple regression procedure has about twenty different reports available. You can select the four or five reports that are useful to you and disable the rest. Your selection can be saved as a template. Each time you use the multiple regression procedure, you simply load your template and run the analysis. You do not have to set all the options every time.

Note that up to six Procedure windows can be opened at a time.

Default Template

Whenever you close a procedure, the current settings are automatically saved in a default template file named Default. This template file is automatically loaded when the procedure is next opened. This allows you to continue using the template without resetting all of the options.

Menus

File Menu

The File Menu is used for initializing, loading, and saving a copy of a template. Each set of options for a procedure, called a template, may be saved for future use. In this way, you do not have to set the options every time you use a procedure. Instead, you set the options the first time, save them as a template, and re-use the template whenever you re-use the procedure.

- **New Template (Reset)**

  This menu item resets all options to their default values.

- **Open Template Panel**

  This option sets the Template panel as the active procedure panel. This panel lets you load or save template files. It displays all templates associated with this procedure along with the Template Id (the optional phrase at the bottom of the window).

Saving a Template
To save a template, enter the name you want to give the template file in the File Name box. You may also enter an identifying phrase in the box at the bottom of the window since this will be displayed along side of the file names. Finally, press the Save Template button to save the file.

Note that there is no automatic connection between the template in memory and the copy on the disk. If you want to save the changes you have made to a template, you must use the Save Template option to save them.

Loading a Template
To load a template file, select it from the list of files given in the Template Files box. Press the Load Template button to load the template.

- **Close Procedure**

  This option closes this Procedure window.

- **Save Template**

  This option saves the current option settings to the template file that is currently specified on the Template panel. You can be viewing any panel of the procedure when you issue this command—you do not have to be viewing the Template panel.

  You should note that the templates for each procedure have different file name extensions. Thus, you can use the same name for a template in the t-test procedure as for a template used in the multiple regression procedure.

  Also note that the templates are stored in a subdirectory of the **NCSS** directory. You can erase any of the templates you want by deleting them from this directory.

  Note the Save button on the toolbar provides this same operation. It may be more convenient than selecting this menu item.

- **Printer Setup**

  This option allows you to change the printer settings.

- **Exit NCSS**

  This option terminates the **NCSS** system. Before using this option, you should save all datasheets and output documents.

Run Menu

The Run Procedure option runs the analysis, displaying the output in the Output document of the word processor. After you have set all options to their appropriate values, select this option to perform the analysis.

Note that the procedure may alternatively be run by pressing the F9 function key or by pressing the left-most key on the toolbar.

Analysis and Graphics Menus

These menus let you load any of the analysis or graphics procedure windows.

Tools Menu

From this menu you can load the Macro Command Center, the Data Simulator window, or the Merging Two Databases window.

Window Menu

This menu lets you display any of the other windows in the **NCSS** system that are currently open such as the Output window, the Data window, the Navigator window, or any procedure window.

Help Menu

From this menu you can launch the **NCSS** Help System and view PDF documentation, tutorials, and references. From this menu you can also view serial numbers and licensing information.

Entering Options

Your settings and selections are entered on the panel. The panel consists of several types of windows objects such as text boxes, check boxes, list boxes, and buttons. Each of these is used in the normal fashion.

Entering Text

When text (either numeric or letters) is needed for a particular option, you will be allowed to type text in the box. Many of these text boxes also have a pull-down button on the right. Pressing this button will allow you to select an option from a list of typical values, rather than type in the value.

Selecting from a List

Some options require you to select from a list. In this case, a dropdown list will allow you to choose from the selections available.

Selecting One or More Variables

When an option needs one or more variable names, you can type the names directly into the box or you can double-click in the box to bring up the Variable Selection window. This window lets you select one or more variables from those on the current database.

When using the Variable Selection window, you select one or more variables from the top window and they are listed in the button window. You can use the Shift key to select a list of contiguous variables. Use the Ctrl key to select disjoint (non-contiguous) variables.

At times, it may be more convenient to store the variable numbers rather than the variable names. Use the Store as Number button to indicate how you want the variables stored.

Toolbar

The toolbar is a series of small buttons that appear just below the menus at the top of the Procedure window. Each of these buttons provides quick access to a menu item. For example, the first button performs the same action as selecting the Run Procedure item from the Run menu.

Near the end of these buttons is a series of eight buttons that you can customize to represent your favorite procedures. This customization is done in the Navigator window. Pressing any of these buttons will load the corresponding Procedure window.

Chapter 106

Output

Introduction

NCSS sends all statistics and graphics output to its built-in word processor from where they can be viewed, edited, printed, or saved. Reports and graphs are saved in rich text format (RTF). Since RTF is a standard document transfer format, these files may be loaded directly into your word processor for further processing. You can also cut and paste data onto an **NCSS** datasheet for further analysis. This chapter covers the basics of our built-in word processor.

Documents

The **NCSS** word processor maintains two documents: *Output* and *Log*. Although both of these documents allow you to view your data, the *Output* document serves as a viewer while the *Log* document serves as a recorder.

You can load additional documents as well. For example, you might want to view the output from a previous analysis to compare the results with the current analysis. To do this, you open a third document that is actually the log file from a previous analysis.

All **NCSS** documents are stored in the RTF format. This is a common format that is used by most word processors, including Word and WordPerfect. When you save an **NCSS** report, you will be able to load that report directly into your own word processor. All text, formatting, and graphics will appear in your word processor ready for further editing. You can then save the document in your word processor's native format. In this way, you can easily transfer the output of an **NCSS** procedure to almost any format you desire.

Output Document

The Output document displays the output report from the current analysis. Whenever you run an **NCSS** procedure (like t-test or histogram), the resulting reports and graphs are displayed in the Output document. Each new run clears the existing Output document, so if you want to save a report, you must do so before running the next report.

The Output document provides four main functions: display, print, save to the Log document, and save as an RTF file.

Log Document

The Log document provides a place to store a permanent record of your analysis. Since the Output document is erased by each new analysis, you need a place to store your permanent work. The Log document serves this purpose. When you have a report or graph that you want to keep, copy it from the Output document to the Log document.

The Log document provides four main word processing functions: load, display and edit, print, and save. When you load a file into the Log document, you can add new output to it. In this way, you can record your work on a project in a single file, even though your work on that project is spread out over several days.

Output Menus

You should be familiar with the operation of pull-down menus. We will discuss the various options that are on these menus

File Menu

The File Menu is used for opening, saving, and printing **NCSS** word processor files. All options apply to the currently active document (the document whose title bar is selected). We will now discuss each of the options on this menu.

- **New**

  This option opens an empty document. You might use this when you want to make notes about your analysis.

- **New Log**

  This option opens an empty log document. You might use this when you want to start a new project.

- **Open**

  This option opens an existing file. When this item is selected, the Open Report File dialog box appears. Note that no connection is maintained between a loaded file and its image on the disk. If you make changes to a file, you must save those changes to the disk.

- **Open Log**

  This option opens an existing log file. When this item is selected, the Open Report File dialog box appears. The requested file is loaded into the Log document. Note that no connection is maintained between a loaded file and its image on the disk. If you make changes to a file, you must save those changes to the disk.

  You might use this option when you want to continue using a certain file as the Log file.

- **Toggle Auto-Log**

  When this menu item is checked, the output is automatically added to the bottom of the log file. If it is not checked, you must manually add the output to the log file by selecting “Add Output to Log.” To change this item from off to on or on to off, select it from the menu.

- **Add Output to Log**

  Selecting this option automatically copies the contents of the Output document to the Log document. The Output document remains unchanged. This allows you to save the current output document for further use.

- **Save As**

  This option lets you save the contents of the currently active document to a designated file using the RTF format. Note that only the active document is saved. Also note that all file names should have the “RTF” extension so that other systems can recognize their format.

- **Printer Setup**

  This option brings up a window that lets you set parameters of your printer(s).

- **Print Preview**

  This option allows you to preview the document before printing.

- **Print**

  This option lets you print the entire document or a range of pages. When you select this option, a Print Dialog box will appear that lets you control which pages are printed.

- **Close Output Window**

  Clears and minimizes the document. Note that this option will clear the Output and Log documents, but it will not close them since these two documents must remain open.

- **Exit NCSS**

  This option exits the **NCSS** system. All documents and databases are closed.

Edit Menu

This menu contains options that let you edit a document.

- **Undo**

  This item reverses the last edit action. It is particularly useful for replacing something that was accidentally deleted.

- **Cut**

  This item copies the currently selected text to the Windows clipboard and erases it from the document. You can paste the information from the clipboard to a different location in the current document, into another document, into a datasheet in the spreadsheet, or into another application. The selected text is erased.

- **Copy**

  This item copies the currently selected text from the document to the Windows clipboard. You can paste this information from the clipboard to a different location in the current document, into another document, into a datasheet in the spreadsheet, or into another application. The selected text is not modified.

- **Paste**

  This item copies the contents of the clipboard to the current document at the insertion point. This command is especially useful for moving selected information from the Output document to the Log document.

- **Select All**

  This item selects the entire document. Although you can select a portion of the document using the mouse or a shift-arrow key, this is much faster if you want to select the entire document.

- **Toggle Page Break**

  Changes the status of the page break on the line at which the insertion point resides. If a page break exists (shown by a horizontal line), it is removed. If a page break does not currently exist at that point, one is added.

  Note that **NCSS** does not repaginate your document for you. Once you make changes, it will be up to you to repaginate your document.

- **Find**

  This item opens the Search dialog box. You can specify text that you want to search for. This is especially useful when you are looking for a certain topic or data value in a large report.

- **Find Next**

  This item continues finding the text you entered in the Search Dialog box.

- **Replace**

  This item opens the Search and Replace dialog box. This allows you to quickly make repetitive changes. For example, you might want to change the name of one of the variables to a more useful name.

- **Goto Section**

  This item does not modify the document. Instead, it lets you reposition the insertion point to one of the major topics. When **NCSS** runs a procedure, it stores the major report topics in this list box. You can quickly position the view to a desired topic using this screen.

View Menu

This menu lets you designate which editing tools you want to use.

- **Ruler**

  This option controls whether the ruler and the tabs bar are displayed. The ruler displays the physical dimensions of the document. The tabs bar, found just below the ruler bar, lets you set the margins and tabs of your document. Only the currently selected part of your document is affected by a change in the tabs and margins.

- **Format Toolbar**

  The Format Toolbar lets you make formatting changes to the currently selected text. The function of each of the buttons is shown below.

- **Status Bar**

  The Status Bar shows the current position of the insertion point (cursor).

- **Show All**

  Selecting this menu item causes the ruler, tabs bar, format toolbar, and status bar to be displayed.

- **Hide All**

  Selecting this menu item causes the ruler, tabs bar, format toolbar, and status bar to be hidden. This gives you more screen space to view your output.

- **Redraw**

  Selecting this menu item causes the output to be redrawn.

Format Menu

This menu lets you set the format for a selected block of text.

- **Font**

  This option displays the Replace Font dialog box, which lets you specify the font and style of the selected text.

- **Paragraph**

  This option displays the Paragraph dialog box, which lets you specify the tabs and margins of the selected text.

- **Format Markers**

  Indicates whether the (usually hidden) tab arrows and the end-of-paragraph marks are displayed in the document. Note that these characters are never printed.

Window Menu

This menu lets you designate how you want the documents arranged on the screen and which **NCSS** window you want displayed on top of your output desktop.

- **Cascade**

  This item arranges the documents in a cascading display from the upper left to the lower right of the screen.

- **Tile Horizontally**

  This item arranges the documents horizontally across the word processor window.

- **Tile Vertically**

  This item arranges the documents vertically down the word processor window.

- **Arrange Icons**

  When a document is minimized, it is represented as an icon at the bottom of the word processor window. This option arranges all document icons. It is usually applied when the word processor window has been resized.

- **Current Output**
  This item causes the Output window to be displayed on top of your desktop.
- **Log**
  This item causes the Log window to be displayed on top of your desktop.
- **View Data**
  Causes the Spreadsheet window to be displayed on top of your desktop.
- **Navigator**
  Causes the **NCSS** Navigator window to be displayed on top of your desktop.
- **PASS Home**
  Causes the **PASS** Home window to be displayed on top of your desktop.
- **Quick Launch**
  Causes the Quick Launch window to be displayed on top of your desktop.

Help Menu

From this menu you can launch the **NCSS** Help System and view PDF documentation, tutorials, and references. From this menu you can also view serial numbers and licensing information.

Chapter 107

Navigator and Quick Launch

Introduction

The **NCSS** Navigator window lets you quickly and easily find the appropriate statistical or graphical procedure. Designed in an outline format, it lists every procedure in the system along with a brief paragraph that describes what the procedure is for and when it might be used.

The Navigator window also lets you configure the eight procedure buttons that appear on the toolbars of the Data, Output, and Procedure windows. These buttons give you immediate access to your favorite procedures.

The **NCSS** Quick Launch window is an alternative method for finding statistical and graphical procedures. The Quick Launch window shows all the procedures of the program on a single tab. This window may also be used for configuring the eight procedure buttons of the toolbar.

Using the Navigator

The Navigator window is very easy to use. The Navigator window may be loaded by selecting the Navigator item from any Window menu or by clicking the globe icon on any of the toolbars.

The Navigator window has a set of menus, a toolbar, and then a large display area. On the left side of the display area is an outline list of all the statistical and graphical procedures in the system. On the right side of the display area is a window that will display a brief paragraph explaining the main purpose of the currently selected procedure.

Navigator Menus

Below is a description of each of the four menus that appear at the top of the Navigator window.

Outline Menu

- **Collapse Outline**

  This option collapses the outline so that only the main heading is displayed.

- **Expand to First Level**

  This option expands the outline so that the main headings and first-level subheadings are displayed.

- **Expand All**
  This option completely expands the outline so that all entries are displayed.
- **Bold Text**
  This option toggles the bolding of the text.
- **Goto Selected Procedure**
  This option loads the currently selected procedure's window.
- **Close the Navigator**
  This option closes the Navigator window.

Tools Menu

This menu allows you to open tool procedures such as Macros or Merge Databases.

Window Menu

This menu allows you to open other windows such as the Data window or the Output window.

Help Menu

This menu provides access to the help system, release date/version information, the serial number editing window, and printable electronic (PDF) documentation.

- **Help**
  This option launches the help system.
- **About**
  This option provides information about the release date and versions of ***GESS***, ***NCSS***, and ***PASS*** that you are currently using, the current filter status, and instructions for citing this software in publications.
- **View PDF**
  This option launches your PDF viewer to display the appropriate electronic documentation file.

Using the Quick Launch Window

The Quick Launch window may be loaded by selecting the Quick Launch item from any Window menu or by clicking the launch icon on any of the toolbars.

The Quick Launch contains a button corresponding to each statistical and graphical procedure in the system. As you mouse over each button, a brief paragraph explaining the main purpose of the currently selected procedure will appear in the message box to the right. The procedure name will also appear near the button.

A procedure window is launched by clicking on the corresponding button.

Documentation Access through Quick Launch

The Quick Launch window may be used to access the complete set of documentation in pdf format. Once the Quick Launch window is open, click on the Documentation tab. The pdf files are loaded by clicking on the corresponding pdf file button.

Toolbar

The toolbar gives you one-click access to several of the menu items. The menu item assigned to each button on the toolbar is displayed when the mouse is held over the button for a few seconds.

Customizing the Toolbars – Navigator

The eight procedure buttons that show up on all toolbars throughout the program may be changed from the Navigator. The process of assigning one of these eight buttons a new procedure is as follows:

1. Find and select the procedure in the outline section (left-side of main window) of the Navigator window.
2. Click on the button you want to assign the procedure to with the **right** mouse button.

The icon of the selected procedure will now appear in all toolbars throughout the program.

Customizing the Toolbars – Quick Launch

The eight procedure buttons that show up on all toolbars throughout the program may be changed from the Quick Launch window. To add (or change) a procedure to one of the eight toolbar buttons, click on the desire procedure, and drag it to the desired button on the Quick Launch toolbar. Release the mouse button. The new icon will replace the previous icon.

The icon of the selected procedure that was dragged and dropped will now appear in all toolbars throughout the program.

107-4 Navigator and Quick Launch

Chapter 115

Importing Data

Introduction

Data from a wide variety of spreadsheets, databases, and statistical systems can be imported. Following is a list of the types of files that can be imported. Other file types and more current versions are being added. If the file you need to import is not on this list, check the online help. It may have been added since this manual was produced.

List of Imported Files

File Name	File Type	Versions	File Extension
Access	Database	1.0, 2.0	MDB
Alpha Four	Database		DBF
ASCII Delimited	Text File		TXT
ASCII Fixed Format	Text File		TXT
BMDP	Stat System	Classic	POR
Clipper	Database		DBF
dBase	Database	II,III,III+,IV	DBF
DIF	Spreadsheet		DIF
Excel	Spreadsheet	2, 3, 4, 5	XLS
Gauss	Stat System		DAT
Lotus 123	Spreadsheet	2.1, 3.x, 4.0	WKS, WK1
NCSS	Stat System	5.0 (DOS)	LAB, DAT
Paradox	Database		DB
Quattro	Spreadsheet		WKQ
SAS	Stat System	Native or Export	SSD or TPT
Solo	Stat System	4.0 (DOS)	LAB, DAT
SPSS	Stat System	Export	SAV or POR
Stata	Stat System		DTA
Symphony	Spreadsheet	1.0, 1.1, 2.0, 2.1,2.2	WRK, WR1
Systat	Stat System		SYS

Import Limitations

The only limitation in the importing files is that **NCSS** spreadsheet files can only accept 16,384 rows of data. Hence the files that you import must be smaller than this. If the files are larger, use an NCSS S0Z database to receive your data.

How to Import a File

With a few exceptions (that will be discussed later in this chapter), the steps you need to take to import a file are as follows:

Step 0 – Display the Import a File Window

Load the Import a File window by selecting **Import** from the File menu of the NCSS Data window.

Step 1 – Select the File

Select the type of file

Indicate the type of file to be imported. Several files share common characteristics, so you must tell the program what the file type is. Depending on the type of file you select, the screen will be modified to present additional options.

Select a file name

Use the Select a File to Import button to bring up a dialog box that designates the file you want to import. Once selected, the name will appear in the Current File Name box.

You may specify the number of rows imported by changing the entry in the Rows box. Enter the number of rows or 'All' to import all rows in the file. This option allows you to test the importing of long files with many thousands of rows.

If you are importing an ASCII file, two additional options are available. The Names Row box lets you indicate on which row the variable names reside. This option is especially useful in the case of spreadsheet files in which the data are not located at the top of the file. The Lines/Obs'n box lets you specify the number of rows in the file that make up one observation.

Next button

Press the Next button to proceed to step 2.

Step 2 – (Additional Options)

This window allows additional options to be set when necessary. For example, you might need to specify a format or delimiter for text files or a table for Access databases. These options will be discussed in detail below.

Step 3 – Run the Import Procedure

Finally, press the Finish button to run the analysis. The data is imported into the first empty datasheet of the current database. If all datasheets currently have data, another datasheet is added to the database to accept the imported data.

Once you have imported a data file, you will probably want to save it as an **NCSS** database. Remember, the file exists only in your computer's memory until it is saved. If you want to avoid importing the file over and over, save it after you have imported it.

Importing Fixed Format ASCII Files

NCSS can import data (both text and numeric) from two types of ASCII files (ASCII files are text files). This section discusses importing files that have a fixed format. The next section will cover importing free format (delimited) ASCII files.

A fixed format ASCII file is one in which each variable occurs at the same position on a row. The data may appear as a solid string of numbers. A format statement is needed to tell the program how to break the data apart into variables.

The format statement is entered in the Format box found on the Step 2 Specify the Fixed Format window.

Specify the Fixed Format

The fixed format syntax is based on three single-letter commands and the slash character. These commands are combined to form the format statement. These commands will be discussed next, followed by examples of format statements.

Fixed Format Syntax

C is for Variable

The character C is used to designate a variable. The actual syntax is rCn. The r indicates the number of times the format segment is repeated. The n represents the number of positions (characters) that are used. If r is omitted, it is assumed to be one. Following are some examples of this type of format.

Format	Meaning
C1	The variable is the next single character on the row.
C3	The variable is the next three characters on the row.
2C4	Two variables, each four characters long.
3C1,2C2	Three variables that are each one character long followed by two variables that are each two characters long.

X is for Skip

The character X is used to designate the skipping of certain character positions on the row. The actual syntax is Xn. The n represents the number of positions (characters) that are skipped. Following are some examples of this type of format.

Format	Meaning
X1	Skip the next character position along the row.
X2	Skip the next two character positions along the row.
X25	Skip the next twenty-five character positions along the row.
X2,X8	Skip two and then eight character positions. Of course, you would usually write X10 instead.

T is for Transfer

The character T is used to transfer to a specific character position on the current row. This character position becomes the next position processed by the format decoder.

If your format includes multiple rows, you should remember that you cannot use the T command to move back to a previous row or ahead to the next row.

Format	Meaning
T1	Transfer to the first position.
T22	Transfer to position twenty-two.

/ is for Next Row

The character / is used to transfer to the beginning of the next row.

Examples of Fixed Format Statements

The above format commands, except for the slash, are put together using commas. The slash serves as its own separator and does not need to be combined with a comma. Note that the values are assigned to the datasheet variables in sequence. Following are some examples of how these format commands can be placed together to form the format statement.

File Segment	Format Statement	Interpreted Values
12345 7890	10C1	1, 2, 3, 4, 5, missing value, 7, 8, 9, 10
12 4567890	5C2	12, 4, 56, 78, 90
1234567890	C2,X4,C3	12, 789
1234567890	C2,T7,C3	12, 789
1234567890	C3,T1,3C1	123, 1, 2, 3,
1234567890	C2,C3,C5	12, 345, 67890
1234567890	(combined with the next line)	
2345678901	2C5/C3	12345, 67890, 234

Importing Delimited ASCII Files

NCSS can import data from two types of ASCII files (ASCII files are text files). This section discusses importing files that have a free, or delimited, format. The previous section covered importing fixed format ASCII files.

The delimiter statement is entered in the Delimiter(s) box found on the Step 2 Specify the Delimiter(s) window.

To use this format, simply specify the character(s) that are to be used to separate the values on the file. For example, if your data consists of values separated by spaces, you would enter “blank” as the delimiter. This is all you need to do.

Some of the common delimiters do not display, so you can type in their names. This is the case for “comma,” “blank,” and “tab.” Other than these, you can indicate the delimiter character directly.

A few other comments may be helpful. Delimiters may only be one character long. Multiple spaces are treated as a single blank. Text with imbedded spaces, such as a street address, should be enclosed within double quotation marks.

Importing SAS or Access Files

Each Microsoft Access database contains several tables. Each table may be thought of as an independent database. When you are importing data from an Access database, you will have to designate a table. Each table has its own set of variables.

Only one Access table may be imported at a time. If an Access database has several tables, you can import them one at a time. Each will go into a separate datasheet in the current database.

Some SAS files also include multiple tables. You can only import one table at a time from these files as well.

Chapter 116

Exporting Data

Introduction

Both text and numeric data may be exported to a wide variety of spreadsheets, databases, and statistical systems. Following is a list of the file types that can be exported. Other file types and more current versions are being added. If the file type you need to create is not on this list, check the online help. It may have been added since this manual was printed.

List of Exported Files

File Name	File Type	Versions	File Extension
Access	Database	2.0	MDB
Alpha Four	Database		DBF
ASCII Delimited	Text File		TXT
ASCII Fixed Format	Text File		TXT
BMDP	Stat System	Classic	POR
Clipper	Database		DBF
Crunch	Stat System		CSC
dBase	Database	II, III, III+, IV	DBF
DIF	Spreadsheet		DIF
Excel	Spreadsheet	(all)	XLS
Gauss	Stat System		DAT
Lotus 123	Spreadsheet	2.1, 3.x, 4.0	WKS, WK1,WK4
NCSS	Stat System	5.0 (DOS)	LAB, DAT
Paradox	Database		DB
Quattro	Spreadsheet		WKQ
SAS	Stat System	Native and Transport	SSD and TPT
Solo	Stat System	4.0 (DOS)	LAB, DAT
SPlus	Stat System		DAT
SPSS	Stat System	Native and Transport	SAV and POR
Stata	Stat System		DTA
Symphony	Spreadsheet	1.0, 1.1, 2.0, 2.1,2.2	WRK, WR1
Systat	Stat System		SYS

Export Limitations

The first limitation in exporting files is that some file types can only receive a limited number of rows and columns. For example, most spreadsheet formats are limited to 256 columns and less than 16,384 rows. Also, formatting and transformations are not exported.

Text or Numeric Field Type

The database (and some spreadsheet) formats require that each field (variable) has a designated type such as numeric or text. Since **NCSS** variables can be either text, numeric, or both, some assumptions have to be made during the export. **NCSS** checks the first row of data for each variable that is exported. If the value in the first row is numeric, then the field is designated numeric. If the first value is text, the field is designated text. Once a database field (variable) has been designated as numeric, any non-numeric values are changed to missing values during the export. Once a field has been designated as text, any numeric values are changed to text values.

How to Export a File

With a few exceptions (that will be discussed later), the steps you need to take are as follows.

Step 0 – Display the Export a File Window

Load the Export a File window by selecting Export from the File menu of the NCSS Data window.

Step 1 – Select Variables to be Exported

First, select the variables from the currently open **NCSS** datasheet that you want to export. Many of the file types have a limitation of 256 variables, so often this will be the maximum number of variables that you can export in a single run.

Variables Exported

Specify the variables to be exported by clicking on the small down-arrow button on the right or by double clicking the box. Either action will bring up the Select Variables to Export window. Once you have selected those variables you wish to export, press the Ok button. The selected variables will be listed in the Variables Exported box.

Rows Exported

Normally, you will leave this box set to All. However, occasionally you may want to run a test of a large export operation. This box lets you limit the number of rows that are exported.

Next button

Press the Next button to proceed to step 2.

Step 2 – Specify the File to be Created

This step specifies the type and name of the new file.

Type of File

Select the type of file that you want to create from this drop-down list of available file types. Note the each file type uses a certain three-letter extension at the end of the file name. For example, Access files use the extension MDB.

Current File Name

This is the name of the currently specified export file. This name may only be changed by clicking the Specify File Name button. You cannot edit the name directly.

Specify File Name

This button activates the Select a File to Open for Exporting window. This window allows you to specify the file name and the directory into which it will be placed. Note that the file name extension such not be changed! Once you have specified the file name, press the Save button to return to the Step 2 window.

Next button

Press the Next button to proceed to step 3.

Step 3 – Specify Parameters

Some file types require additional information. This is the case for ASCII files (covered later in this chapter). Specify any additional information on this window.

Step 4 – Run the Export Procedure

Finally, press the Finish button to export the data to the indicated file.

Exporting Fixed Format ASCII Files

Data may be exported to two types of ASCII files (ASCII files are text files). This section discusses exporting files that are to have a fixed format.

A fixed format ASCII file is one in which each variable occurs at the same position on a row. The data may appear as a solid string of numbers. A format statement is needed to tell the program how to list the data on a row of the file. Besides the format statement, you must also designate what indicator should be used for missing values and what the total row length of the file should be.

Specify Parameters

Format Statement

This box contains the format statement that specifies how the data will be positioned in the output file.

The fixed format syntax is based on three single-letter commands and the slash character. These items are combined to form the format statement. These commands will be discussed next, followed by examples of format statements.

- **L is for a Left-Justified Variable**

  The character L is used to designate a left-justified variable. The actual syntax is rLn.d, where r indicates the number of times the format segment is repeated, n represents the number of positions (characters) used, and d represents the number of decimal places. If r is omitted, it is assumed to be one. Note the n must include space for the decimal point (which is one character) and a minus sign (if necessary).

- **R is for a Right-Justified Variable**

  The character R is used to designate a right-justified variable. The actual syntax is rRn.d, where r indicates the number of times the format segment is repeated, n represents the number of positions (characters) used, and d represents the number of decimal places. If r is omitted, it is assumed to be one. Note the n must include space for the decimal point (which is one character) and a minus sign (if necessary).

- **X is for Skip**

  The character X is used to designate the skipping of certain character positions on the row. The actual syntax is Xn. The n represents the number of positions (characters) that are left blank.

- **/ is for Next Row**

  The character / is used to transfer to the beginning of the next row.

Examples of Fixed Format Statements

The above format commands, except for the slash, are combined using commas. The slash serves as its own separator and does not need to be combined with a comma. Note that the values are assigned to the datasheet variables in sequence. Following are some examples of how these format commands can be placed together to form the format statement.

Variable Values	Format Statement	Resulting Row
12, 45, 32	3L6.1	12.0 45.0 32.0
12, 45, 32	3L6.0	12 45 32
12.23, 45.56, 32.14	2R8.1,X10,R8.3	12.2 45.6 32.140

Missing Value Indicator

This box specifies the character(s) to be inserted for missing values.

Characters Per Line

This box specifies the number of characters per line . If ‘Variable’ is selected, each line will be long enough to hold its data, but no longer.

Export Variable Names

Checking this option causes the variable names to be stored in the first row.

Exporting Delimited ASCII Files

This section discusses exporting files that have a free (delimited) format.

Specify Parameters

Delimiter

You specify the character that is to be used to separate the values on the file. For example, if you want to separate the values with spaces (blanks), you would select “space” as the delimiter.

Some of the common delimiters won’t display, so you can type in their names. This is the case for “comma,” “space,” and “tab.” Other than these, you enter the delimiter character directly.

Missing Indicator

This box specifies the character(s) to be inserted for missing values. We suggest that you do not use space for both the missing value indicator and the delimiter.

Enclose Text Between

Text may be enclosed between double or single quotation marks so that text variables with imbedded spaces, such as street addresses, can be processed correctly by other programs. This option specifies how text data should be treated.

Characters Per Line

This box specifies the number of characters per line. If ‘Variable’ is selected, each line will be long enough to hold its data, but no longer.

Export Variable Names

Checking this option causes the variable names to be stored in the first row.

Chapter 117

Data Report

Introduction

This procedure generates a report of the data on a database. It is used when you want to maintain a printed copy of your data.

Data Structure

The procedure prints rows of selected variables. The rows printed may be selected using a filter.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables displayed on the report.

Data Variables

Data Variables

Select at least one variable to be printed. Both numeric and text data may be printed.

Report Options

Decimal Places

This option specifies the number of decimal places displayed for each variable. The number of decimal places is entered as a list of items separated by blanks or commas. For example, suppose you have selected variables X1, X2, and X3 for printing. If you enter “1,2,0” for this option, X1 will be printed with one decimal place, X2 with two decimal places, and X3 with no decimal places.

The number of decimal places can range from 0 to 9. In addition to this, you can enter one of three special formatting codes: **S**, **D**, and **F**.

S is used to indicate that numbers should be displayed in *single* precision.

D is used to indicate that numbers should be displayed in *double* precision.

F is used to indicate that numbers should be displayed using the *format* that is specified for the variable in the Variable Info sheet of the database. This allows you to specify commas, date conversions, etc.

For example, suppose you entered “F,S,2” here. X1 would be displayed using its format as specified on the Variable Info sheet, X2 would be displayed as a single precision number, and X3 would be displayed to two decimal places.

Note that if this statement is too short to account for all the variables, it is repeated. Hence, if you enter a “1” here and have three variables to display, all three will show a single decimal place.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want the table to automatically attach labels to the values (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Precision

Specify the precision of numbers in the report. This is used when the format statement is left blank.

Label Justification

This option specifies whether the column labels should be right or left justified.

Data Justification

This option specifies whether the data should be right, left, or decimal justified.

Split Column Headings

Check this option to split the column headings into two headings instead of one.

Double Space

Check this option to add a blank row after each row.

Tabs

First

Specifies the position of the first item in inches. Note that the left-hand label always begins at 0.5 inches. Hence, the distance between this tab and 0.5 is the width provided for the row information.

Maximum

Specifies the right border of the report. The number of tabs is determined based on the First Tab, the Tab Increment, and this option. If you set this value too large, your table may not be printed correctly.

Increment

Specifies the width of an item in inches.

Offset

The labels are left justified. The data in the report are decimal tabbed (centered at the decimal place). Using two tabbing styles will cause the labels to be out of alignment with the data. Each data tab is moved to the right by this amount so that the data will line up with the column labels.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a List of Variables from a Database

This section presents an example of how to create a list of variables from a database. Data from the RESALE database will be used to generate the sample report.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Data Report window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **RESALE.s0**.
- Click **Open**.

2 Open the Data Report window.

- On the menus of the NCSS Data window, select **Data**, then **Data Report**. The Data Report procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Report window, select the **Variables tab**.
- Double-click in the **Data Variables** box. This will bring up the variable selection window.
- Select **State, City, Price, Year, Bedrooms,** and **Bathrooms** from the list of variables and then click **Ok**.
- Enter **0 0 0 0 0 0** in the **Decimal Places** box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data List Section

Data List Section

Row	State	City	Price	Year	Bedrooms	Bathrooms
1	Nev	2	260000	1972	2	3
2	Nev	2	66900	1942	3	3
3	Vir	4	127900	1975	2	2
4	Nev	1	181900	1984	3	3
5	Nev	2	262100	1970	2	3
6	Nev	1	147500	1986	2	3
7	Nev	2	167200	1987	2	2
8	Nev	1	395700	1991	2	2
9	Vir	5	106600	1976	3	4
10	Nev	3	78700	1963	2	2

(report continues)

This report lists the data in the selected variables.

Chapter 118

Data Screening

Introduction

This procedure performs a screening of data in a database, reporting on the:

1. Type of data (discrete or continuous)
2. Normality of each variable
3. Missing-value patterns
4. Presence of outliers

When you have missing values in your database, this program estimates the missing values using either a simple average or a more elaborate multiple regression technique.

Data screening should be carried out prior to any statistical procedure. Often data screening procedures are so tedious that they are skipped. Then, after an analysis produces unanticipated results, the data are scrutinized. This program automates the whole data screening process. When used in conjunction with histograms and scatter plots, you will be able to verify most of your data assumptions before beginning the actual analysis.

Data Structure

The data are entered in one or more variables. Only numeric values are allowed. Missing values are represented by blanks. Text values are treated as missing values.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables to be analyzed.

Data Variables

Variables to Screen

Specify the variables to be screened. Only numeric values are analyzed.

Options

Max Discrete Levels

The maximum number of unique values that a variable can have and still be designated as *discrete* rather than *continuous*.

Missing Value Estimation

Specify the type of missing value estimation (imputation) if any.

- **None**

  No missing value estimation is carried out.

- **Average**

  Estimate the missing value using the average of the variable.

  Although this method is fast and simple, it does have disadvantages. First, the variance of the variable will be understated since adding values near the overall mean has little impact on the variance. Second, correlations with other variables may be incorrect since they involve the variances of the two variables.

- **Multivariate Normal**

  Estimate missing values using the multivariate normal procedure. This method takes extra time but results in much more reliability estimates.

  A regression analysis is conducted using the variable containing the missing value as the dependent variable and all variables with nonmissing data in this row as independent variables. The values of these nonmissing variables from the row containing the missing value are used in the regression equation to compute a predicted value for the missing value. Finally, if you are estimating a discrete value, the predicted value is rounded to the nearest possible discrete value. This process is iterated by using the imputed missing values from one run during the estimation phase of the next.

  This procedure provides reasonable estimates of the missing values. It does have a few disadvantages. First, it assumes a multivariate normal distribution which may not be accurate. Second, it tends to provide estimates that understate the size of the variance of the variable. Third, it relies on the correlations between the variable with the missing value and the other variables in the database. If these correlations are all small, the resulting regression equation may not be very reliable.

Number of Iterations

This option specifies the number of iterations used during the estimation of missing values. Usually, only three or four iterations are necessary.

Zero

Specify the value used as zero by the numerical routines. Because of round-off problems, values less than this amount (in absolute value) are changed to zero during the calculations.

Treatment of Blanks at the End

This option specifies how to treat blanks at the end of each variable. Since the number of observations in each variable (column) can vary, an option is needed to control how the blanks at the end of each column are treated.

Let *MAXN* represent the row number of the largest, non-blank row in the database. Let *Ni* represent the largest, non-blank row number in column *i*. Two options are available:

- **Spreadsheet**
  With this option, blanks at the end (bottom) of each variable are ignored. That is, the *Ni* of each variable is determined separately.

- **Database**
  With this option, blanks at the end (bottom) of each variable are included. That is, each *Ni* is set equal to *MAXN*.

Store Estimated Values

Checking this option will cause missing values in the database to be replaced by their estimates. Remember, these new values are not stored permanently until you manually save the database.

Reports Tab

The following options control the format of the reports that are displayed.

Select Reports

Descriptive Statistics ... Iteration Report

Indicate whether to display the indicated reports.

Report Options

T2 Alpha

This is the probability value used to identify outliers. Observations with a T2 probability less than this are designated as outliers.

Normality Test Alpha

This is the probability value used to identify variables that are not normally distributed. Variables with a normal test probability less than this are designated as being not normal.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Screening Data

This section presents an example of how to screen the data in the PCA2 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Data Screening window.

1 Open the PCA2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **PCA2.s0**.
- Click **Open**.

2 Open the Data Screening window.

- On the menus of the NCSS Data window, select **Descriptive Statistics**, then **Data Screening**. The Data Screening procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Screening window, select the **Variables tab**.
- Double-click in the **Variables to Screen** box. This will bring up the variable selection window.
- Select **X1** to **Normal** from the list of variables and then click **Ok**.
- Enter **Multivariate Normal** in the **Missing Value Estimation** box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Descriptive Statistics Section

Descriptive Statistics Section

Data Type	Variable	Value Count	Missing Count	Minimum	Maximum	Mean	Standard Deviation
Continuous	X1	30	0	3	85	44.2	24.66241
Continuous	X2	30	0	1	102	51.53333	30.57803
Continuous	X3	30	0	2	105	54.93333	29.05753
Continuous	X4	30	0	5	91	41.7	25.3175
Continuous	X5	30	0	6	91	43.66667	26.65143
Continuous	X6	30	0	0	102	47.63334	34.18962
Discrete	Q1	30	0	1	5	3.466667	1.407696
Discrete	Q2	30	0	1	5	3.033333	1.098065
Discrete	Q3	29	1	1	5	2.827586	1.559967
Discrete	Q4	30	0	1	5	2.833333	1.620628
Discrete	Q5	27	3	1	5	2.62963	1.445102
Discrete	Q6	29	1	1	5	3.241379	1.50369
Discrete	Q7	30	0	1	5	2.966667	1.188547
Discrete	Q8	28	2	1	5	2.785714	1.397276
Discrete	Q9	30	0	1	5	2.9	1.470398
Continuous	Normal	30	0	79.98992	119.8588	100.6198	10.17271

This report gives descriptive statistics and counts for each variable. Note that using the missing value imputation option will not influence the values in this report. Most of these statistics have been defined in the Descriptive Statistics chapter.

Data Type

The type of data contained in each variable. If the number of unique values is less than the cutoff value given in the Max Discrete Levels option, the variable will be categorized as *Discrete*. Otherwise, it is categorized as *Continuous*. It is important to know the data type of each variable early in an analysis.

Value Count

This is the number of rows for which there were valid numeric values.

Missing Count

This is the number of rows for which there were missing values.

Normality Tests Section

Normality Tests Section

	------ Skewness Test -------			-------- Kurtosis Test --------			- Omnibus Test -		Variable
Variable	Value	Z	Prob	Value	Z	Prob	K2	Prob	Normal?
X1	0.11	0.30	0.7662	1.93	-1.80	0.0715	3.34	0.1885	Yes
X2	0.11	0.30	0.7664	1.88	-2.01	0.0446	4.12	0.1274	No
X3	-0.02	-0.06	0.9552	2.14	-1.14	0.2534	1.31	0.5201	Yes
X4	0.41	1.05	0.2918	2.31	-0.72	0.4712	1.63	0.4425	Yes
X5	0.39	0.99	0.3221	2.02	-1.50	0.1327	3.24	0.1978	Yes
X6	0.29	0.75	0.4535	1.63	-3.18	0.0015	10.65	0.0049	No
Q1	-0.50	-1.26	0.2072	1.97	-1.65	0.0980	4.33	0.1148	Yes
Q2	0.41	1.05	0.2925	2.27	-0.79	0.4269	1.74	0.4191	Yes
Q3	0.12	0.31	0.7595	1.53	-3.67	0.0002	13.57	0.0011	No
Q4	0.17	0.46	0.6483	1.52	-3.85	0.0001	15.03	0.0005	No
Q5	0.36	0.89	0.3732	1.81	-2.09	0.0368	5.15	0.0761	No
Q6	-0.23	-0.58	0.5597	1.59	-3.33	0.0009	11.41	0.0033	No
Q7	-0.31	-0.80	0.4208	2.19	-1.00	0.3186	1.64	0.4398	Yes
Q8	0.22	0.57	0.5713	1.75	-2.43	0.0153	6.20	0.0450	No
Q9	0.11	0.28	0.7765	1.67	-2.93	0.0034	8.64	0.0133	No
Normal	-0.28	-0.72	0.4715	2.44	-0.43	0.6701	0.70	0.7047	Yes

This report shows the results of the three D'Agostino normality tests. These tests are described in detail in the Descriptive Statistics chapter. If any of the three probability values are less than the user supplied Normality Test Alpha, the variable is designated as *not normal* (Variable Normal = No). Otherwise, the variable is designated as *normal* (Variable Normal = Yes).

We should remind you that the results of these tests depends heavily on sample size. If you have a small sample size (less than 50), these tests may fail to reject normality because the sample size is too small--not because the data are actually normal. Likewise, if your sample size is very large (greater than 1000), these tests may reject normality even though the data are nearly normal. When in doubt, you should supplement these tests with additional tests and graphs.

Pair-wise Missing Data Counts Section

Pair-wise Missing Data Counts Section

	X1	X2	X3	X4	X5	X6
X1	0	0	0	0	0	0
X2	0	0	0	0	0	0
X3	0	0	0	0	0	0
X4	0	0	0	0	0	0
X5	0	0	0	0	0	0
X6	0	0	0	0	0	0
Q1	0	0	0	0	0	0
Q2	0	0	0	0	0	0
Q3	1	1	1	1	1	1
Q4	0	0	0	0	0	0
Q5	3	3	3	3	3	3
Q6	1	1	1	1	1	1
Q7	0	0	0	0	0	0
Q8	2	2	2	2	2	2
Q9	0	0	0	0	0	0
Normal	0	0	0	0	0	0

(report continues)

This report provides a pair-wise break down of the number of rows with missing values in at least one of each pair of variables. This is the number of observations that would be omitted from the calculation of the correlation coefficient between these two variables.

An understanding of the distribution of missing values is extremely important when conducting an analysis that is based on correlations such as factor analysis or multiple regression. You may determine that much more data would be used if you omit two or three variables that have high counts of missing values.

Pair-wise Missing Data Percentages Section

Pair-wise Missing Data Percentages Section

	X1	X2	X3	X4	X5	X6
X1	0.0	0.0	0.0	0.0	0.0	0.0
X2	0.0	0.0	0.0	0.0	0.0	0.0
X3	0.0	0.0	0.0	0.0	0.0	0.0
X4	0.0	0.0	0.0	0.0	0.0	0.0
X5	0.0	0.0	0.0	0.0	0.0	0.0
X6	0.0	0.0	0.0	0.0	0.0	0.0
Q1	0.0	0.0	0.0	0.0	0.0	0.0
Q2	0.0	0.0	0.0	0.0	0.0	0.0
Q3	3.3	3.3	3.3	3.3	3.3	3.3
Q4	0.0	0.0	0.0	0.0	0.0	0.0
Q5	10.0	10.0	10.0	10.0	10.0	10.0
Q6	3.3	3.3	3.3	3.3	3.3	3.3
Q7	0.0	0.0	0.0	0.0	0.0	0.0
Q8	6.7	6.7	6.7	6.7	6.7	6.7
Q9	0.0	0.0	0.0	0.0	0.0	0.0
Normal	0.0	0.0	0.0	0.0	0.0	0.0

(report continues)

This report provides a pair-wise break down of the percentage of rows with missing values in at least one of each pair of variables. This is the percentage of observations that would be omitted from the calculation of the correlation coefficient between these two variables.

An understanding of the distribution of missing values is extremely important when conducting an analysis that is based on correlations such as factor analysis or multiple regression. You may determine that much more data would be used if you omit two or three variables that have high counts of missing values.

List of Discrete Variables and Values Section

List of Discrete Variables and Values Section

Variable	Value1(Count1) Value2(Count2) etc.
Q1	1(4) 2(4) 3(5) 4(8) 5(9)
Q2	1(1) 2(10) 3(10) 4(5) 5(4)
Q3	1(9) 2(4) 3(5) 4(5) 5(6)
Q4	1(10) 2(3) 3(7) 4(2) 5(8)
Q5	1(8) 2(6) 3(5) 4(4) 5(4)
Q6	1(5) 2(6) 3(3) 4(7) 5(8)
Q7	1(5) 2(4) 3(10) 4(9) 5(2)
Q8	1(6) 2(8) 3(4) 4(6) 5(4)
Q9	1(7) 2(6) 3(6) 4(5) 5(6)

This report lists each of the discrete variables (as defined by Max Discrete Levels) followed by a list of the discrete values and corresponding counts of those values. For example, the first entry of 1(4) means that four 1's occurred in this variable.

This report is particular useful in helping you find out-of-range values in discrete data.

Multivariate Outlier Section

Multivariate Outlier Section

Row	T2 Value	T2 Prob	Outlier?
1			
2	27.97	0.6310	
3	28.03	0.6295	
4	11.99	0.9729	
5			
6	11.30	0.9790	
7	20.40	0.8250	
8			
9	11.86	0.9741	
10	13.64	0.9544	
.	.	.	
.	.	.	
.	.	.	

This report tests each observation to determine if it is a multivariate outlier. The program uses a T^2 test based on the Mahalanobis distance of each point from the variable means. The formula for T^2 is:

$$T_i^2 = (n-1)(X_i - \bar{X})' \left[(X - \bar{X})' (X - \bar{X}) \right]^{-1} (X_i - \bar{X})$$

The following mathematical relationship between the T^2 and the F-distribution is used to calculate the probability levels:

$$T^2_{p,n,\alpha} = \frac{p(n-1)}{n-p} F_{p,n-p,\alpha}$$

Note that as the number of variables, *p*, approaches the sample size, n, the denominator degrees of freedom approaches zero. As *n-p* approaches zero, the power of the test also approaches zero.

This test is only calculated for rows that have no missing values. To test rows with missing values, you will need to store imputed values on the database and rerun the analysis.

When the probability level is less than the value indicated in the T2 Alpha box, the observation is starred.

Rows With Missing Values Section

Rows With Missing Values Section

Row	Pattern of Missing Values (| = data, . = missing)
1	|.||
5	.|||
8	|||.
12	|||.
14	||.|
16	|.||
19	|.||

Variables With Missing Values

Q3
Q5
Q6
Q8

This report presents a list of only those variables and rows that had missing values. It lets you consider the pattern of missing values more closely.

For each row, missing values are represented by a period and valid values are represented by a vertical bar. These symbols were selected because the have about the same width in most fonts.

Missing Values Estimation Iteration Section

Missing Values Estimation Iteration Section

Iteration No.	Count	Covariance Matrix Trace	Percent Change
0	23	4815.0995	0.00
1	30	5029.5552	4.45
2	30	5029.4391	0.00
3	30	5029.3897	0.00

This report shows the percent change in the trace of the variance-covariance matrix as you progress from one iteration to the next during the estimation of missing values. You would use the report to determine if enough iterations have been run during the estimation of missing values. Once the percent change is less than four percent after the first two iterations, you could terminate the procedure. If the last two iterations show very different values, you should rerun the analysis with a higher number of iterations.

Chapter 119

Transformations

Introduction

Transformations generate new data. They can be used to perform mathematical operations such as addition, multiplication, and exponentiation on variables. For example, you might want to create a new variable that is the logarithm of a variable, the total of three variables, or the recoding of a variable. A rich set of text transformations is also available.

A transformation formula specifies how the new variable is constructed from the existing variables. The formula is made up of mathematical operators, variables, constants, and common mathematical functions.

A transformation is stored in the appropriate position of the Variable Info datasheet of the database. You can type the formula directly, or you can enter it using the Transformation Window that will be discussed shortly.

It is important to understand that these transformations modify variables (columns) of data, not individual cells. This is one place where the NCSS spreadsheet is different from traditional spreadsheets.

Example

The following example will fill the third variable on the database with the result of adding variables C1 and C2 together. Note that entering a transformation on the database causes no immediate calculations. You must select Recalc Current or Recalc All from the Data menu or toolbar to cause the transformations to take effect.

Step 1 Move to the Variable Info datasheet by clicking on the Variable Info tab at the bottom of the database screen.

Step 2 Move to the third column, which is labeled Transformation.

Step 3 Move down to the third row and type **C1+C2**. Note that you could have double-clicked on this cell to bring up the special Transformation Window.

Step 4 Create the transformed values by selecting **Recalc All** from the Data menu or clicking the calculator button on the toolbar.

Transformation Window

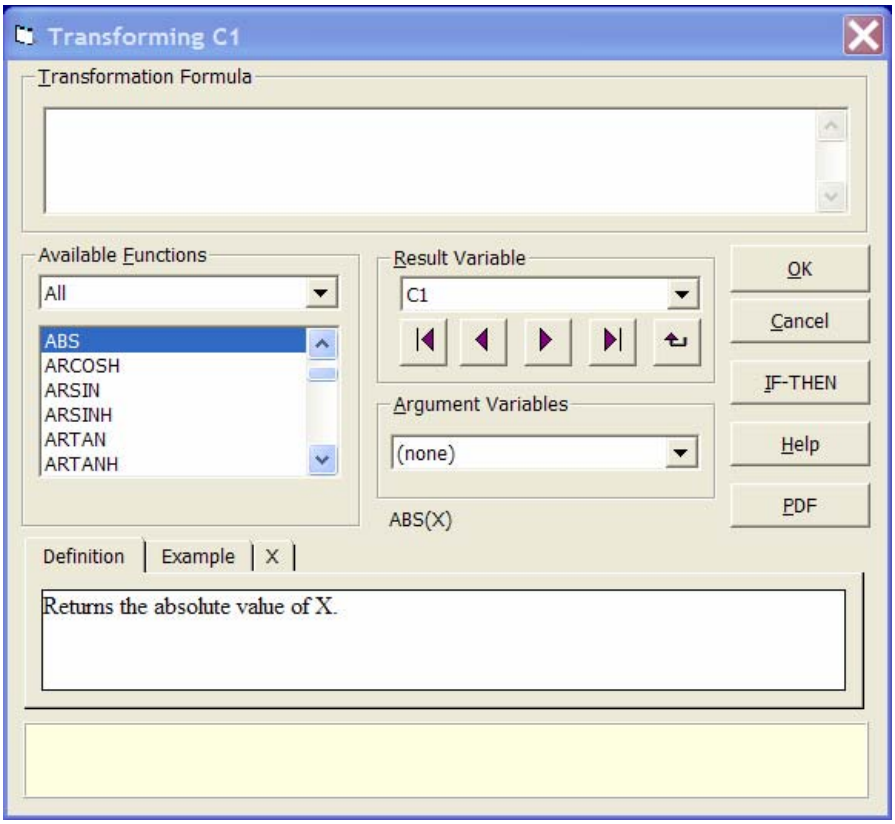
The Transformation Window may be used to facilitate the entry of a transformation formula. You can think of this window as a special viewer of the Transformation column of the Variable Info sheet. As you change the Result Variable in this window, you are actually moving up and down the Transformation column of the Variable Info sheet.

Remember, however, that you do not have to use the Transformation Window to enter a transformation. You may enter the formula directly into the appropriate cell of the Variable Info sheet.

The Transformation Window may be activated in several ways:

1. Select the menu sequence: **Data**, then **Enter Transform** from the spreadsheet window.
2. Select the menu sequence: **Edit**, then **Variable Info**, then **Transformation** from the spreadsheet window.
3. Select the **Enter Transformation** icon from the Toolbar of the spreadsheet window.
4. Right-click on a cell. From the pop-up menu select **Variable Info**, then **Transformation**.
5. Double-click on a transformation cell in the Variable Info sheet.
6. Choose **Transformation** from the Data node of the NCSS Navigator window.

When you take one of these actions, the following window will appear.



Transformation Formula

Enter a formula here, either by typing or making selections. The syntax of the formula is discussed later.

Available Functions

Select a function type from the drop-down list on the top field. Select a particular function from the list on the bottom field. Double-clicking on an item will cause it to be added to the current formula at the insertion point.

Result Variable

Specify the current Result Variable here. The transformation will be associated with this variable. Selecting another Result Variable will cause the current transformation to be saved before moving to the selected variable. You can use the Toolbar to navigate quickly among the possible Result Variables.

Argument Variables

You can select argument variables from this list or just type them in directly. Double-clicking on a variable will cause it to be added to the current formula at the insertion point.

Ok

Store the current transformation and close the Transformation Window.

Cancel

Close the Transformation Window without saving the current transformation.

IF-THEN

Open the IF-THEN transformation Window.

Help

Loads the Help file.

PDF

Load a printable copy of the documentation for this module.

Definition, Example, Etc.

This section provides on-line help about the currently selected function. Click on the tabs to see help about each of the arguments in the function.

Transformation Syntax

Result Variable

The results of a transformation are stored in the designated Result Variable. In the example below, a transformation has been used to create PercentA and PercentB from the original variables YieldA and YieldB.

Recalculating

Entering a transformation on the database causes no immediate calculations. Instead, select Recalc Current or Recalc All from the Data Menu to cause the transformations to take effect.

Note that when a transformed variable is recalculated, existing data (including cell formulas) are replaced by the new values. Hence, you should never enter data directly into a variable that is associated with a transformation because upon recalculation, these data will be replaced.

Recalc Current instructs the program to recalculate the variable on which the cursor is positioned. This is convenient for testing a single transformation in a database with several transformations.

Recalc All instructs the program to recalculate all transformations. Note that the program moves across the database from left to right, running each transformation as it is encountered.

Number of Rows

The number of rows processed by the transformations is controlled by an option displayed above the spreadsheet in the toolbar directly to the right of the calculator icon. The default value of this option is zero. This option controls the number of rows that are processed. A zero indicates that the largest row containing data is used to determine the number of transformed rows. When you set this value to an integer greater than zero, this is the number of rows that will be processed regardless of the number of rows with data in other variables.

Arguments

Variable Transformations are used to formulate new variables by manipulating existing data. This section describes the basic syntax of a transformation formula. The syntax of a transformation formula is the sequence of symbols, numbers, and characters in the formula.

The arguments of a function are the individual entities that make up the formula. For example, in the formula

3.2 + 54 / 22

the numbers 3.2, 54, and 22 are the arguments. Arguments may be numbers, text, variables, or another formula contained within parentheses. Each argument must produce a valid value or an error or missing value will result. When one formula is contained within another formula, it is said to be nested. You can nest as many functions within each other as you like.

Following are a few examples of transformation formulas:

C1+C2+C3

C1*4+2

LOG(C2)

100*C1/(SIN(C2+4))

Operators

Formulas are made by combining arguments using operators. Common operators are addition, subtraction, multiplication, and division. A complete list of the operators available is given below.

Note that no two operators can appear consecutively. A formula like **C1+/C2** is illegal. However, you can enter **C4*-5** because **-5** is treated as a constant. The operators and parentheses serve as delimiters for the arguments of a function.

The standard order of calculation is used. This means that embedded functions are calculated first, then exponentiation, multiplication, division, addition, and subtraction. Hence, the formula **4+2*6** would yield **16**, not **36**!

Addition and Subtraction Operators

+ This is the addition symbol. In the formula **X + Y**, it signifies adding **X** to **Y**. Multiplications and divisions are performed before additions and subtractions. If either **X** or **Y** is missing, the result is missing.

- This is the subtraction symbol. In the formula **Y - X**, it signifies subtracting **X** from **Y**. If either **X** or **Y** is missing, the result is missing.

Multiplication and Division Operators

***** This is the multiplication symbol. In the formula **X * Y**, it signifies the product of **X** and **Y**. If either **X** or **Y** is missing, the result is missing.

/ This is the division symbol. In the formula **Y / X**, it signifies dividing **Y** by **X**. If either **X** or **Y** is missing, the result is missing. Also note that **X** cannot be equal to zero (in which case a missing value results).

Exponentiation Operator

^ This is the exponentiation symbol. In the formula **X ^ Y**, it signifies raising **X** to the power **Y**. If either **X** or **Y** is missing, the result is missing. Note that the result must be less than 10^308 in absolute value. Also note that **X** must be a nonnegative number.

Examples: 12^2 = 144; 2^3 = 8; 9^(1/2) = 3; 10^(-1) = 0.10.

Logic Operators

< This is the less-than symbol used in a logic expression. If **(X<Y)** is true, a one results. If **(X<Y)** is false, a zero results.

Examples: (4<6) = 1 and (6<4) = 0.

<= This is the less-than or equal-to symbol used in a logic expression. If **(X<=Y)** is true, a one results. If **(X<=Y)** is false, a zero results.

Examples: (4<=6) = 1; (6<=6) = 1; and (6<=4) = 0.

> This is the greater-than symbol used in a logic expression. If **(X>Y)** is true, a one results. If **(X>Y)** is false, a zero results.

Examples: (4>6) = 0 and (6>4) = 1.

>= This is the greater-than or equal-to symbol used in a logic expression. If **(X>=Y)** is true, a one results. If **(X>=Y)** is false, a zero results.

Examples: (4>=6) = 0; (6>=6) = 1; and (6>=4) = 1.

= This is the equal-to symbol used in a logic expression. If **(X=Y)** is true, a one results. If **(X=Y)** is false, a zero results.

Examples: (4=6) = 0 and (6=6) = 1.

<> This is the not-equal-to symbol used in a logic expression. If **(X<>Y)** is true, a one results. If **(X<>Y)** is false, a zero results.

Examples: (4<>6) = 1 and (6<>6) = 0.

Arrangement Operators

( ) The parentheses are used to force the order of precedence in an expression. Expressions inside a set of parentheses are evaluated first. Note that they must be used as a pair; each left parenthesis must have a corresponding right parenthesis or an error will result. For example, 3+4*2 = 11; while (3+4)*2 = 14.

{ } The braces are used interchangeably with the parentheses. You would usually use them to avoid confusion when multiple sets of parentheses are called for.

[ ] The brackets are used interchangeably with the parentheses. You would usually use them to avoid confusion when multiple sets of parentheses are called for.

Numeric Functions

Numeric functions have a name and one or more arguments contained within parentheses. Inside an expression, a function is treated just like a number or a variable. The following functions may be used in combination with other expressions. You must be careful to include the correct number of arguments and to make sure that the values of the arguments are within appropriate limits. For example, you cannot have **SQRT(-5)**.

Date Functions

Day (Julian)

Returns the number of the day in a Julian date. **Julian** is a Julian date.

Julian (Month,Day,Year)

Returns a Julian date. **Month** is the number of the month (1 to 12). **Day** is the number of the day (1 to 31). **Year** is the number of the year (0 to 2050).

Month (Julian)

Returns the number of the month in a Julian date. **Julian** is a Julian date.

Year (Julian)

Returns the number of the year in a Julian date. **Julian** is a Julian date.

Fill Functions

Sequence (Inc)

Returns a series beginning at **Inc** that adds **Inc** at each subsequent row. **Inc** is the increment added at each row to the amount of the last row.

Example: Sequence(5) yields 5,10,15,20,25,...

Ser (Length,Restart)

Returns a series beginning at 1 that adds 1 at each new row. **Length** is the number of rows before the series is increased by one. **Restart** is the number of rows before the series is restarted.

Example: **Ser(2,6)** yields 1,1,2,2,3,3,1,1,2,2,3,3,...

Mathematical Functions

Abs (X)

Returns the absolute value of **X**.

Example: **ABS(-4)** yields **4**.

Cum (X)

Returns the sum of the values of **X** up to and including the current row.

Exp (X)

Returns the value of e raised to the power **X**. The opposite of log (base e) of **X**.

Fraction (X)

Returns the fractional (noninteger) portion of **X**. **X** is a number.

Example: **Fraction(5.1234)** yields 0.1234

Int (X)

Returns the integer portion (the whole number part) of **X**.

Ln(X)

Returns the logarithm (base e) of **X**. **X** is a number greater than zero.

Log (X)

Returns the log (base 10) of **X**. **X** is a number greater than zero.

LogBase(X,Base)

Returns the logarithm of **X** in base **Base**. **X** is a number > 0. **Base** is the base, a number > 0.

Logit (X)

Returns the logit of **X** which is **Ln(X/(1-X))**. **X** is a number between 0 and 1.

Mod (X,Y)

Returns the remainder after dividing **X** by **Y**. This is sometimes written as **X Mod Y**. **X** is a number. **Y** is a positive integer less than 255.

Example: **Mod(15,4)** yields 3.

Round (X,D)

Returns a value rounded off to the nearest roundoff unit. **X** is the number to be rounded. **D** is the roundoff unit.

Example: **Round(10.432,.1)** yields 10.4.

Short (X)

Returns a single-precision version of **X**.

X is the number to be changed.

Sign (X)

Returns the sign of **X**. If **X** is negative, the result is -1. If **X** is positive, the result is +1. If **X** is zero, the result is zero.

Example: **SIGN(-4.2)** yields -1.

Sqrt (X)

Returns the positive square root of **X**. **X** is a non-negative number.

Probability Functions

BetaProb (X,A,B)

Returns the probability of being less than **X** when **X** follows the Beta distribution. This is the left-tail probability. **X** is a real number between 0 and 1. **A** is a nonnegative number. **B** is a nonnegative number.

BetaValue (Prob,A,B)

Returns the inverse of the Beta distribution. It gives the value, **X**, such that the area to the left is equal to **Prob**. **Prob** is a real number between 0 and 1. **A** is a nonnegative number. **B** is a nonnegative number.

BinomProb (R,N,P)

Returns the probability of being less than or equal to **R** when **R** follows the binomial distribution with sample size, **N**, and prob of success, **P**. This is the left-tail probability. **R** is a non-negative number less than or equal to **N**. **N** is the number of trials (sample size) and must be a positive integer. **P** is probability of a success on a particular trial. It must be between 0 and 1.

BinomValue (Prob,N,P)

Returns the inverse of the binomial distribution. It gives the number, **R**, such that the cumulative binomial probability is less than or equal to **Prob**. **Prob** is a real number between 0 and 1. **N** is the number of trials (sample size) and must be a positive integer. **P** is probability of a success on a particular trial. It must be between 0 and 1.

BinormProb(X,Y,Rho)

Returns the probability of being greater than **X** and greater than **Y** when **X** and **Y** follow the standardized bivariate-normal distribution. **X** is any real number. Usually, |**X**|<3. **Y** is any real number. Usually, |**Y**|<3. **Rho** is the population correlation coefficient between **X** and **Y**. **-1 < Rho < 1**.

CorrProb (R,N,Rho)

Returns the probability of being less than **R** when **R** follows the correlation distribution. This is the left-tail probability. **R** is any real number. **N** is the sample size. It must be a positive number. **Rho** is the population correlation coefficient. **-1 < Rho < 1**.

CorrValue(Prob,N,Rho)

Returns the inverse of the correlation distribution. It gives the value, **R**, such that the area to the left is equal to **Prob**. **Prob** is a probability value. It must be between 0 and 1. **N** is the sample size. It must be a positive number. **Rho** is the population correlation coefficient. **-1 < Rho < 1**.

CsProb (X,Df)

Returns the probability of being less than **X** when **X** follows the Chi-Square distribution. This is the left-tail probability. **X** is a positive real number. **Df** is the degrees of freedom. It must be a positive number.

CsValue(Prob,Df)

Returns the inverse of the Chi-Square distribution. It gives the value, X, such that the area to the left is equal to **Prob**. **Prob** is a probability value. It must be between 0 and 1. **Df** is the degrees of freedom. It must be a positive number.

ExpoProb (T,Scale)

Returns the probability of being less than or equal to **T** when **T** follows the Exponential distribution. This is the left-tail probability. **T** is a positive number. **Scale** is the scale parameter.

ExpoValue (Prob,Scale)

Returns the inverse of the Exponential distribution. It gives the number, **T**, such that the Exponential probability is less than or equal to **Prob**. **Prob** is a real number between 0 and 1. **Scale** is the scale parameter.

Fprob (F,Df1,Df2)

Returns the probability of being less than **F** when **F** follows the F distribution. This is the left-tail probability. **F** is a positive real number. **Df1** is the numerator degrees of freedom. It must be a positive number. **Df2** is the denominator degrees of freedom. It must be a positive number.

Fvalue (Prob,Df1,Df2)

Returns the inverse of the F distribution. It gives the value, **F**, such that the area to the left is equal to **Prob**. **Prob** is a probability value. It must be between 0 and 1. **Df1** is the numerator degrees of freedom. It must be a positive number. **Df2** is the denominator degrees of freedom. It must be a positive number.

GammaProb (X,A)

Returns the probability of being less than **X** when **X** follows the Gamma distribution. This is the left-tail probability. **X** is a non-negative number. **A** is a positive number.

GammaValue (Prob,A)

Returns the inverse of the Gamma distribution. It gives the value, **X,** such that the area to the left is equal to **Prob**. **Prob** is a real number between 0 and 1. **A** is a nonnegative number.

HypergeoProb (X,N,R,M)

Returns the probability of the hypergeometric distribution. This is the left-tail probability, including the current value of **X**. **X** is the number of successes, where **X>=0** and **X<=R**. **N** is the population size, where **N>=1**. **R** is the sample size, where **R>=1** and **R<=N**. **M** is the subgroup size, where **M>=1** and **M<=N.**

LogGamma (X)

Returns the natural logarithm of the gamma function. **X** is a positive number.

NcBetaProb (X,A,B,Ncp)

Returns the probability of being less than **X** when **X** follows the noncentral-Beta distribution. This is the left-tail probability. **X** is a real number between 0 and 1. **A** is a non-negative number. **B** is a non-negative number. **Ncp** is the noncentrality parameter.

NcBetaValue (Prob,A,B,Ncp)

Returns the inverse of the noncentral-Beta distribution. It gives the value, **X**, such that the area to the left is equal to **Prob**. **Prob** is a real number between 0 and 1. **A** is a non-negative number. **B** is a non-negative number. **Ncp** is the noncentrality parameter.

NcCsProb (X,Df,Ncp)

Returns the probability of being less than **X** when **X** follows the noncentral-Chi-Square distribution. This is the left-tail probability. **X** is a positive real number. **Df** is the degrees of freedom. It must be a positive number. **Ncp** is the noncentrality parameter.

NcCsValue (Prob,Df,Ncp)

Returns the inverse of the noncentral-Chi-Square distribution. It gives the value, **X**, such that the area to the left is equal to **Prob**. **Prob** is a probability value. It must be between 0 and 1. **Df** is the degrees of freedom. It must be a postive number. **Ncp** is the noncentrality parameter.

NcFprob (F,Df1,Df2,Ncp)

Returns the probability of being less than **F** when **F** follows the noncentral-F distribution. This is the left-tail probability. **F** is a positive real number. **Df1** is the numerator degrees of freedom. It must be a positive number. **Df2** is the denominator degrees of freedom. It must be a positive number. **Ncp** is the noncentrality parameter.

NcFvalue (Prob,Df1,Df2,Ncp)

Returns the inverse of the noncentral-F distribution. It gives the value, F, such that the area to the left is equal to **Prob**. **Prob** is a probability value. It must be between 0 and 1. **Df1** is the numerator degrees of freedom. It must be a positive number. **Df2** is the denominator degrees of freedom. It must be a positive number. **Ncp** is the noncentrality parameter.

NcTprob (T,Df,Ncp)

Returns the probability of being less than **T** when **T** follows the noncentral-t distribution. This is the left-tail probability. **T** is any real number. **Df** is the degrees of freedom. It must be a positive number. **Ncp** is the noncentrality parameter.

NcTvalue (Prob,Df,Ncp)

Returns the inverse of the noncentral-t distribution. It gives the value, **X**, such that the area to the left is equal to **Prob**. **Prob** is a probability value. It must be between 0 and 1. **Df** is the degrees of freedom. It must be a positive number. **Ncp** is the noncentrality parameter.

NegBinomProb (X,R,P)

Returns the probability of the negative binomial distribution with number of successes: **R** and probability of success: **P**. This is the left-tail probability. **X** is a nonnegative number less than or equal to **R**. **R** is the number of trials resulting in a success and must be a positive integer. **P** is the probability of a success on a particular trial. It must be between 0 and 1.

NormalProb (Z)

Returns the probability of being less than **Z** when **Z** follows the standard-normal distribution. This is the left-tail probability. **Z** is any real number.

NormalValue (Prob)

Returns the inverse of the standard normal distribution. It gives the value, **X**, such that the area to the left is equal to **Prob**. **Prob** is a probability value. It must be between 0 and 1.

PoissonProb (X,Mean)

Returns the probability of the Poisson distribution with given mean. This is the left-tail probability. **X** is a non-negative number. **Mean** is the mean number of occurrences per unit time.

StdRangeProb (R,Nm,Df)

Returns the probability of being less than or equal to **R** when **R** follows the Studentized-Range distribution. This is the left-tail probability. **R** is a positive number. **Nm** is the number of means included in the range. It must be between 2 and 200. **Df** is the degrees of freedom of the standard-error term used in the denominator.

StdRangeValue (Prob,Nm,Df)

Returns the inverse of the Studentized-Range distribution. It gives the number, **R**, such that the Studentized-Range probability is less than or equal to **Prob**. **Prob** is a real number between 0.90 and 0.99. **Nm** is the number of means included in the range. **Df** is the degrees of freedom of the standard-error term used in the denominator.

Tprob (T,Df)

Returns the probability of being less than **T** when **T** follows the Student's t distribution. This is the left-tail probability. **T** is any real number. **Df** is the degrees of freedom. It must be a positive number.

Tvalue (Prob,Df)

Returns the inverse of the Student's t distribution. It gives the value, **X**, such that the area to the left is equal to **Prob**. **Prob** is a probability value. It must be between 0 and 1. **Df** is the degrees of freedom. It must be a postive number.

WeibullProb (T,Scale,Shape)

Returns the probability of being less than or equal to **T** when **T** follows the Weibull distribution. This is the left-tail probability. **T** is a positive number. **Scale** is the scale parameter. **Shape** is the shape parameter.

WeibullValue (Prob,Scale,Shape)

Returns the inverse of the Weibull distribution. It gives the number, **T**, such that the Weibull probability is less than or equal to **Prob**. **Prob** is a real number between 0 and 1. **Scale** is the scale parameter. **Shape** is the shape parameter.

Random-Number Functions

RandomNormal (X)

Returns a random number from the standard normal distribution (mean 0, sigma 1). **X** is the seed of the random generator. It must be an integer greater than zero and less than 32,000.

Uniform (X)

Returns a random number from the uniform distribution. The number is between 0 and 1. **X** is ignored.

Rearrangement Functions

Collate (Type,X1,X2)

Returns a single variable containing the data in the variables **X1, X2**, ... The data are put into the new variable one row at a time. This function cannot be used with any other function. **Type** determines whether the data **(Type=0)**, the row number **(Type=1)**, or the column number **(Type=2)** is stored. **X1,X2, ...** is a list of variables to be combined.

Example: **Collate(0,X1:X2,X3)**

X1	X2	X3	Result
1	5	15	1
2	4	32	5
3	3	55	15
			2
			4
			32
			3
			3
			55

Sort (X)

Returns a sorted version of the data in **X**. Note that only this variable is sorted. This function cannot be used with any other function. **X** is the variable to be sorted.

Splice (Type,X1,X2, ...)

Returns a single variable containing the data in the variables **X1, X2, ...** The data are put into the new variable one column at a time. This function cannot be used with any other function. **Type** determines whether the data **(Type=0),** the row number **(Type=1)**, or the column number **(Type=2)** is stored. **X1,X2, ...** is a list of variables to be combined.

Example: **Splice(0,X1:X2,X3)**

X1	X2	X3	Result
1	5	15	1
2	4	32	2
3	3	55	3
			5
			4
			3
			15
			32
			55

UnCollate (N,I,X)

Returns a portion of a variable. The opposite of Collate. This function cannot be used with any other function. **N** is the total number of variables to be created. **I** is the sequence number of the variable created. **1 <= I <= N**. Every $\mathbf{I}^{th}$ number is copied. **X** is the variable to be partitioned.

Example: the following three transformations were used to generate variables **X2**, **X3**, and **X4**, respectively.

UnCollate(3,1,X1)

UnCollate(3,2,X1)

UnCollate(3,3,X1)

X1	X2	X3	Result
1	1	2	3
2	4	5	6
3	7	8	9
4			
5			
6			
7			
8			
9			

Uniques (X)

Returns the ordered unique values of a variable. This function cannot be used with any other function. **X** is the variable from which the unique values are obtained.

Example: the following transformation was used to generate variable **X2**.

Uniques(X1)

X1	X2
1	1
3	2
2	3
3	4
2	
4	
3	
4	
2	

UnSplice(N,I,X)

Returns a portion of a variable. The opposite of **Splice**. This function cannot be used with any other function. **N** is the total number of variables to be created. **I** is the sequence number of the variable created. **1 <= I <= N**. The **I^{th}** group of values is copied. **X** is the variable to be partitioned.

Example: the following three transformations where used to generate variables **X2**, **X3**, and **X4,** respectively.

UnSplice(3,1,X1)

UnSplice(3,2,X1)

UnSplice(3,3,X1)

X1	X2	X3	X4
1	1	4	7
2	2	5	8
3	3	6	9
4			
5			
6			
7			
8			
9			

Recode Functions

Lookup (X1,X2,X3,Compare)

Returns the value in **X3** that is on the same row as the value in **X2** that matches the current value of **X1**. This function cannot be used with any other function. **X1** is a text or numeric variable to be matched with **X2**. **X2** is a variable comprised of unique values to be used as a cross-reference table. **X3** is a variable containing values that will be placed in the current variable. Each value is a label for the value on the same row in **X2**. **Compare** is 0 if the match is case-sensitive, 1 if the match is not sensitive to the case of **X2** and **X1**.

Example: In the example below, a lookup transformation creates the variable **CMonthName** from variables **Cmonth**, **Month_Number,** and **Month_Name**. The following expression was placed in the **CmonthName** transformation:

Lookup(Cmonth,Month_Number,Month_Name,0)

Cmonth	CMonthName	Month_Number	Month_Name
2	Feb	1	Jan
1	Jan	2	Feb
5	May	3	Mar
6	Jun	4	Apr
11	Nov	5	May
5	May	6	Jun
12	Dec	7	Jul
1	Jan	8	Aug
5	May	9	Sep
2	Feb	10	Oct
8	Aug	11	Nov
5	May	12	Dec

Recode (X; (A=B) ... (Missing="na") (Else=0))

Recodes the values of **X**. It lets you change the values of one variable according to rules you supply. It provides a type of If-Then transformation. Note that the recode statements (the phrases inside a pair of parentheses) are processed sequentially, so that the last true condition is the one used. This function cannot be used with any other function. **X** is a variable to be recoded. It cannot be an expression. **A** represents the current values of **X** that are to be recoded. These values can be numeric or text. If the value is text, it must be enclosed in double-quotes. A range may be specified using the format **Smallest : Largest**. The colon is the "range" operator. All values between, and including, the two boundaries specified are in the range. If the first boundary is omitted, negative infinity is assumed. If the second boundary is omitted, positive infinity is assumed. The keyword **Else** (for otherwise) may be used to indicate what to enter when none of the conditions are met. The keyword **Missing** may be used to indicate a missing value. **B** is the new value to be assigned if the value of **X** is equal to, or falls in the range, designated as **A**. It may be a numeric value, a text value (between double-quotes), or a variable. It cannot be an expression.

Example: **Recode(Q1;(:0="Below") (1=5) (2:4=6) (5:="Above") (Missing="NA") (Else=Q2))**

Q1	R1
0	Below
1	5
5	Above
4	6
2	6
10	Above
	NA
-1	Below
3	6

File Function

File(XYZ.txt)

This function returns the contents of the file specified between the parentheses. A path to the file may be included in the file name. If no path is specified, the path to the currently open database is used.

The transformation in the file may be a segment or a complete transformation. The contents of the file simply replace the file statement before the transformation is processed. The file may include several lines. When the file is loaded, all carriage returns and line feeds are removed as the file is read, so a multi-line file becomes a single line of text.

This statement is especially useful for handling long transformations that might not easily fit on a single line.

Statistical Functions

Average (X1, X2, X3, X5:X8)

Returns the average of the list of numbers and variables separated with commas. Note that a colon may be used to signify a contiguous set of variables. Hence, **X5:X8** means **X5, X6, X7, X8**.

Count (X1, X2, X5:X8)

Returns the number of non-missing items in the list.

Lagk (X)

Returns the **k**-period lag of a variable. This function cannot be used with any other function. **k** is number of rows to lag. **X** is a variable.

Example: **Lag2(X)**

Ledk (X)

Returns the **k**-period lead of a variable. This function cannot be used with any other function. **k** is number of rows to lead. **X** is a variable.

Example: **Led3(X)**

Mavk (X)

Returns the **k**-period moving average of a variable. This function cannot be used with any other function. **k** is number of rows to be averaged. The current row is the last item in the average. **X** is a variable.

Example: **Mav5(X)**

Max (X1,X2, X5:X8)

Returns the maximum of the items in the list.

Min (X1,X2, X5:X8)

Returns the minimum of the items in the list.

NormScore (X)

Returns the inverse-normal quantiles of **X**. These are standard-normal values, not probabilities. This function cannot be used with any other function. **X** is a variable.

Rank(X)

Returns the rank of the values in **X**. Ties are assigned the average rank. This function cannot be used with any other function. **X** is a variable to be ranked.

Smooth (X)

Returns the 3RSSH-smooth of the values in **X**. This function cannot be used with any other function. **X** is a variable to be smoothed. The 3RSSH is John Tukey's famous median smoother. It behaves much like a moving average operator that down plays the influence of large jumps in the series.

Standardize (X)

Returns a standardized version of the data in **X**. The formula is **New=(X-Mean)/Sigma**. This function cannot be used with any other function. **X** is a variable to be standardized.

Stddev (X1, X2, X5:X8)

Returns the standard deviation of the list of numbers.

Example: **Stddev(X1:X20)** yields the standard deviation of variables **X1** through **X20**.

Sum (X1, X2, X5:X8)

Returns the sum of the list of numbers. Missing values are treated as zeros.

Trigonometric Functions

ArCosh (X)

Returns the inverse hyperbolic cosine of **X**. **X** is a number >= 1.

ArSinh (X)

Returns the inverse hyperbolic sine of **X**. **X** is a number.

ArSin(X)

Returns the arc sine (the angle in radians whose sine is **X**) of **X**. Note that **X** is a number between **-1** and **1**.

ArTan (X)

Returns the arc tangent (the angle in radians whose tangent is **X**) of **X**.

ArTanh(X)

Returns the inverse hyperbolic tangent of **X**. **X** is a number between -1 and 1.

Cos (X)

Returns the cosine of the angle **X** (in radians).

Cosh (X)

Returns the hyperbolic cosine of **X**. **X** is a number whose absolute value is less than 705.

Sin (X)

Returns the sine of the angle **X**. **X** is an angle in radians.

Sinh (X)

Returns the hyperbolic sine of **X**. **X** is a number whose absolute value is less than 705.

Tan (X)

Returns the tangent of the angle **X**. **X** is an angle in radians.

Tanh (X)

Returns the hyperbolic tangent of **X**. **X** is a number whose absolute value is less than 705.

Text Functions

Text functions have a name and one or more arguments contained within parentheses. Inside an expression, a function is treated just like a number or a variable. The following functions are designed to work specifically with text data.

Contains (X,Chars,Logic)

Returns a numeric value indicating the position of **Chars** in **X. X** is a value or variable. **Chars** is the text to be searched for. **Logic** is 0 if case-sensitive, 1 if the case of the letters does not matter.

Example: **Contains("The box of oranges","BOX",1)** yields **5.**

Extract (X,First,Nchar)

Returns a value made by extracting from **X** the next **Nchar** characters, beginning at position **First**. **X** is text or a numeric value. **First** is the beginning position. Each character in **X** takes up one position. **Nchar** is the number of characters. If the end is reached before **Nchar** characters are obtained, the number of characters is reduced to the number available.

Example: **Extract("HINTZE",2,3)** yields **"INT"**

Join (X,Y)

Returns the text made by connecting the text in X (on the left) to that in Y (on the right). **X** is text. If it is a number, it is used as a text value. **Y** is text. If it is a number, it is used as a text value.

Example: **Join("LOG","CABIN")** yields **"LOGCABIN"**

Lcase (X)

Returns the lower case of the letters contained in **X**. Non-letters are unchanged. **X** is a variable.

Example: **Lcase("CaT")** yields **"cat"**

Left (X,Nchar)

Returns a value made of the first **Nchar** characters of **X**. **X** is text or a numeric value. **Nchar** is the number of characters. If the end is reached before **Nchar** characters are obtained, the number of characters is reduced to the number available.

Example: **Left("Howdy",3)** yields **"How"**

Length (X)

Returns the number of characters (letters, numbers, spaces, etc.) in **X**. **X** is a variable.

Example: **Length("Computer Program")** yields **16**.

Remove (X,Old)

Returns a value made by removing the characters in **Old** from **X**. **X** is a value or variable. **Old** is the text (enclosed in double quotes) to be removed.

Example: **Remove("SMILES","MILE")** yields **"SS"**

Repeat (X,N)

Returns a value made by repeating **X** a total of **N** times. **X** is a numeric value, a text value (enclosed in double quotes), or a variable. **N** is the number of times **X** is repeated.

Example: **Repeat("AB",3)** yields **"ABABAB"**

Replace (X,Old,New)

Returns a value made by replacing the characters in **OLD** with the characters in **NEW** in the variable **X**. **X** is a variable. **Old** is the text (enclosed in double quotes) to be replaced. **New** is the text (enclosed in double quotes) that will be inserted.

Example: **Replace("XOXOX","O","AA")** yields **"XAAXAAX"**

Right (X,Nchar)

Returns a value made of the last **Nchar** characters of **X**. **X** is text or a numeric value. **Nchar** is the number of characters. If the beginning is reached before **Nchar** characters are obtained, the number of characters is reduced to the number available.

Example: **Right("Howdy",2)** yields **"dy"**

Ucase (X)

Returns the upper case of the letters contained in **X**. Non-letters are unchanged. **X** is a variable.

Example: **Ucase("cat")** yields **"CAT"**

Indicator Variables

The *Recode* function will usually suffice for recoding data. However, there are times when a richer set of transformations is needed. The techniques described here will let you combine logic statements in any way you like to form very complex, logical transformations.

An indicator function is one if a condition is true and zero if the condition is false. Indicator variables are used a great deal in regression analysis. The following two examples would give identical results (note which one is simpler):

(X<4)

Recode(X; (:4=1)(Else=0))

This technique is based on the logic operators which were discussed above. To review, the statement **(X<4)** results in a "0" if the statement is false and a "1" if the statement is true. Note that the less-than symbol may be replaced with any of the other logic symbols. The following table presents a few examples of the types of expressions that may be generated.

Transformation	X=0	X=1	X=2
(X<1)	1	0	0
(X<=1)	1	1	0
(X<1)*4	4	0	0
(X<1)+(X<2)	2	1	0
(X=1)	0	1	0
(X=1)*2+(X=2)*4	0	2	4

The next concept that you need to understand is that the logical "OR" is represented by adding two statements and the logical "AND" is formed by multiplying two statements. Hence, the statement

If X is equal to two or six, recode it to four

would be represented by the expression

4*[(X=2)+(X=6)].

Likewise, the statement

if X is less than four and Y is greater than 2, set the result equal to 2

might be written

2*[(X<4) * (Y>2)].

Now, building upon these few simple concepts you can build any logic operator you want.

Chapter 120

If-Then Transformations

Introduction

If-Then transformations, accessed from the Data menu, are used to generate data when certain conditions are met. They are similar to regular transformations. However, there two important differences. First, the transformation is only executed when the conditions are met. Second, the transformation does not reside with the database. Rather, it stays in the If-Then window.

For example, you might want to use the following set of if-then transformations to create a new variable called Index based on the values of two existing variables: Gender and State.

```
IF (Gender = “M”) AND (State = “Florida”) THEN Index = 1
IF (Gender = “M”) AND (State <> “Florida”) THEN Index = 2
IF (Gender = “F”) AND (State = “Florida”) THEN Index = 3
IF (Gender = “F”) AND (State <> “Florida”) THEN Index = 4
```

Notice that the syntax of these transformations is similar to that of the regular transformations.

Note that the If-Then Transformation window may be used to specify unconditional transformations. This is done by leaving the condition blank. A blank condition is assumed to be true.

If-Then Syntax

The basic syntax if the If-Then transformation is

```
IF condition THEN variable=expression
```

The syntax of the expressions is the same as for regular transformations. You might review the Transformations chapter if you want more details on writing transformations.

If-Then Transformation Specification

If (these conditions are true)

A *condition* is an expression made up of two quantities separated by a logic operator. The result of a condition is either *true* or *false*. Usually, a condition is enclosed in parentheses, but this is not always necessary with simple expressions. Examples of conditions are

(X < 2)

(Y = “Apple”)

(X + Y < X * W)

A *false* expression is one that results in a zero. For example, the expression (5 < 3) is false and thus (5 < 3) = 0. A *true* expression results in a nonzero value, usually negative one. Thus, (5 > 3) = -1.

All you have to remember is that if the condition is true or results in a nonzero number, the transformation is executed. Otherwise, it is not.

AND/OR Statements

You may use the AND and OR statements to string together several logical statements. When you do, you must enclose the individual logic statements in parentheses. Note that the NOT statement is not supported.

Examples using these statements are

(X < 2) AND (Y = “Apple”)

(X < 2) OR (Y = “Apple”)

(Max(X1,X2,X3) < 2) OR (Average(X3,X4,X5)< 70)

Then (result)

The second portion of the If-Then transformation defines the *result variable* that will receive the new value and the *expression* that defines that new value.

The *result variable* must exist on your database. You can define new result variables by changing the variable names on your database. For example, suppose you want to define a new variable called “Status” on your database. By default, the variables are named sequentially from C1 to C256. Suppose your existing data uses the first ten columns of the database. Hence, you decide to change the name of the eleventh column from C11 to Status. This is done in the Variable Info screen.

The *expression* is any valid transformation as described in the Transformation chapter with a few exceptions. A few examples of valid statements are

C1 = 4

Status = “On”

X = Y + 4

Pct = (X1 + X2) / 100

Symbols and Functions

The expression may include any of the following symbols and functions.

Symbols

*, /, +, -, ^, <, >, =, <>, <=, >=.

Date Functions

Day(J), Julian(M,D,Y), Month(J), Year(J).

File Function

File(filename.txt).

Math Functions

Abs(X), Exp(X), Fraction(X), Int(X), Ln(X), Log(X), LogBase(X,Base), Logit(X), Mod(X,Y), Round(X, D), Short(X), Sign(X), Sqrt(X).

Probability Functions

BetaProb(X,A,B), BetaValue(P,A,B), BinomProb(R,N,P), BinomValue(Prob,N,P), BiNormProb(X,Y,Rho), CorrProb(R,N,Rho), CorrValue(Prob,N,Rho), CSProb(X,DF), CSValue(Prob,DF), ExpoProb(X, Scale), ExpoValue(Prob,Scale), Fprob(F,DF1,DF2), Fvalue(Prob,DF1,DF2), GammaProb(X,A), GammaValue(Prob,A), HyperGeoProb(X,N,R,M), LogGamma(X), NCBetaProb(X,A,B,NCP), NCBetaValue(Prob,A,B,NCP), NCCSProb(X,DF,NCP), NCCSValue(Prob,DF,NCP), NCFProb(F,DF1,DF2,NCP), NCFValue(Prob,DF1,DF2,NCP), NCTProb(T,DF,NCP), NCTValue(Prob,DF,NCP), NegBinomProb(X,R,P), NormalProb(Z), NormalValue(Prob), PoissonProb(X,Mean), StdRangeProb(R,N,DF), StdRangeValue(Prob,N,DF), Tprob(X,DF), Tvalue(Prob,DF), WeibullProb(X,Scale,Shape), WeibullValue(Prob,Scale,Shape).

Random-Number Functions

RandomNormal(X), Uniform(X).

Statistical Functions

Average(X1,X2,X3), Count(X1,X2,X3), Max(X1,X2,X3), Min(X1,X2,X3), StdDev(X1,X2,X3), Sum(X1,X2,X3).

Text Functions

Contains(X,Chars,Logic), Extract(X,First,Nchar), Join(X,Y), Lcase(X), Left(X,Nchar), Length(X), Remove(X,Old), Repeat(X,N), Replace(X,Old,New), Right(X,Nchar), Ucase(X)

Trigonometric Functions

ArCosh(X), ArSinh(X), ArSin(X), ArTan(X), ArTanh(X), Cos(X), Cosh(X), Sin(X), Sinh(X), Tan(X), TanH(X).

Calculation Order

If-Then transformations are applied row-by-row, from first to last. It is important to understand the calculation order that is used when there are several If-Then transformations. The program reads in the values of all variables used in all of the transformation expressions. These initial values are not changed while all of the conditions and expressions are calculated, even if they are also result variables. The result variables then receive their new values depending on the whether the conditions are true or false.

Number of Rows

The number of rows processed by the transformations is controlled by an option displayed above the spreadsheet in the toolbar directly to the right of the calculator icon. The default value of this option is zero. A zero indicates that the largest row among the active variables containing data is

used to determine the number of transformed rows. When you set this value to an integer greater than zero, this is the number of rows that will be processed regardless of the number of rows with data in other variables.

Missing Statement

A missing value is specified by using the word MISSING in the expression.

ELSE Statement

There is no ELSE statement supported by If-Then transformations because it is not necessary. To implement an ELSE statement, simply list the ELSE result first without a condition. In the following example, the user wants a zero to be inserted when both of the other two If-Then conditions are false.

```
IF  THEN C4 = 0
IF (C1<5) AND (C2 = 4) THEN C4 = 1
IF (C1<5) AND (C2 = 3) THEN C4 = 2
```

Examples

Following are several sets of If-Then transformations to give you an idea of how they look.

```
IF (C1 < C6+C2) THEN C10 = C1
IF (C1 = 4) THEN C10 = C2+LOG(C3)
IF (C1 = “Apple”) THEN C10 = “Fruit”
IF (C1 = Missing) THEN C10 = 999
IF (C1 < -999) THEN C10 = 14*C1 + 23.4*C2 +13.7
```

Chapter 121

Filter

Introduction

This chapter explains how to use *filters* to limit which rows (observations) are used by the other procedures and which are skipped. For example, you might want to limit an analysis to those weighing over 200 pounds. You would use a filter to accomplish this.

Procedure Options

This section describes the options available in this procedure.

Statements Tab

This tab is used to enter the desired filter statements, as well as all filter options.

Filter Specification

Filter System Active

This statement must be checked for the Filter to be activated.

Note: You must RUN this screen to activate (or de-activate) the Filter System.

Keep Spreadsheet Row If:

You can specify several filter expressions. This option specifies how these expressions are combined.

- **OR**

  If you select the 'OR' option, the condition on only one of the expressions must be met to retain the row in the analysis.

- **AND**

  If you select the 'And' option, the conditions in all of the expressions must be met in order to retain the row in the analysis.

Filter Statements

These boxes contain the filter statements. Each box may contain several filter expressions separated by semicolons.

Note that text must be enclosed in double quotes.

Examples: C1<5; C2=4; C3=1,2,3; C4<C5; C6<C7+C8; C1<>Missing

Syntax

The basic syntax of a filter statement is

VARIABLE LOGIC OPERATOR VALUE

where VALUE is an expression the yields a text value or a number constructed from variable names, numbers, and the symbols +, -, *, and / (add, subtract, multiply, and divide). Note that parentheses are not allowed. A list of values may be used.

Examples of valid VALUE expressions are

C1

Height

1.0

C1+5

Height/100

1,2,3,4,5

X+Y/100+4

Missing (may be used to indicate a missing value)

“John” (must be enclosed in DOUBLE quotes)

Logic Operator

The Logic Operator is one of the following operators:

= Equal to

<> Not equal

< Less than

<= Less than or equal

> Greater than

>= Greater than or equal

Note that only one operator may be specified in an expression.

Variable Name Locator (for pasting variable names into filter statements)

The selected variable name may be copied to the clipboard and pasted into the filter statement.

This field provides no active options. It is here to let you have easy access to a list of all variable names on the current database.

Filter Specification – Filter Statement Comparison Option

Comparison Fuzz Factor

When you make a comparison, you may want to allow for a certain amount difference between two numbers that may occur because of rounding error, etc. For example, you may want the statement .3333=.3334 to evaluate to true instead of false. If the fuzz factor is set to .000001, this expression will be false. However, if the fuzz factor is set to .0001, then this expression will be true.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Using the Filter

Using the MAMMALS database, we will setup up a filter so that only those animals with a body weight greater than 200 kilograms are used in the statistical calculations.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Filter window.

1 Open the MAMMALS dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **MAMMALS.S0**.
- Click **Open**.

2 Open the Filter window.

- On the menus, select **Data**, then **Filter**. The Filter procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the filter options.

- Select the **Statements tab**.
- Check the **Filter System Active** box.
- Set **Keep Spreadsheet Row If** to **If at least one statement is true (OR)**.
- Under **Filter Statements** enter **Body_Weight>200**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

The filter is now active. This procedure does not produce any output. All analyses run after this filter is activated (run) will use only the first 6 rows of the database (those with weights greater than 200).

Disabling the Filter

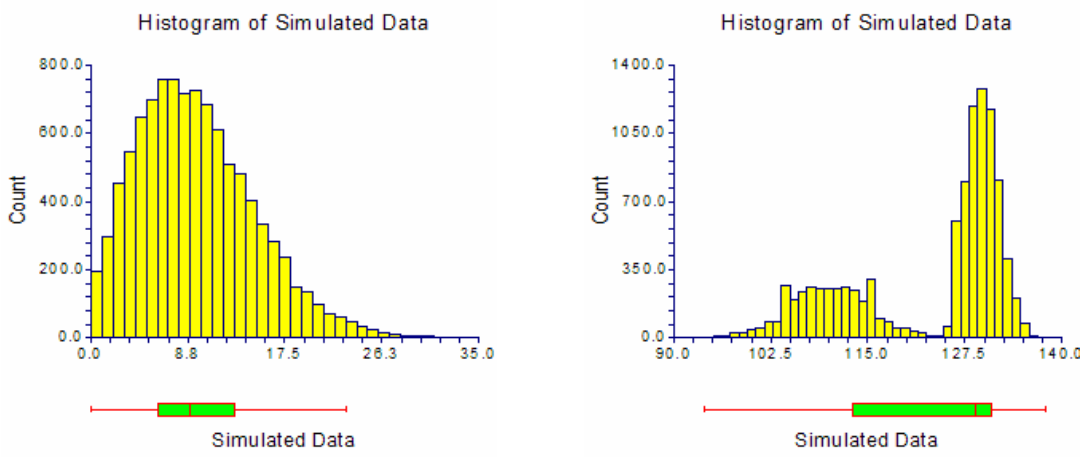
When you are finished using a filter, you can bring up the Filter procedure window, click the **Filter System Active** button so that it is not checked, and press the **Run** button to run the filter procedure. This will deactivate the filter.

Chapter 122

Data Simulator

Introduction

Because of mathematical intractability, it is often necessary to investigate the properties of a statistical procedure using *simulation* (or *Monte Carlo*) techniques. In power analysis, *simulation* refers to the process of generating several thousand random samples that follow a particular distribution, calculating the test statistic from each sample, and tabulating the distribution of these test statistics so that the significance level and power of the procedure may be investigated. This module creates a histogram of a specified distribution as well as a numerical summary of simulated data. By studying the histogram and the numerical summary, you can determine if the distribution has the characteristics you desire. The distribution formula can then be used in procedures that use simulation, such as the new t-test procedures. Below are examples of two distributions that were generated with this procedure.



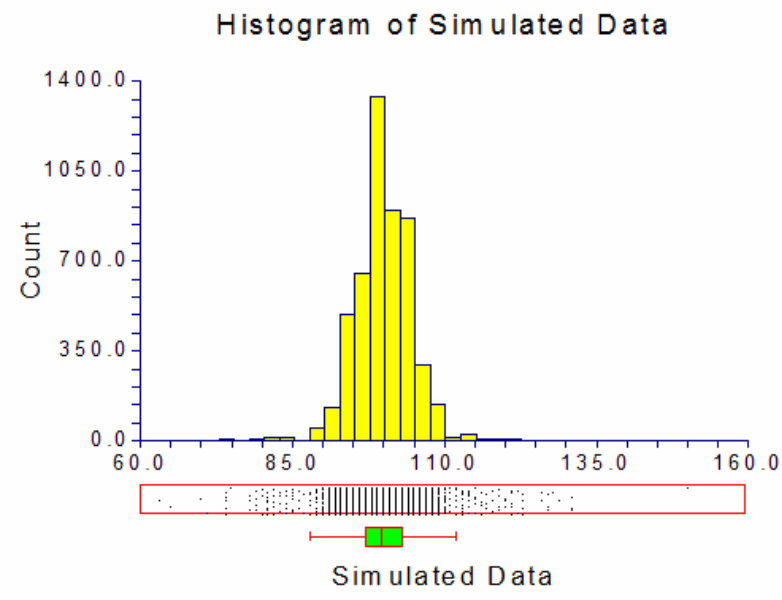
Technical Details

A random variable's probability distribution specifies its probability over its range of values. Examples of common continuous probability distributions are the normal and uniform distributions. Unfortunately, experimental data often do not follow these common distributions, so other distributions have been proposed. One of the easiest ways to create distributions with desired characteristics is to combine simple distributions. For example, outliers may be added to a distribution by mixing it with data from a distribution with a much larger variance. Thus, to simulate normally distributed data with 5% outliers, we could generate 95% of the sample from a normal distribution with mean 100 and standard deviation 4 and then generate 5% of the sample from a normal distribution with mean 100 and standard deviation 16. Using the standard notation

for the normal distribution, the composite distribution of the new random variable Y could be written as

$$Y \sim \delta(0 \le X < 0.95)N(100,4) + \delta(0.95 \le X \le 1.00)N(100,16)$$

where X is a uniform random variable between 0 and 1, $\delta(z)$ is 1 or 0 depending on whether z is true or false, N(100,4) is a normally distributed random variable with mean 100 and standard deviation 4, and N(100,16) is a normally distributed random variable with mean 100 and standard deviation 16. The resulting distribution is shown below. Notice how the tails extend in both directions.



The procedure for generating a random variable, *Y*, with the mixture distribution described above is

1. Generate a uniform random number, *X*.
2. If *X* is less than 0.95, *Y* is created by generating a random number from the *N*(100,4) distribution.
3. If *X* is greater than or equal to 0.95, *Y* is created by generating a random number from the *N*(100,16) distribution.

Note that only one uniform random number and one normal random number are generated for any particular random realization from the mixture distribution.

In general, the formula for a mixture random variable, *Y*, which is to be generated from two or more random variables defined by their distribution function $F_i(Z_i)$ is given by

$$Y \sim \sum_{i=1}^{k} \delta(a_i \le X < a_{i+1}) F_i(Z_i), \quad a_1 = 0 < a_2 < \cdots < a_{K+1} = 1$$

Note that the a_i's are chosen so that weighting requirements are met. Also note that only one uniform random number and one other random number actually need to be generated for a particular value. The $F_i(Z_i)$'s may be any of the distributions which are listed below.

Since the test statistics which will be simulated are used to test hypotheses about one or more means, it will be convenient to parameterize the distributions in terms of their means.

Beta Distribution

The beta distribution is given by the density function

$$f(x)=\frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)}\left(\frac{x-C}{D-C}\right)^{A-1}\left(1-\frac{x-C}{D-C}\right)^{B-1},\quad A,B>0,C\le x\le D$$

where *A* and *B* are shape parameters, *C* is the minimum, and *D* is the maximum. In statistical theory, *C* and *D* are usually zero and one, respectively, but the more general formulation used here is more convenient for simulation work. In this program module, a beta random variable is specified as ***A(M, A, B, C)***, where *M* is the mean which is

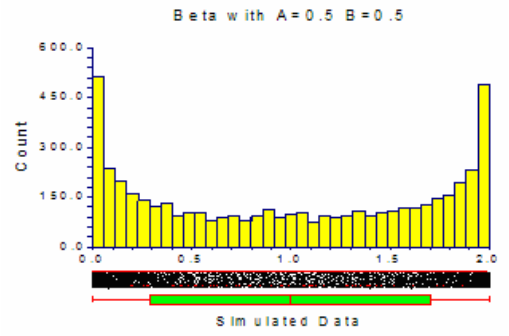
$$E(X)=M=(D-C)\left[\frac{A}{A+B}\right]+C$$

The parameter *D* is obtained from *M*, *A*, *B*, and *C* using the relationship

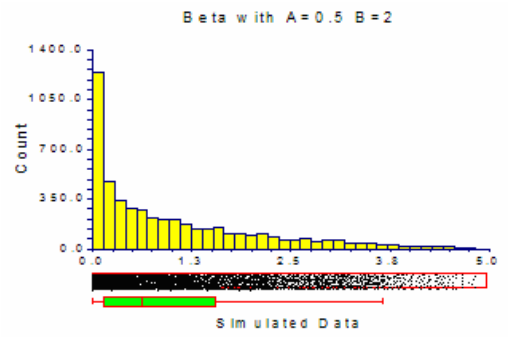
$$D=\frac{(M-C)(A+B)}{A}+C.$$

The beta density can take a number of shapes depending on the values of *A* and *B*:

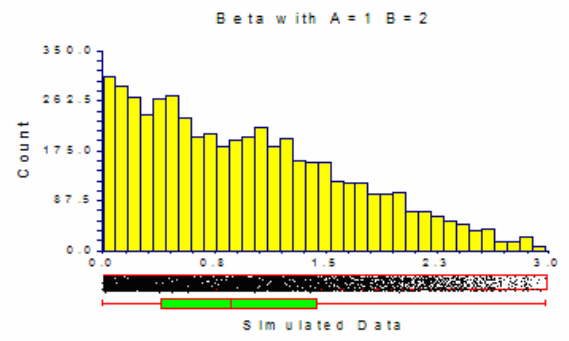
1. When *A*<1 and *B*<1 the density is U-shaped.



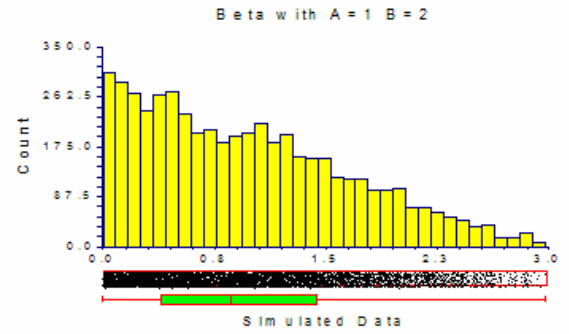
2. When $0<A<1\le B$ the density is J-shaped.



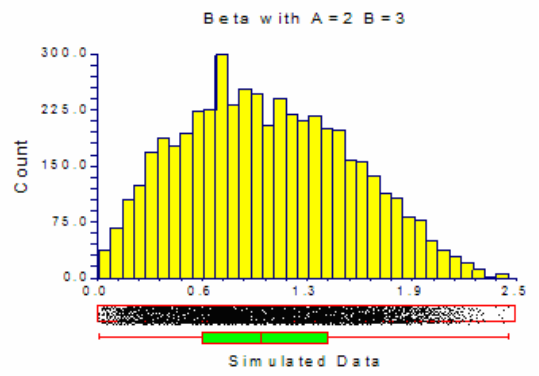
3. When A=1 and B>1 the density is bounded and decreases monotonically to 0.



4. When A=1 and B=1 the density is the uniform density.



5. When A>1 and B>1 the density is unimodal.



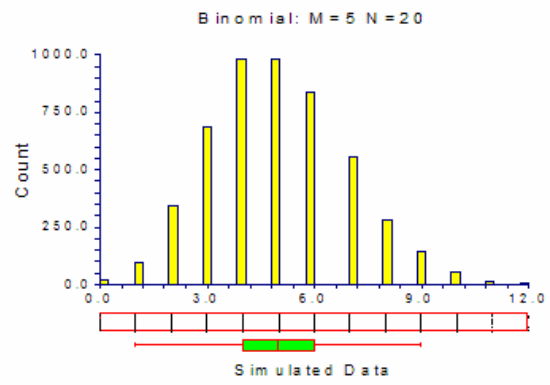
Beta random variates are generated using Cheng's rejection algorithm as given on page 438 of Devroye (1986).

Binomial Distribution

The binomial distribution is given by the function

$$\Pr(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad x = 0,1,2,\ldots,n$$

In this program module, the binomial is specified as ***B(M, n)***, where *M* is the mean which is equal to $n\pi$ and *n* is the number of trials. The probability of a positive response, π, is not entered directly, but is obtained using the relationship $\pi = M / n$. For this reason, $0 < M < n$.



Binomial random variates are generated using the inverse CDF method. That is, a uniform random variate is generated, and then the CDF of the binomial distribution is scanned to determine which value of *X* is associated with that probability.

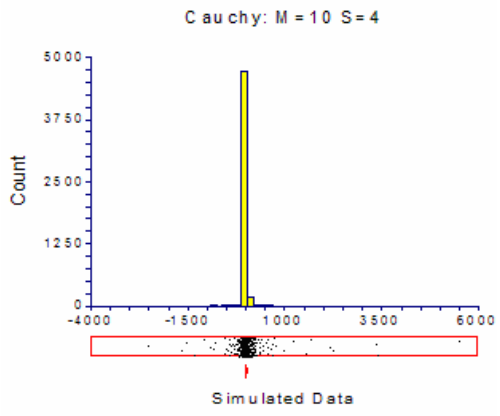
Cauchy Distribution

The Cauchy distribution is given by the density function

$$f(x) = \left[S\pi \left(1 + \left\{ \frac{X - M}{S} \right\}^2 \right) \right]^{-1}, \quad S > 0$$

Although the Cauchy distribution does not possess a mean and standard deviation, *M* and *S* are treated as such. Cauchy random numbers are generated using the algorithm given in Johnson, Kotz, and Balakrishnan (1994), page 327.

In this program module, the Cauchy is specified as ***C(M, S)***, where *M* is a location parameter (median), and *S* is a scale parameter.



Constant Distribution

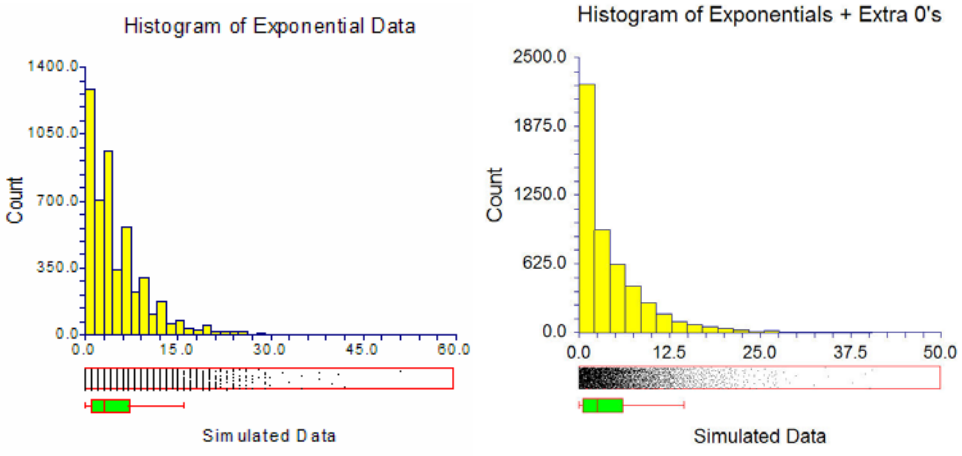
The *constant* distribution occurs when a random variable can only take a single value, *X*. The constant distribution is specified as ***K*(*X*)**, where *X* is the value.

Data with a Many Zero Values

Sometimes data follow a specific distribution in which there is a large proportion of zeros. This can happen when data are counts or monetary amounts. Suppose you want to generate exponentially distributed data with an extra number of zeros. You could use the following simulation model:

```
K(0)[2]; E(5)[9]
```

The exponential distribution alone was used to generate the histogram below on the left. The histogram below on the right was simulated by adding extra zeros to the exponential data.

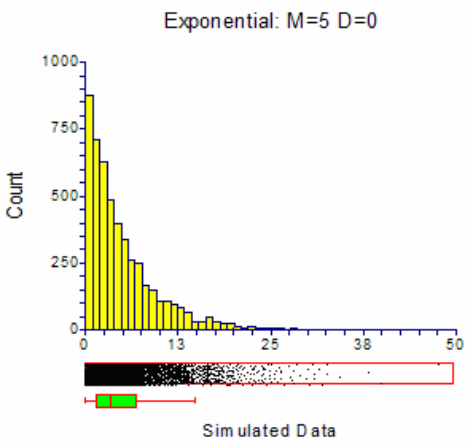


Exponential Distribution

The exponential distribution is given by the density function

$$f(x) = \frac{1}{M} e^{-\frac{x}{M}}, \quad x > 0$$

In this program module, the exponential is specified as ***E*(*M*)**, where *M* is the mean.



Random variates from the exponential distribution are generated using the expression $-M\ln(U)$, where U is a uniform random variate.

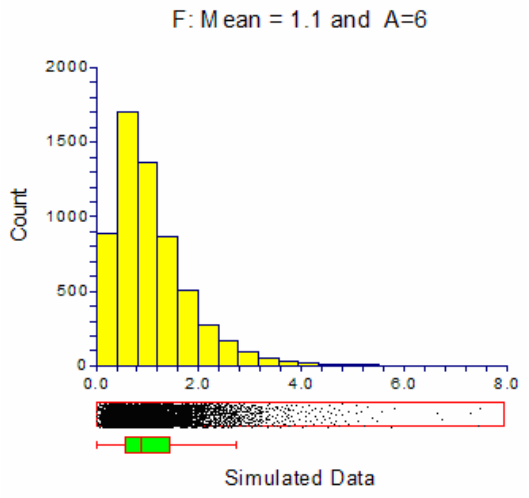
F Distribution

Snedecor's *F* distribution is the distribution of the ratio of two independent chi-square variates. The degrees of freedom of the numerator chi-square variate is *A*, while that of the denominator chi-square is *D*. The *F* distribution is specified as ***F(M, A)***, where *M* is the mean and *A* is the degrees of freedom of the numerator chi-square. The value of *M* is related to the denominator chi-square degrees of freedom using the relationship *M=D/(D-2)*.

F variates are generated by first generating a symmetric beta variate, *B*(*A*/2,*D*/2), and transforming it into an *F* variate using the relationship

$$F_{A,D} = \frac{BD}{A - BA}$$

Below is a histogram for data generated from an *F* distribution with a mean of 1.1 and *A* = 6.



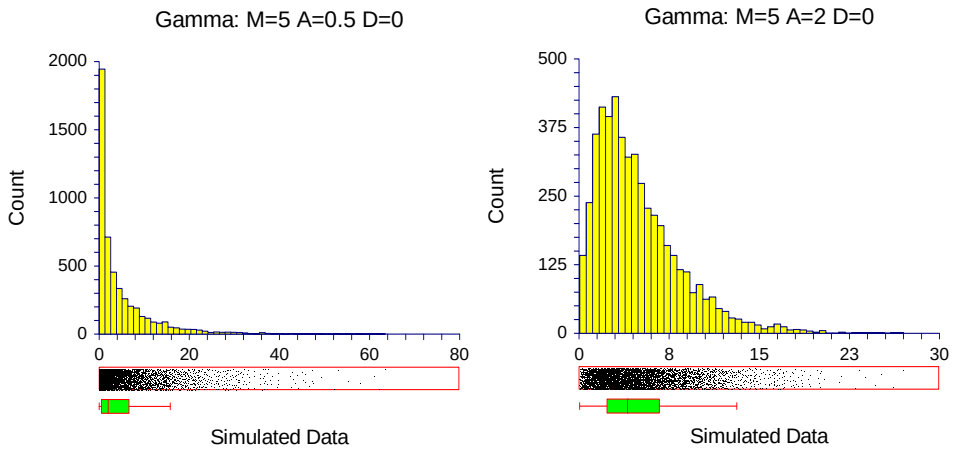
Gamma Distribution

The three parameter gamma distribution is given by the density function

$$f(x) = \frac{(x)^{A-1}}{B^A \Gamma(A)} e^{-\frac{x}{B}}, \quad x > 0,\ A > 0,\ B > 0$$

where *A* is a shape parameter and *B* is a scale parameter. In this program module, the gamma is specified as ***G(M, A)***, where *M* is the mean, given by *M=AB*. The parameter *B* may be obtained using the relationship *B* = *M*/*A*.

Gamma variates are generated using the exponential distribution when *A* = 1, Best's XG algorithm given in Devroye (1986), page 410, when *A* > 1, and Vaduva's algorithm given in Devroye (1986), page 415, when *A* < 1.



Multinomial Distribution

The *multinomial* distribution occurs when a random variable has only a few discrete values such as 1, 2, 3, 4, and 5. The multinomial distribution is specified as ***M(P1, P2, …, Pk)***, where *Pi* is the is the probability of that the integer *i* occurs. Note that the values start at one, not zero.

For example, suppose you want to simulate a distribution which has 50% 3's and 1's, 2's, 4's, and 5's all with equal percentages. You would enter *M(1 1 4 1 1)*.

As a second example, suppose you wanted to have a equal percentage of 1's, 3's, and 7's, and none of the other percentages. You would enter *M(1 0 1 0 0 0 1)*.

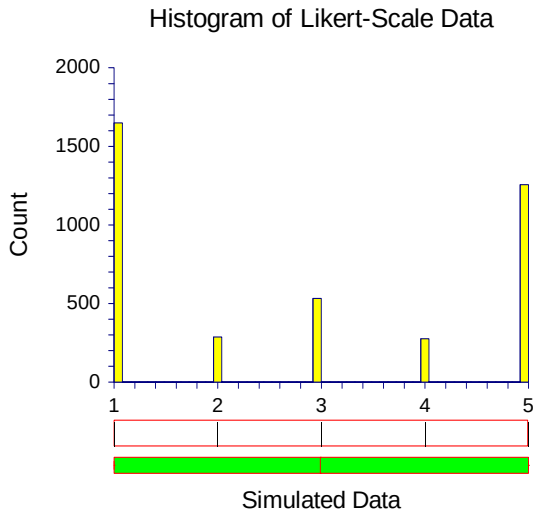
Likert-Scale Data

Likert-scale data are common in surveys and questionnaires. To generate data from a five-point Likert-scale distribution, you could use the following simulation model:

M(6 1 2 1 5)

Note that the weights are relative—they do not have to sum to one. The program will make the appropriate weighting adjustments so that they do sum to one.

The above expression generated the following histogram.



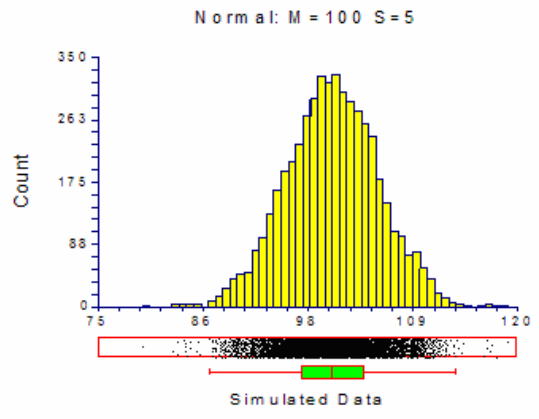
Normal Distribution

The normal distribution is given by the density function

$$f(x) = \phi\left(\frac{x-\mu}{\sigma}\right), \quad -\infty \le x \le \infty$$

where $\phi(z)$ is the usual standard normal density. The normal distribution is specified as ***N(M, S)***, where *M* is the mean and *S* is the standard deviation.

The normal distribution is generated using the Marsaglia and Bray algorithm as given in Devroye (1986), page 390.



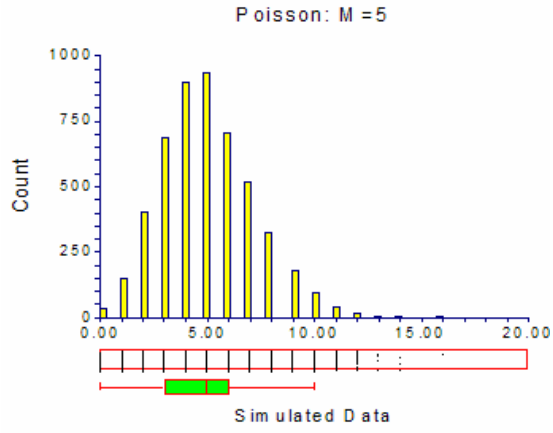
Poisson Distribution

The Poisson distribution is given by the function

$$\Pr(X = x) = \frac{e^{-M} M^{x}}{x!}, \quad x = 0,1,2,\ldots,\ M > 0$$

In this program module, the Poisson is specified as ***P(M)***, where *M* is the mean.

Poisson random variates are generated using the inverse CDF method. That is, a uniform random variate is generated and then the CDF of the Poisson distribution is scanned to determine which value of *X* is associated with that probability.



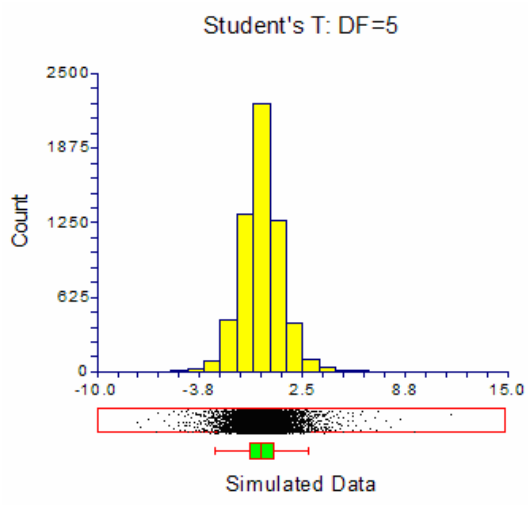
Student's T Distribution

Student's *T* distribution is the distribution of the ratio of a unit normal variate and the square root of an independent chi-square variate. The degrees of freedom of the chi-square variate are the degrees of freedom of the *T* distribution. The *T* is specified as ***T(M, A)***, where *M* is the mean and *A* is the degrees of freedom. The central *T* distribution generated in this program has a mean of zero, so, to obtain a mean of *M*, *M* is added to every data value.

T variates are generated by first generating a symmetric beta variate, B(A/2,A/2), with mean equal to 0.5. This beta variate is then transformed into a *T* variate using the relationship

$$T = \sqrt{A}\frac{X - 0.5}{\sqrt{X(1-X)}}$$

Here is a histogram for data generated from a *T* distribution with mean 0 and 5 degrees of freedom.



Tukey's Lambda Distribution

Hoaglin (1985) presents a discussion of a distribution developed by John Tukey for allowing the detailed specification of skewness and kurtosis in a simulation study. This distribution is extended in the work of Karian and Dudewicz (2000). Tukey's idea was to reshape the normal distribution using functions that change the skewness and/or kurtosis. This is accomplished by multiplying a normal random variable by a skewness function and/or a kurtosis function. The general form of the transformation is

$$X = A + B\left\{G_g(z)H_h(z)z\right\}$$

where *z* has the standard normal density. The skewness function Tukey proposed is

$$G_g(z) = \frac{e^{gz} - 1}{gz}$$

The range of *g* is typically -1 to 1. The value of $G_0(z) \equiv 1$. The kurtosis function Tukey proposed is

$$H_h(z) = e^{hz^2/2}$$

The range of *h* is also -1 to 1.

Hence, if both *g* and *h* are set to zero, the variable *X* follows the normal distribution with mean *A* and standard deviation *B*. As *g* is increased toward 1, the distribution is increasingly skewed to the right. As *g* is decreased towards -1, the distribution is increasingly skewed to the left. As *h* is increased toward 1, the data are stretched out so that more extreme values are probable. As *h* is decreased toward -1, the data are concentrated around the center—resulting in a beta-type distribution.

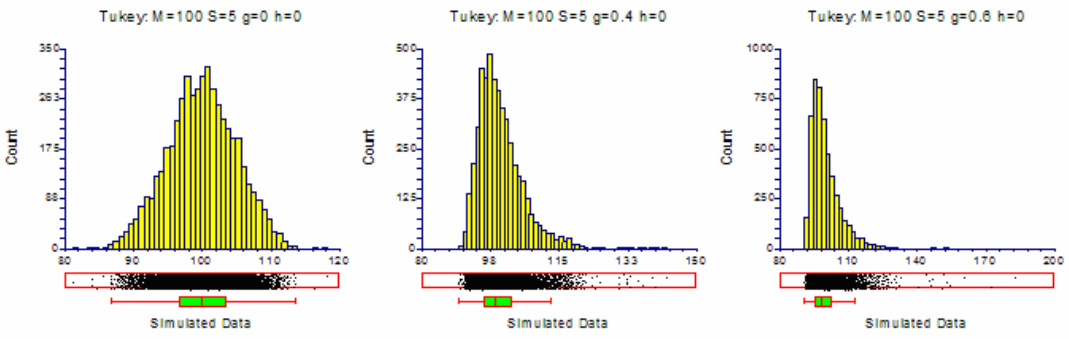
The mean of this distribution is given by

$$M = A + B\left(\frac{e^{g^2/2(1-h)} - 1}{g\sqrt{1-h}}\right), \quad 0 \le h < 1$$

which may be easily solved for *A*.

Tukey's lambda is specified in the program as ***L(M, B, g, h)*** where *M* is the mean, *B* is a scale factor (when *g*=*h*=0, *B* is the standard deviation), *g* is the amount of skewness, and *h* is the amount of kurtosis.

Random variates are generated from this distribution by generating a random normal variate and then applying the skewness and kurtosis modifications. Here are some examples as *g* is varied from 0 to 0.4 to 0.6. Notice how the amount of skewness is gradually increased. Similar results are achieved when *h* is varied from 0 to 0.5.

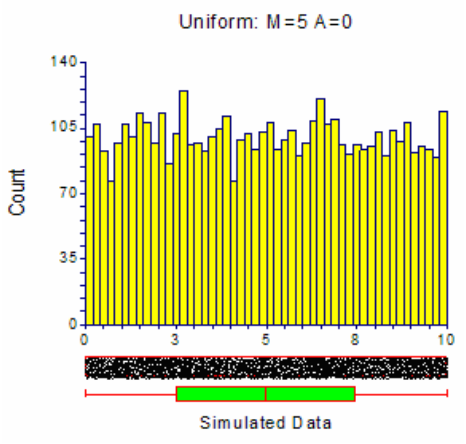


Uniform Distribution

The uniform distribution is given by the density function

$$f(x) = \frac{1}{B - A}, \quad A \le x \le B$$

In this program module, the uniform is specified as ***U(M, A)***, where *M* is the mean which is equal to (*A*+*B*)/2 and *A* is the minimum of *x*. The parameter *B* is obtained using the relationship *B=2M-A*.



Uniform random numbers are generated using Makoto Matsumoto's Mersenne Twister uniform random number generator which has a cycle length greater than 10E+6000 (that's a one followed by 6000 zeros).

Weibull Distribution

The Weibull distribution is indexed by a shape parameter, *B*, and a scale parameter, *C*. The Weibull density function is written as

$$f(x|B,C) = \frac{B}{C}\left(\frac{z}{C}\right)^{(B-1)} e^{-\left(\frac{x}{C}\right)^{B}}, \quad B>0, C>0, \; x>0.$$

Shape Parameter - B

The shape parameter controls the overall shape of the density function. Typically, this value ranges between 0.5 and 8.0. One of the reasons for the popularity of the Weibull distribution is that it includes other useful distributions as special cases or close approximations. For example, if

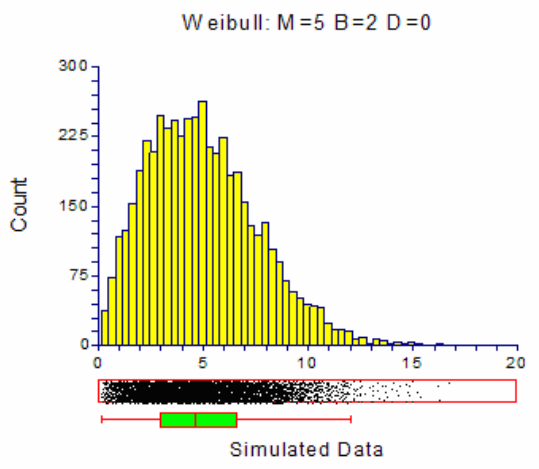
B = 1	The Weibull distribution is identical to the exponential distribution.
B = 2	The Weibull distribution is identical to the Rayleigh distribution.
B = 2.5	The Weibull distribution approximates the lognormal distribution.
B = 3.6	The Weibull distribution approximates the normal distribution.

Scale Parameter - C

The scale parameter only changes the scale of the density function along the *x* axis. Some authors use 1/*C* instead of *C* as the scale parameter. Although this is arbitrary, we prefer dividing by the scale parameter since that is how one usually scales a set of numbers.

The Weibull is specified in the program as ***W(M, B)***, where *M* is the mean which is given by

$$M = C\,\Gamma\left(1+\frac{1}{B}\right).$$



Combining Distributions

This section discusses how distributions may be combined to form new distributions. Combining may be done in the form of algebraic manipulation, mixtures, or both.

Creating New Distributions using Expressions

The set of probability distributions discussed above provides a basic set of useful distributions. However, you may want to mimic reality more closely by combining these basic distributions. For example, paired data is often analyzed by forming the differences of the two original variables. If the original data are normally distributed, then the differences are also normally distributed. Suppose, however, that the original data are exponential. The difference of two exponentials is not a common distribution.

Expression Syntax

The basic syntax is

C1 D1 operator1 C2 D2 operator2 C3 D3 operator3 …

where C1, C2, C3, etc. are coefficients (numbers), D1, D2, D3, etc. are probability distributions, and *operator* is one of the four symbols: +, -, *, /. Parentheses are only permitted in the specification of distributions.

Examples of valid expressions include

N(4, 5) – N(4, 5)

2E(3) – 4E(4) + 2E(5)

N(4, 2)/E(4)-K(5)

Notes about the Coefficients: C1, C2, C3

The coefficients may be positive or negative decimal numbers such as 2.3, 5, or -3.2. If no coefficient is specified, the coefficient is assumed to be one.

Notes about the Distributions: D1, D2, D3

The distributions may be any of the distributions listed above such as normal, exponential, or beta. The expressions are evaluated by generating random values from each of the distributions specified and then combining them according to the operators.

Notes about the operators: +, -, *, /

All multiplications and divisions are performed first, followed by any additions and subtractions.

Note that if only addition and subtraction are used in the expression, the mean of the resulting distribution is found by applying the same operations to the individual distribution means. If the expression involves multiplication or division, the mean of the resulting distribution is usually difficult to calculate directly.

Creating New Distributions using Mixtures

Mixture distributions are formed by sampling a fixed percentage of the data from each of several distributions. For example, you may model outliers by obtaining 95% of your data from a normal distribution with a standard deviation of 5 and 5% of your data from a distribution with a standard deviation of 50.

Mixture Syntax

The basic syntax of a mixture is

D1[W1]; D2[W2]; …; Dk[Wk]

where the D's represent distributions and the *W*'s represent weights. Note that the weights must be positive numbers. Also note that semi-colons are used to separate the components of the mixture.

Examples of valid mixture distributions include

N(4, 5)[19]; N(4, 50)[1]. 95% of the distribution is N(4, 5), and the other 5% is N(4, 50).

W(4, 3)[7]; K(0)[3]. 70% of the distribution is W(4, 3), and the other 30% is made up of zeros.

N(4, 2)-N(4,3)[2]; E(4)*E(2)[8]. 20% of the distribution is N(4, 2)-N(4,3), and the other 80% is E(4)*E(2).

Notes about the Distributions

The distributions D1, D2, D3, etc. may be any valid distributional expression.

Notes about the Weights

The weights w1, w2, w3, etc. need not sum to one (or to one hundred). The program uses these weights to calculate new, internal weights that do sum to one. For example, if you enter weights of 1, 2, and 1, the internal weights will be 0.25, 0.50, and 0.25.

When a weight is not specified, it is assumed to have the value of '1.' Thus

N(4, 5)[19]; N(4,50)[1]

is equivalent to

N(4, 5)[19]; N(4,50)

Special Functions

A set of special functions is available to modify the generator number after all other operations are completed. These special functions are applied in the order they are given next.

Square Root (Absolute Value)

This function is activated by placing a ^ in the expression. When active, the square root of the absolute value of the number is used.

Logarithm (Absolute Value)

This function is activated by placing a ~ in the expression. When active, the logarithm (base e) of the absolute value of the number is used.

Exponential

This function is activated by placing an & in the expression. When active, the number is exponentiated to the base e. If the current number x is greater than 70, exp(70) is used rather than exp(x).

Absolute Value

This function is activated by placing a | in the expression. When active, the absolute value of the number is used.

Integer

This function is activated by placing a # in the expression. When active, the number is rounded to the nearest integer.

Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

Data Tab

The Data tab contains the parameters used to specify a probability distribution.

Data Simulation

Probability Distribution to be Simulated

Enter the components of the probability distribution to be simulated. One or more components may be entered from among the continuous and discrete distributions listed below the data-entry box.

The *W* parameter gives the relative weight of that component. For example, if you entered

P(5)[1];K(0)[2], about 33% of the random numbers would follow the P(5) distribution, and 67% would be 0. When only one component is used, the value of *W* may be omitted. For example, to generate data from the normal distribution with mean of five and standard deviation of one, you would enter N(5, 1), not N(5, 1)[1].

Each of the possible components were discussed earlier in the chapter.

Data Simulation – For Summary and Histogram

Number of Simulated Values

This is the number of values generated from the probability distribution for display in the histogram. We recommend a value of about 5000.

Note that the histogram, box plot, and dot plot row limits must be set higher than this amount or the corresponding plot will not be displayed. These limits are modified by selecting Edit, Options, and Limits from the spreadsheet menu.

Data Simulation – Storage of Simulated Values

Storage of Simulated Values

This is the variables in which the simulated values will be stored. Any data already in this variable will be replaced.

Numbers of Values Stored

This is the number of generated values that are stored in the current database.

Reports Tab

The following options control the format of the reports.

Select Report

Numerical Summary

This option controls the display of this report.

Select Plot

Histogram

This option controls the display of the histogram.

Report Options

Precision

This allows you to specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Report Options – Percentile Options

Percentile Type

This option specifies which of five different methods is used to calculate the percentiles.

RECOMMENDED: **Ave Xp(n+1)** since it gives the common value of the median.

In the explanations below, p refers to the fractional value of the percentile (for example, for the 75th percentile p = .75), Zp refers to the value of the percentile, X[i] refers to the ith data value after the values have been sorted, n refers to the total sample size, and g refers to the fractional part of a number (for example, if np = 23.42, then g = .42). The options are

- **Ave Xp(n+1)**

  This is the most commonly used option. The 100pth percentile is computed as

  $$Zp = (1-g)X[k1] + gX[k2]$$

  where k1 equals the integer part of p(n+1), k2=k1+1, g is the fractional part of p(n+1), and X[k] is the kth observation when the data are sorted from lowest to highest.

- **Ave Xp(n)**

  The 100pth percentile is computed as

  $$Zp = (1-g)X[k1] + gX[k2]$$

  where k1 equals the integer part of np, k2=k1+1, g is the fractional part of np, and X[k] is the kth observation when the data are sorted from lowest to highest.

- **Closest to np**

  The 100pth percentile is computed as

  $$Zp = X[k1]$$

  where k1 equals the integer that is closest to np and X[k] is the kth observation when the data are sorted from lowest to highest.

- **EDF**

  The 100pth percentile is computed as

  $$Zp = X[k1]$$

  where k1 equals the integer part of np if np is exactly an integer or the integer part of np+1 if np is not exactly an integer. X[k] is the kth observation when the data are sorted from lowest to highest. Note that EDF stands for empirical distribution function.

- **EDF w/Ave**

  The 100pth percentile is computed as

  $$Zp = (X[k1] + X[k2])/2$$

  where k1 and k2 are defined as follows: If np is an integer, k1=k2=np. If np is not exactly an integer, k1 equals the integer part of np and k2 = k1+1. X[k] is the kth observation when the data are sorted from lowest to highest. Note that EDF stands for empirical distribution function.

Smallest Percentile

This option lets you assign a different value to the smallest percentile value shown on the percentile report. The default value is 1.0. You can select any value between 0 and 100, including decimal numbers.

Largest Percentile

This option lets you assign a different value to the largest percentile value shown on the percentile report. The default value is 1.0. You can select any value between 0 and 100, including decimal numbers.

Report Options – Report Decimal Places

Means – Values

Specify the number of decimal places used when displaying this item.

GENERAL: Display the entire number without special formatting.

Histogram Tab

This panel sets the options used to define the appearance of the histogram.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Tick Label Settings…

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Major Ticks – Minor Ticks

These options set the number of major and minor tickmarks displayed on the axis.

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Histogram Settings

Plot Style File

Designate a histogram style file. This file sets all options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Histograms procedure.

Number of Bars

Specify the number of bars to be displayed. Select 'Automatic' to direct the program to select an appropriate number based on the number of values.

Titles

Plot Title

This is the text of the title. The characters *{Y}*, *{X}*, and *{G}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Generating Normal Data

In this example, 5000 values will be generated from the standard normal (mean zero, variance one) distribution. These values will be displayed in a histogram and summarized numerically.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Data Simulator window.

1 Open the Data Simulator window.

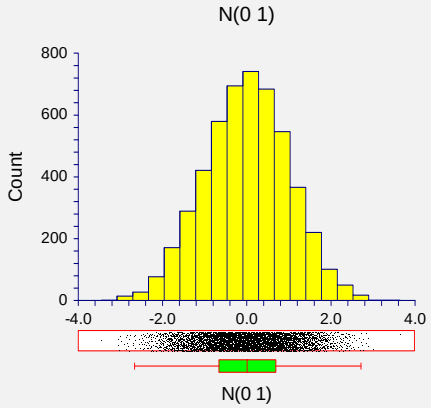
- On the menus of the NCSS Data window, select **Data**, then **Data Simulation**. The Data Simulator procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **N(0 1)** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).



Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	8.627304E-03	Minimum	-3.453244
Standard Deviation	0.9910903	1st Percentile	-2.289698
Skewness	-6.010016E-02	5th Percentile	-1.65369
Kurtosis	2.847444	10th Percentile	-1.298245
Coefficient of Variation	114.8783	25th Percentile	-0.6598607
Count	5000	Median	2.659537E-02
		75th Percentile	0.6922649
		90th Percentile	1.276735
		95th Percentile	1.620931
		99th Percentile	2.274097
		Maximum	3.645533

This report shows the histogram and a numerical summary of the 5000 simulate normal values. It is interesting to check how well the simulation did. Theoretically, the mean should be zero, the standard deviation one, the skewness zero, and the kurtosis three. Of course, your results will vary from these because these are based on generated random numbers.

Example 2 – Generating Data from a Contaminated Normal

In this example, we will generate data from a contaminated normal. This will be accomplished by generating 95% of the data from a N(100,3) distribution and 5% from a N(110,15) distribution.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Data Simulator window.

1 Open the Data Simulator window.

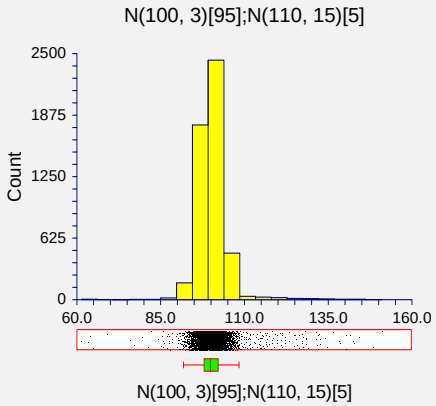
- On the menus of the NCSS Data window, select **Data**, then **Data Simulation**. The Data Simulator procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **N(100, 3)[95];N(110, 15)[5]** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).



Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	100.3549	Minimum	71.11177
Standard Deviation	4.716588	1st Percentile	92.40327
Skewness	2.67692	5th Percentile	94.82752
Kurtosis	23.11323	10th Percentile	96.01687
Coefficient of Variation	4.699908E-02	25th Percentile	97.88011
Count	5000	Median	100.0328
		75th Percentile	102.18
		90th Percentile	104.266
		95th Percentile	105.9543
		99th Percentile	120.4605
		Maximum	145.2731

This report shows the data from the contaminated normal. The mean is close to 100, but the standard deviation, skewness, and kurtosis have non-normal values. Note that there are now some very large outliers.

Example 3 – Likert-Scale Data

In this example, we will generate data following a discrete distribution on a Likert scale. The distribution of the Likert scale will be 30% 1's, 10% 2's, 20% 3's, 10% 4's, and 30% 5's.

You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Data Simulator window.

1 Open the Data Simulator window.

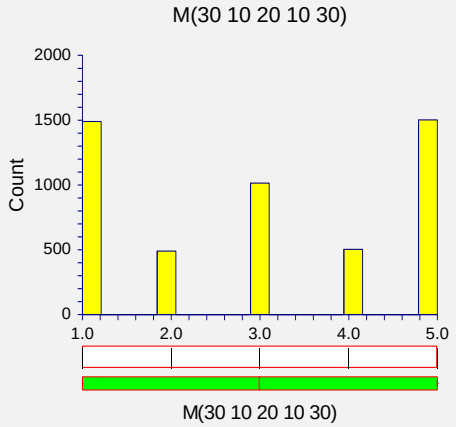
- On the menus of the NCSS Data window, select **Data**, then **Data Simulation**. The Data Simulator procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **M(30 10 20 10 30)** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).



Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	2.9858	Minimum	1
Standard Deviation	1.622506	1st Percentile	1
Skewness	1.169226E-02	5th Percentile	1
Kurtosis	1.436589	10th Percentile	1
Coefficient of Variation	0.5434074	25th Percentile	1
Count	5000	Median	3
		75th Percentile	5
		90th Percentile	5
		95th Percentile	5
		99th Percentile	5
		Maximum	5

This report shows the data from a Likert scale.

Example 4 – Bimodal Data

In this example, we will generate data that have a bimodal distribution. We will accomplish this by combining data from two normal distributions, one with a mean of 10 and the other with a mean of 30. The standard deviation will be set at 4.

You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Data Simulator window.

1 Open the Data Simulator window.

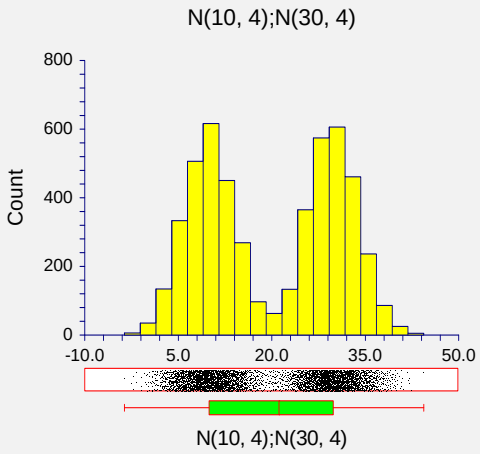
- On the menus of the NCSS Data window, select **Data**, then **Data Simulation**. The Data Simulator procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **N(10, 4);N(30, 4)** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).



Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	19.82392	Minimum	-7.067288
Standard Deviation	10.81724	1st Percentile	1.819361
Skewness	3.503937E-02	5th Percentile	4.949873
Kurtosis	1.500108	10th Percentile	6.473907
Coefficient of Variation	0.545666	25th Percentile	9.863808
Count	5000	Median	17.73235
		75th Percentile	29.94195
		90th Percentile	33.4243
		95th Percentile	35.11861
		99th Percentile	38.04005
		Maximum	43.71141

This report shows the results for the simulated bimodal data.

Example 5 – Gamma Data with Extra Zeros

In this example, we will generate data that have a gamma distribution, except that we will force there to be about 30% zeros. The gamma distribution will have a shape parameter of 5 and a mean of 10.

You may follow along here by making the appropriate entries or load the completed template **Example5** from the Template tab of the Data Simulator window.

1 Open the Data Simulator window.

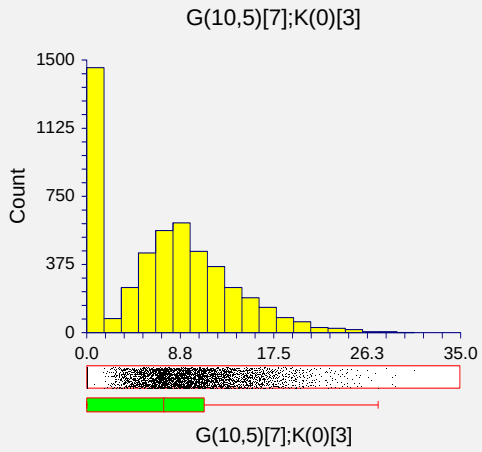
- On the menus of the NCSS Data window, select **Data**, then **Data Simulation**. The Data Simulator procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **G(10,5)[7];K(0)[3]** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).



Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	7.0965	Minimum	0
Standard Deviation	5.960041	1st Percentile	0
Skewness	0.4635893	5th Percentile	0
Kurtosis	2.730829	10th Percentile	0
Coefficient of Variation	0.8398563	25th Percentile	0
Count	5000	Median	7.24275
		75th Percentile	11.09861
		90th Percentile	14.94062
		95th Percentile	17.41686
		99th Percentile	22.18812
		Maximum	36.13172

This report shows the results for the simulated bimodal data.

Example 6 – Mixture of Two Poisson Distributions

In this example, we will generate data that have a mixture of two Poisson distributions. 60% of the data will be from a Poisson distribution with a mean of 10 and 40% from a Poisson distribution with a mean of 20.

You may follow along here by making the appropriate entries or load the completed template **Example6** from the Template tab of the Data Simulator window.

1 Open the Data Simulator window.

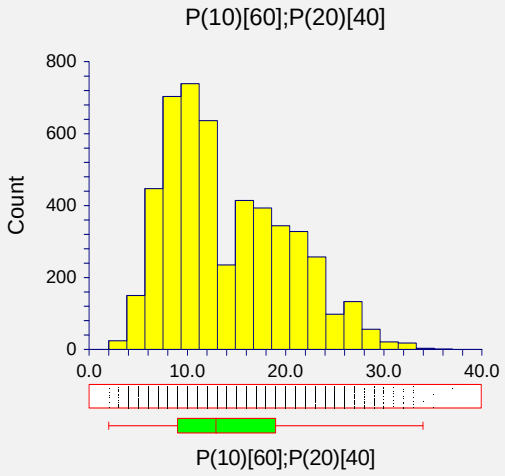
- On the menus of the NCSS Data window, select **Data**, then **Data Simulation**. The Data Simulator procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **P(10)[60];P(20)[40]** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).



Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	13.8662	Minimum	0
Standard Deviation	6.088589	1st Percentile	4
Skewness	0.5726398	5th Percentile	6
Kurtosis	2.580317	10th Percentile	7
Coefficient of Variation	0.4390957	25th Percentile	9
Count	5000	Median	13
		75th Percentile	18
		90th Percentile	23
		95th Percentile	25
		99th Percentile	29
		Maximum	36

This report shows the results for the simulated mixture-Poisson data.

Example7 – Difference of Two Identically Distributed Exponentials

In this example, we will demonstrate that the difference of two identically distributed exponential random variables follows a symmetric distribution. This is particularly interesting because the exponential distribution is skewed. In fact, the difference between any two identically distributed random variables follows a symmetric distribution.

You may follow along here by making the appropriate entries or load the completed template **Example7** from the Template tab of the Data Simulator window.

1 Open the Data Simulator window.

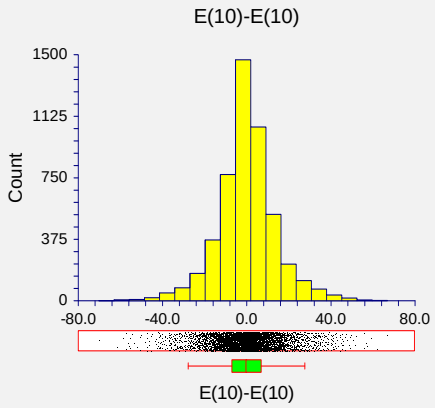
- On the menus of the NCSS Data window, select **Data**, then **Data Simulation**. The Data Simulator procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **E(10)-E(10)** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).



Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	-0.3431585	Minimum	-99.37727
Standard Deviation	14.20288	1st Percentile	-41.02162
Skewness	-7.391038E-02	5th Percentile	-23.19279
Kurtosis	6.115685	10th Percentile	-16.70725
Coefficient of Variation	-41.38868	25th Percentile	-7.131404
Count	5000	Median	-0.1380393
		75th Percentile	6.575472
		90th Percentile	15.78555
		95th Percentile	22.12444
		99th Percentile	38.82129
		Maximum	72.36995

This report shows demonstrates that the distribution of the difference is symmetric.

Chapter 123

Data Matching – Optimal and Greedy

Introduction

This procedure is used to create treatment-control matches based on propensity scores and/or observed covariate variables. Both optimal and greedy matching algorithms are available (as two separate procedures), along with several options that allow the user to customize each algorithm for their specific needs. The user is able to choose the number of controls to match with each treatment (e.g., 1:1 matching, 1:*k* matching, and variable (full) matching), the distance calculation method (e.g., Mahalanobis distance, propensity score difference, sum of rank differences, etc.), and whether or not to use calipers for matching. The user is also able to specify variables whose values must match exactly for both treatment and controls in order to assign a match. ***NCSS*** outputs a list of matches by match number along with several informative reports and optionally saves the match numbers directly to the database for further analysis.

Matching Overview

Observational Studies

In observational studies, investigators do not control the assignment of treatments to subjects. Consequently, a difference in covariates may exist between treatment and control groups, possibly resulting in undesired biases. Matching is often used to balance the distributions of observed (and possibly confounding) covariates. Furthermore, in many observational studies, there exist a relatively small number of treatment group subjects as compared to control group subjects, and it is often the case that the costs associated with obtaining outcome or response data is high for both groups. Matching is used in this scenario to reduce the number of control subjects included in the study. Common matching methods include Mahalanobis metric matching, propensity score matching, and average rank sum matching. Each of these will be discussed later in this chapter. For a thorough treatment of data matching for observational studies, the reader is referred to chapter 1.2 of D'Agostino, Jr. (2004).

The Propensity Score

Ideally, one would match each treatment subject with a control subject (or subjects) that was an exact match on each of the observed covariates. As the number of covariates increases or the ratio of the number of control subjects to treatment subjects decreases, it becomes less and less likely that an exact match will be found for each treatment subject. *Propensity scores* can be used in this situation to simultaneously control for the presence of several covariate factors. The propensity score was introduced by Rosenbaum and Rubin (1983). The propensity score for subject i (i = 1, …, N) is defined as the conditional probability of assignment to a treatment (Z_i = 1) versus the control (Z_i = 0), given a set (or vector) of observed covariates, $\mathbf{x}_i$. Mathematically, the propensity score for subject i can be express as

$$e(\mathbf{x}_i) = \text{pr}(Z_i = 1 \,|\, \mathbf{X}_i = \mathbf{x}_i) .$$

It is assumed that the Z_i's are independent, given the X's. The observed covariates, $\mathbf{x}_i$, are not necessarily the same covariates used in the matching algorithm, $\mathbf{y}_i$, although they could be. Rosenbaum and Rubin (1985a) suggest using the logit of the estimated propensity score for matching because the distribution of transformed scores is often approximately normal. The logit of the propensity score is defined as

$$q(\mathbf{x}) = \log\left(\frac{1 - e(\mathbf{x})}{e(\mathbf{x})}\right),$$

Matching on the observed propensity score (or logit propensity score) can balance the overall distribution of observed covariates between the treatment and control groups. The propensity score is often calculated using logistic regression or discriminant analysis with the treatment variable as the dependent (group) variable and the background covariates as the independent variables. Research suggests that care must be taken when creating the propensity score model (see Austin et al. (2007)). For more information about logistic regression or discriminant analysis, see the corresponding chapters in the ***NCSS*** manuals.

Optimal vs. Greedy Matching

Two separate procedures are documented in this chapter, *Optimal Data Matching* and *Greedy Data Matching*. The goal of both algorithms is to produce a matched sample that balances the distribution of observed covariates between the treatment and matched-control groups. Both algorithms allow for the creation of 1:1 or 1:k matched pairings. Gu and Rosenbaum (1993) compared the greedy and optimal algorithms and found that "optimal matching is sometimes noticeably better than greedy matching in the sense of producing closely matched pairs, sometimes only marginally better, but it is no better than greedy matching in the sense of producing balanced matched samples." The choice of the algorithm depends on the research objectives, the desired analysis, and cost considerations. We recommend using the optimal matching algorithm where possible.

The optimal and greedy algorithms differ in three fundamental ways:

1. Treatment of Previously-Matched Subjects
2. Complete vs. Incomplete Matched-Pair Samples
3. Variable (Full) Matching

Treatment of Previously-Matched Subjects

Optimal matching refers to the use of an optimization method based on the Relax-IV algorithm written by Dimitri P. Bertsekas (see Bertsekas (1991)), which minimizes the overall sum of pair-wise distances between treatment subjects and matched control subjects. The Relax-IV algorithm is based on network flow theory, and matching is just one of its many uses. Optimal matching is not a linear matching algorithm in the sense that as the algorithm proceeds, matches are created, broken, and rearranged in order to minimize the overall sum of match distances.

Greedy matching, on the other hand, is a linear matching algorithm: when a match between a treatment and control is created, the control subject is removed from any further consideration for matching. When the number of matches per treatment is greater than one (i.e., 1:*k* matching), the greedy algorithm finds the best match (if possible) for each treatment before returning and creating the second match, third match, etc. Once a treatment subject has been matched with the user-specified number of control subjects, the treatment subject is also removed from further consideration. A familiar example of a greedy algorithm is forward selection used in multiple regression model creation.

Complete vs. Incomplete Matched-Pair Samples

Optimal matching only allows for *complete matched-pair samples*, while greedy matching also allows for *incomplete matched-pair samples*. A complete matched-pair sample is a sample for which every treatment is matched with at least one control. An incomplete matched-pair sample is a sample for which the number of treatment subjects matched is less than the total number of treatment subjects in the reservoir. Rosenbaum and Rubin (1985b) present strong reasons for avoiding incomplete matched-pair samples.

Variable (Full) Matching

Variable (or "Full") matching is only available using the optimal matching algorithm. In variable matching, a different number of controls may be matched with each treatment. Each control is used only once, and each treatment receives at least one control. All eligible controls (e.g. all controls for which at least one treatment-control distance is non-infinite) are matched. Results from Gu and Rosenbaum (1993) suggest that in terms of bias reduction, full matching performs much better than 1:*k* matching. If we require that every treatment have the same number of controls, and the distributions between the two groups of covariates are not the same, then some treatments will be paired with controls that are not good matches. Variable matching, on the other hand, is more flexible in allowing control subjects to pair with the closest treatment subject in every case.

The gains in bias reduction for variable matching over 1:*k* matching, however, must be weighed against other considerations such as simplicity and aesthetics. The analysis after 1:*k* matching would arguably be more simple; a more complex analysis method (e.g. stratified analysis) would be employed after variable matching than would be after 1:*k* matching.

The Distance Calculation Method

Several different distance calculation methods are available in the matching procedures in ***NCSS***. The different methods are really variations of three common distance measures:

1. Mahalanobis Distance
2. Propensity Score Difference
3. Sum of Rank Differences

The variations arise when using calipers for matching or when using forced match variables. A *caliper* is defined in this context a restricted subset of controls whose propensity score is within a specified amount (c) of the treatment subject's propensity score. A *forced match variable* contains values which must match exactly in the treatment and control for the subjects to be considered for matching. If the values for the forced match variables do not agree, then the distance between the two subjects is set equal to ∞ (infinity), and a match between the two is not allowed.

Distance Measures

The complete list of possible distance measures available in ***NCSS*** is as follows:

1. Mahalanobis Distance within Propensity Score Calipers (no matches outside calipers)
2. Mahalanobis Distance within Propensity Score Calipers (matches allowed outside calipers)
3. Mahalanobis Distance including the Propensity Score (if specified)
4. Propensity Score Difference within Propensity Score Calipers (no matches outside calipers)
5. Propensity Score Difference
6. Sum of Rank Differences within Propensity Score Calipers (no matches outside calipers)
7. Sum of Rank Differences within Propensity Score Calipers (matches allowed outside calipers)
8. Sum of Rank Differences including the Propensity Score (if specified)

Distance measures #2 and #7, where matches are allowed outside calipers in caliper matching, are only available with greedy matching. All others can be used with both the greedy and optimal matching algorithms.

For distance measures that involve propensity score calipers, the caliper size is determined by the user-specified radius, c. For any treatment subject, i, the j^{th}, control subject is included in the i^{th} treatment caliper if

$$|\, q(\mathbf{x}_i) - q(\mathbf{x}_j) \,| \leq c$$

where $q(\mathbf{x}_i) = e(\mathbf{x}_i)$ is the propensity score based on the covariates $\mathbf{x}_i$. If the logit transformation is used in the analysis, then $q(\mathbf{x}) = \log((1 - e(\mathbf{x}))/e(\mathbf{x}))$. The width of each caliper is equal to $2c$.

Which Distance Measure to Use?

The best distance measure depends on the number of covariate variables, the variability within the covariate variables, and possibly other factors. Gu and Rosenbaum (1993) compared the imbalance of Mahalanobis distance metrics versus the propensity score difference in optimal 1:1 matching for numbers of covariates (P) between 2 and 20 and control/treatment subject ratios between 2 and 6. Mahalanobis distance within propensity score calipers was always best or second best. When there are many covariates ($P = 20$), the article suggests that matching on the propensity score difference is best. The use of Mahalanobis distance (with or without calipers) is best when there are few covariates on which to match ($P = 2$). In all cases considered by Gu and Rosenbaum (1993), the Mahalanobis distance within propensity score calipers was never the worst method of the three. Rosenbaum and Rubin (1985a) conducted a study of the performance of three different matching methods (Mahalanobis distance, Mahalanobis distance within propensity score calipers, and propensity score difference) in a greedy algorithm with matches

allowed outside calipers and concluded that the Mahalanobis distance within propensity score calipers is the best technique among the three. Finally, Rosenbaum (1989) reports parenthetically that he has had "unpleasant experiences using standard deviations to scale covariates in multivariate matching, and [he] is inclined to think that either ranks or some more resistant measure of spread should routinely be used instead."

Based on these results and suggestions, we recommend using the Mahalanobis Distance within Propensity Score Calipers as the distance calculation method where possible. The caliper radius to use is based on the amount of bias that you want removed.

What Caliper Radius to Use?

The performance of distance metrics involving calipers depends to some extent on the caliper radius used. For instances in the literature where we found reports, comparisons, or studies based on caliper matching, Cochran and Rubin (1973) was nearly always mentioned as the literature used in determining the caliper radius (or "caliper width" as they call it) for the study. The following table (Table 2.3.1 from Cochran and Rubin (1973)) can be used to determine the appropriate coefficient and/or caliper radius to use:

Table 2.3.1 from Cochran and Rubin (1973). Percent Reduction in bias of *x* for caliper matching to within $\pm a\sqrt{(\sigma_1^2+\sigma_2^2)/2}$

a	$\sigma_1^2/\sigma_2^2 = 1/2$	$\sigma_1^2/\sigma_2^2 = 1$	$\sigma_1^2/\sigma_2^2 = 2$
0.2	0.99	0.99	0.98
0.4	0.96	0.95	0.93
0.6	0.91	0.89	0.86
0.8	0.86	0.82	0.77
1.0	0.79	0.74	0.69

The caliper radius to use depends on the desired bias reduction (table body), the coefficient *a*, and the ratio of the treatment group sample variance of $q(\mathbf{x})$, σ_1^2, to the control group sample variance of $q(\mathbf{x})$, σ_2^2. "Loose Matching" corresponds to $a \geq 1.0$, while "Tight Matching" corresponds to $a \leq 0.2$. The caliper radius is calculated as

$$c = a\sqrt{(\sigma_1^2+\sigma_2^2)/2} = a \times SIGMA$$

NCSS allows you to choose the caliper radius using the syntax "*a**SIGMA", where you specify the value for *a* (e.g. "0.2*SIGMA") or by entering the actual value directly for *c* (e.g. "0.5"). In the case of the former, the program calculates the variances of the treatment and control group propensity scores for you and determines the pooled standard deviation, sigma. You may want to run descriptive statistics on the treatment and control group propensity scores to determine the variance ratio of your data in order to find the appropriate value of *a* (from the table above) for your research objectives.

Data Structure

The propensity scores and covariate variables must each be entered in individual columns in the database. Only numeric values are allowed in propensity score and covariate variables. Blank cells or non-numeric (text) entries are treated as missing values. If the logit transformation is

used, values in the propensity score variable that are not between zero and one are also treated as missing. A grouping variable containing two (and only two) unique groups must be present. A data label variable is optional. The following is a subset of the PROPENSITY dataset, which illustrates the data format required for the greedy and optimal data matching procedures.

PROPENSITY dataset (subset)

ID	Exposure	X1	...	Age	Race	Gender	Propensity
A	Exposed	50	...	45	Hispanic	Male	0.7418116515
B	Not Exposed	4	...	71	Hispanic	Male	0.01078557025
C	Not Exposed	81	...	70	Caucasian	Male	0.0008716385678
D	Exposed	31	...	33	Hispanic	Female	0.5861360724
E	Not Exposed	65	...	38	Black	Male	0.1174339761
F	Exposed	22	...	29	Black	Female	0.07538899371
G	Not Exposed	36	...	57	Black	Female	0.008287371892
H	Not Exposed	31	...	52	Caucasian	Male	0.4250166047
I	Not Exposed	46	...	39	Hispanic	Female	0.2630767334
J	Exposed	3	...	58	Hispanic	Male	0.4858799526
K	Not Exposed	84	...	24	Black	Female	0.1251753736

Procedure Options

This section describes the options available in both the optimal and greedy matching procedures.

Variables Tab

Specify the variables to be used in matching and storage, along with the matching options.

Data Variables

Grouping Variable

Specify the variable that contains the group assignment for each subject. The values in this variable may be text or numeric, but only two unique values are allowed. One value should designate the treatment group and the other should designate the control group. The value assigned to the treatment group should be entered under “Treatment Group” to the right.

Treatment Group

Specify the value in the Grouping Variable that is used to designate the treatment group. This value can be text or numeric.

Propensity Score Variable

Specify the variable containing the propensity scores to be used for matching. This variable is optional if the Distance Calculation Method is not specifically based on the propensity score. If no covariate variables are specified, then you must specify a propensity score variable. If caliper matching is used, this variable must be specified. Only numeric values are allowed. Text values are treated as missing values in the reports. If the logit transformation is used, all values in this variable must be between zero and one, otherwise they are treated as missing. Propensity scores are often obtained using logistic regression or discriminant analysis.

Use Logit

This option specifies whether or not to use the logit transformation on the propensity score. If selected, all calculations and reports will be based on the logit propensity score (where applicable).

Forced Match Variable(s)

Specify variables for which the treatment and control values must match exactly in order to create a match. More than one variable may be specified. This variable is optional. Variables such as gender and race are commonly used as forced match variables.

The use of forced match variables may greatly restrict the number of possible matches for each treatment. If you are using greedy matching, the use of this variable may result in unmatched treatment subjects. If you are using optimal matching, the use of forced match variable(s) may result in an infeasible (unsolvable) problem for the matching algorithm. If the optimal matching algorithm is unable to find a solution, try eliminating one or more forced match variable(s).

Covariate Variable(s)

Specify variables to be used in distance calculations between treatment and control subjects. Only numeric values are allowed. Text values are treated as missing. Covariate variables are optional, however, if no propensity score variable is specified you must specify at least one covariate variable. If the distance calculation method involves only the propensity score (e.g. propensity score difference) and one or more covariate variables are specified, then the covariate variables are only used in group comparison reports (they are not used in matching nor to determine whether or not a row contains missing values during matching).

Data Label Variable

The values in this variable contain text (or numbers) and are used to identify each row. This variable is optional.

Storage Variable

Store Match Numbers In

Specify a variable to store the match number assignments for each row. This variable is optional. If no storage variable is specified, the match numbers will not be stored in the database, but matching reports can still be generated.

Optimization Algorithm Options

Maximum Iterations

Specify the number of optimization iterations to perform before exiting. You may choose a value from the list, or enter your own. This option is available in order to avoid an infinite loop. We have found that as the number of Matches per Treatment increases, it takes more and more iterations in order to arrive at a solution.

Matching Options

Distance Calculation Method

Specify the method to be used in calculating distances between treatment and control subjects. If the distance method involves propensity score calipers, then a Propensity Score Variable must

also be specified. Eight different distance measures are available in NCSS. For the formulas that follow, we will adopt the following notation:

1. The subscript i refers to the i^{th} treatment subject.
2. The subscript j refers to the j^{th} control subject.
3. $d(i,j)$ is the estimated distance between subjects i and j.
4. $\mathbf{x}$ is the vector of observed covariates used to estimate the propensity score.
5. $q(\mathbf{x}) = e(\mathbf{x})$ is the propensity score based on the covariates $\mathbf{x}$. If the logit transformation is used in the analysis, then $q(\mathbf{x}) = \log((1-e(\mathbf{x}))/e(\mathbf{x}))$.
6. $\mathbf{y}$ is the vector of observed covariates used in the distance calculation. $\mathbf{y}$ is not necessary equivalent to $\mathbf{x}$, although it could be.
7. $\mathbf{u} = (\mathbf{y}, q(\mathbf{x}))$ is the vector of observed covariates and the propensity score (or logit propensity score).
8. C is the sample covariance matrix of the matching variables (including the propensity score) from the full set of control subjects.
9. c is the caliper radius. The width of each caliper is $2c$.
10. $F\mathrm{M}_{i,l}$ and $F\mathrm{M}_{j,l}$ are the values of the l^{th} forced match variable for subjects i and j, respectively. If no forced match variables are specified, then $F\mathrm{M}_{i,l} = F\mathrm{M}_{j,l}$ for all l.
11. $R_{i,p}$ and $R_{j,p}$ are the ranks of the p^{th} covariate values or propensity score for subjects i and j, respectively. Average ranks are used in the case of ties.

The options are:

- **Mahalanobis Distance within Propensity Score Calipers (no matches outside calipers)**

$$d(i,j) = \begin{cases} (\mathbf{u}_i - \mathbf{u}_j)^T C^{-1} (\mathbf{u}_i - \mathbf{u}_j) & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \le c \text{ and } F\mathrm{M}_{i,l} = F\mathrm{M}_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Mahalanobis Distance within Propensity Score Calipers (matches allowed outside calipers)**

$$d(i,j) = \begin{cases} (\mathbf{u}_i - \mathbf{u}_j)^T C^{-1} (\mathbf{u}_i - \mathbf{u}_j) & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \le c \text{ and } F\mathrm{M}_{i,l} = F\mathrm{M}_{j,l} \text{ for all } l \\ |q(\mathbf{x}_i) - q(\mathbf{x}_j)| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| > c \text{ for all unmatched } j \\ & \quad \text{and } F\mathrm{M}_{i,l} = F\mathrm{M}_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

The absolute difference, $|q(\mathbf{x}_i) - q(\mathbf{x}_j)|$, is only used in assigning matches if there are no available controls for which $|q(\mathbf{x}_i) - q(\mathbf{x}_j)| \le c$.

- **Mahalanobis Distance including the Propensity Score (if specified)**

$$d(i,j) = \begin{cases} (\mathbf{u}_i - \mathbf{u}_j)^T C^{-1} (\mathbf{u}_i - \mathbf{u}_j) & \text{if } F\mathrm{M}_{i,l} = F\mathrm{M}_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Propensity Score Difference within Propensity Score Calipers (no matches outside calipers)**

$$d(i,j) = \begin{cases} |q(\mathbf{x}_i) - q(\mathbf{x}_j)| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \le c \text{ and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Propensity Score Difference**

$$d(i,j) = \begin{cases} |q(\mathbf{x}_i) - q(\mathbf{x}_j)| & \text{if } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Sum of Rank Differences within Propensity Score Calipers (no matches outside calipers)**

$$d(i,j) = \begin{cases} \sum_p |R_{i,p} - R_{j,p}| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \le c \text{ and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Sum of Rank Differences within Propensity Score Calipers (matches allowed outside calipers)**

$$d(i,j) = \begin{cases} \sum_p |R_{i,p} - R_{j,p}| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \le c \text{ and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ |q(\mathbf{x}_i) - q(\mathbf{x}_j)| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| > c \text{ for all unmatched } j \\ & \quad \text{and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

  The absolute difference, $|q(\mathbf{x}_i) - q(\mathbf{x}_j)|$, is only used in assigning matches if there are no available controls for which $|q(\mathbf{x}_i) - q(\mathbf{x}_j)| \le c$.

- **Sum of Rank Differences including the Propensity Score (if specified)**

$$d(i,j) = \begin{cases} \sum_p |R_{i,p} - R_{j,p}| & \text{if } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

In the Greedy Data Matching procedure, two distance calculation methods are available that are not in the Optimal Data Matching procedure (option #2 and option #7). Both involve caliper matching with matches allowed outside calipers. When matches are allowed outside calipers, the algorithm always tries to find matches inside the calipers first, and only assigns matches outside calipers if a match was not found inside. Matches outside calipers are created based solely on the propensity score, i.e., if matches outside calipers are allowed and no available control subject exists that is within c propensity score units of a treatment subject, then the control subject with the nearest propensity score is matched with the treatment. This type of matching algorithm is described in Rosenbaum and Rubin (1985a).

Matches per Treatment

Choose the number of controls to match with each treatment. You may choose one of the values for the list or enter an integer value of your own. For greedy matching, the value you enter can be no larger than controls/treatments rounded up to the next highest integer. When the number of matches per treatment is greater than one, the greedy algorithm finds the best match (if possible) for each treatment before returning and creating the second match, third match, etc. For optimal matching, the value can be no larger than controls/treatments rounded down to the next lowest integer. The options are:

- **Variable (Full Matching) (Optimal Data Matching Only)**

  This option causes the optimal matching algorithm to match a variable number of controls to each treatment. Each control is used only once, and each treatment is matched with at least one control. All eligible controls (e.g. all controls where at least one treatment-control distance is non-infinite) are matched.

- **Maximum Possible**

  This option causes the program to assign the maximum number (k) of matches that can be made between treatments and controls. If greedy matching is used and controls/treatments is not an integer, then using this option will result in incomplete pair-matching.

- **Integer Values**

  If an integer value is entered or selected, then the program attempts to create the specified number of control matches for each treatment.

Order for Matching

This option specifies the order in which subjects are entered into the matching algorithm. In the case of tied distance values, the matches created depend on the order in which the treatment and control subjects are considered. The options are:

- **Random**

  Both treatment and control subjects are randomly ordered before entering into the matching algorithm. When the number of matches per treatment is greater than one, the greedy algorithm finds the best match (if possible) for each treatment before returning and creating the second match, third match, etc. It is likely that match assignments will change from run-to-run when using random ordering.

- **Sorted by Distance (Greedy Data Matching Only)**

  This option causes the program to sort the matrix of all pair-wise treatment-control distances, and assign matches starting with the smallest distance and working toward the largest until all treatments have been matched with the specified number of controls.

- **Sorted by Row Number**

  Both treatment and control subjects are entered into the matching algorithms according to their location in the database. When the number of matches per treatment is greater than one, the greedy algorithm finds the best match (if possible) for each treatment before returning and creating the second match, third match, etc.

Caliper Radius

This option specifies the caliper radius, c, to be used in caliper matching. The caliper radius is calculated as

$$c = a\sqrt{(\sigma_1^2 + \sigma_2^2)/2} = a \times SIGMA$$

where a is a user-specified coefficient, σ_1^2 is the sample variance of $q(\mathbf{x})$ for the treatment group, and σ_2^2 is the sample variance of $q(\mathbf{x})$ for the control group. NCSS allows you to enter the caliper radius using the syntax “a*SIGMA”, where you specify the value for a (e.g. “0.2*SIGMA”) or by entering the actual value directly for c (e.g. “0.5”). In the case of the former, the program calculates the variances of the treatment and control group propensity scores

for you. You may want to run descriptive statistics on the treatment and control group propensity scores to determine the variance ratio of your data in order to find the appropriate value of a (from the table above) for your research objectives.

Reports Tab

The following options control the format of the reports that are displayed.

Select Reports

Data Summary Report ... Matching Detail Report

Indicate whether to display the indicated reports.

Incomplete Matching Report (Greedy Data Matching Only)

Indicate whether to display the incomplete matching report that lists the treatments that were not paired with the specified number of controls.

Report Options

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimals

Propensity Scores/Covariates … Standardized Differences

Specify the number of digits after the decimal point to be displayed for output values of the type indicated.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Optimal (1:1) Matching using the Mahalanobis Distance within Propensity Score Calipers

This tutorial describes how to create 1:1 treatment-control matches using the Mahalanobis Distance within Propensity Score Calipers distance metric. The data used in this example are contained in the PROPENSITY database. The propensity scores were created using logistic regression with Exposure as the dependent variable, X1 – Age as numeric independent variables, and Race and Gender as categorical independent variables. The propensity score represents the probability of being exposed given the observed covariate values. The optimal matching algorithm will always produce a complete matched-pair sample.

You may follow along here by making the appropriate entries or load the completed template **Example 1** from the Template tab of the Data Matching – Optimal window.

1 Open the PROPENSITY dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Propensity.s0**.
- Click **Open**.

2 Open the Data Matching - Optimal window.

- On the menus of the NCSS Data window, select **Tools**, then **Data Matching - Optimal**. The Data Matching - Optimal procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Matching - Optimal window, select the **Variables tab**.
- Enter **Exposure** in the **Grouping Variable** box.
- Enter "**Exposed**" (no quotes) in the **Treatment Group** box.
- Enter **Propensity** in the **Propensity Score Variable** box.
- Make sure **Use Logit** is checked.
- Enter **X1-Age** in the **Covariate Variable(s)** box.
- Enter **ID** in the **Data Label Variable** box.
- Enter **C11** in the **Store Match Numbers In** box.
- Enter **1.5*Sigma** in the **Caliper Radius** box.
- Leave all other options at their default values.

4 Specify the reports.

- On the Data Matching - Optimal window, select the **Reports tab**.
- Put a check mark next to **Matching Detail Report**. Leave all other options at their default values.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

The following reports will be generated for both optimal and greedy matching with slight variations depending on the algorithm selected.

Data Summary Report

Data Summary Report

Rows Read	30
Rows with Missing Data	0
Treatment Rows	8
Control Rows	22
----- Data Variables -----	
Grouping Variable	Exposure
- Treatment Group	"Exposed"
- Control Group	"Not Exposed"
Data Label Variable	ID
----- Variables Used in Distance Calculations -----	
Propensity Score Variable	Logit(Propensity)
Covariate Variable 1	X1
Covariate Variable 2	X2
Covariate Variable 3	X3
Covariate Variable 4	X4
Covariate Variable 5	X5
Covariate Variable 6	X6
Covariate Variable 7	Age
----- Storage Variable -----	
Match Number Storage Variable	C11

This report gives a summary of the data and variables used for matching.

Matching Summary Report

Matching Summary Report

Distance Calculation Method	Mahalanobis Distance within Propensity Score Calipers (no matches outside calipers)
Caliper Radius	2.63288
Order for Matching	Random
Controls Matched per Treatment	1
Sum of Match Mahalanobis Distances	53.94887
Average Match Mahalanobis Distance	6.74361

Group	N	Matched	Percent Matched	Unmatched	Percent Unmatched
Exposed	8	8	100.00%	0	0.00%
Not Exposed	22	8	36.36%	14	63.64%

This report gives a summary of the matches created, as well as a summary of the matching parameters used by the matching algorithm.

Distance Calculation Method

This is the method used to calculate distances between treatment and control subjects.

Caliper Radius

This is the caliper radius entered or calculated by the program. This line is only displayed if caliper matching based on propensity scores was used.

Order for Matching

This is the order used in matching as selected on the procedure window.

Controls Matched per Treatment

This is the target number of controls to match with each treatment. This value is specified on the procedure window.

Sum of Match Mahalanobis Distances (Sum of Match Propensity Score Differences or Sum of Match Rank Differences)

This is the sum of Mahalanobis distances, propensity score differences, or rank differences (depending on the distance calculation method selected) for all matched-pairs.

Average Match Mahalanobis Distance (Average Match Propensity Score Difference or Average Match Rank Differences)

This is the average Mahalanobis distances, propensity score difference, or rank difference (depending on the distance calculation method selected) for all matched-pairs. This is calculated as the [Sum of Match Distances (or Differences)]/[Number of Matches Formed].

Group (e.g. Exposure)

This specifies either the treatment or the control group. The title of this column is the Grouping Variable name (or label).

N

This is the number of candidates for matching in each group, i.e. the number of subjects with non-missing values for all matching variables in each group.

Matched (Unmatched)

This is the number of subjects that were matched (unmatched) from each group.

Percent Matched (Percent Unmatched)

This is the percent of subjects that were matched (unmatched) from each group.

Group Comparison Reports

Group Comparison Report for Variable = Logit(Propensity)

Group Type	Exposure	N	Mean	SD	Mean Difference	Standardized Difference (%)
Before Matching	Exposed	8	-0.18344	1.39		
	Not Exposed	22	2.63066	2.06	-2.81410	-160.32%
After Matching	Exposed	8	-0.18344	1.39		
	Not Exposed	8	0.82159	1.33	-1.00503	-73.88%

Group Comparison Report for Variable = X1

Group Type	Exposure	N	Mean	SD	Mean Difference	Standardized Difference (%)
Before Matching	Exposed	8	39.50000	20.96		
	Not Exposed	22	45.90909	26.11	-6.40909	-27.07%
After Matching	Exposed	8	39.50000	20.96		
	Not Exposed	8	26.50000	13.60	13.00000	73.58%

.
.
.

(output reports continue for each covariate variable specified)

This report provides summary statistics by group for the data in the propensity score variable and each covariate variable both before and after matching. Notice that the matching seemed to improve the balance of the propensity scores (Standardized Difference dropped from –160% to –73%) between the treatment and control groups, but worsened the balance for the covariate X1 (Standardized Difference increased from –27% to 73.58%).

Group Type

This specifies whether the summary statistics refer to groups before or after matching.

Group (e.g. Exposure)

This specifies either the treatment or the control group. The title of this column is the grouping variable name (or label).

N

This is the number of non-missing values in each variable by group. If there are missing values in covariates that were not used for matching, then these numbers may be different from the total number of subjects in each group.

Mean

This is the average value for each variable by group.

SD

This is the standard deviation for each variable by group.

Mean Difference

This is the difference between the mean of the treatment group and the mean of the control group.

Standardized Difference (%)

The standardized difference can be used to measure the balance between the treatment and control groups before and after matching. If a variable is balanced, then the standardized difference should be close to zero. The standardized difference is the mean difference as a percentage of the average standard deviation

$$\text{Standardized Difference}\,(\%) = \frac{100(\bar{x}_{t,p} - \bar{x}_{c,p})}{\sqrt{(s^2_{t,p} - s^2_{c,p})/2}}$$

where $\bar{x}_{t,p}$ and $\bar{x}_{c,p}$ are the treatment and control group means for the p^{th} covariate variable, respectively, and $s^2_{t,p}$ and $s^2_{c,p}$ are the treatment and control group sample variances for the p^{th} covariate variable, respectively.

Matching Detail Report

Matching Detail Report
Treatment = "Exposed", Control = "Not Exposed"

		Treatment			Matched Control		
Match Number	Mahalanobis Distance	Row	Logit Propensity	ID	Row	Logit Propensity	ID
1	4.32807	1	-1.05541	A	8	0.30221	H
2	5.05385	4	-0.34801	D	22	-1.28232	V
3	9.07686	6	2.50671	F	16	3.28652	P
4	3.99318	10	0.05650	J	24	1.73357	X
5	13.85904	14	-1.11718	N	28	-0.07642	BB
6	9.25961	19	-1.31100	S	27	0.85319	AA
7	5.06011	26	1.16584	Z	29	0.72590	CC
8	3.31815	30	-1.36499	DD	9	1.03004	I

This report provides a list of all matches created and important information about each match.

Match

This is the match number assigned by the program to each match and stored to the database (if a storage variable was specified).

Mahalanobis Distance (Propensity Score |Difference| or Sum of Rank |Differences|)

This is the estimated distance between the treatment and matched control. The column title depends on the distance calculation method selected.

Row

This is the row of the treatment or control subject in the database.

Propensity Score (or first covariate variable)

This is the value of the propensity score (or logit propensity score if ‘Use Logit’ was selected). If no propensity score variable was used in distance calculations, then this is the value of first covariate variable specified. The title of this column is based on the propensity score variable name (or label) or the first covariate variable name (or label).

Data Label (e.g. ID)

This is the identification label of the row in the database. The title of this column is the data label variable name (or label).

Example 2 – Greedy (1:2) Matching using the Propensity Score Difference with Forced Match Variables

Continuing with Example 1, we will now use the greedy matching algorithm to create matches while using race and gender as a forced match variables. This will force the algorithm to find control matches for treatments where the gender and race match exactly, i.e., a male can only be matched with a male, and a female can only be matched with a female, etc. Please note that the optimal matching algorithm can also be used with forced match variables, but we use the greedy matching algorithm here to display the incomplete matched-pair sample that results.

You may follow along here by making the appropriate entries or load the completed template **Example 2** from the Template tab of the Data Matching – Greedy window.

1 Open the PROPENSITY dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Propensity.s0**.
- Click **Open**.

2 Open the Data Matching - Greedy window.

- On the menus of the NCSS Data window, select **Tools**, then **Data Matching - Greedy**. The Data Matching - Greedy procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Matching - Greedy window, select the **Variables tab**.
- Enter **Exposure** in the **Grouping Variable** box.
- Enter "**Exposed**" (no quotes) in the **Treatment Group** box.
- Enter **Propensity** in the **Propensity Score Variable** box.
- Make sure **Use Logit** is checked.
- Enter **Race-Gender** in the **Forced Match Variable(s)** box.
- Enter **X1-Age** in the **Covariate Variable(s)** box.
- Enter **ID** in the **Data Label Variable** box.
- Enter **C11** in the **Store Match Numbers In** box.
- Choose **Propensity Score Difference** in the **Distance Calculation Method** box.
- Enter **2** in the **Matches per Treatment** box.
- Leave all other options at their default values.

4 Specify the reports.

- On the Data Matching - Greedy window, select the **Reports tab**.
- Put a check mark next to **Matching Detail Report** and **Incomplete Matching Report**. Leave all other options at their default values.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Greedy Data Matching Output

Data Summary Report

Rows Read	30
Rows with Missing Data	0
Treatment Rows	8
Control Rows	22
----- Data Variables -----	
Grouping Variable	Exposure
- Treatment Group	"Exposed"
- Control Group	"Not Exposed"
Data Label Variable	ID
----- Variables Used in Distance Calculations -----	
Propensity Score Variable	Logit(Propensity)
Forced Match Variable 1	Race
Forced Match Variable 2	Gender
----- Storage Variable -----	
Match Number Storage Variable	C11

Matching Summary Report

Distance Calculation Method	Propensity Score Difference
Order for Matching	Sorted by Distance
Controls Matched per Treatment	2
Sum of Match Propensity Score Differences	14.63954
Average Match Propensity Score Difference	1.46395

Exposure	N	Matched	Percent Matched	Unmatched	Percent Unmatched
Exposed	8	6	75.00%	2	25.00%
Not Exposed	22	10	45.45%	12	54.55%

Group Comparison Report for Variable = Logit(Propensity)

Group Type	Exposure	N	Mean	SD	Mean Difference	Standardized Difference (%)
Before Matching	Exposed	8	-0.18344	1.39		
	Not Exposed	22	2.63066	2.06	-2.81410	-160.32%
After Matching	Exposed	6	0.11751	1.50		
	Not Exposed	10	1.48046	1.55	-1.36296	-89.21%

Group Comparison Report for Variable = X1

Group Type	Exposure	N	Mean	SD	Mean Difference	Standardized Difference (%)
Before Matching	Exposed	8	39.50000	20.96		
	Not Exposed	22	45.90909	26.11	-6.40909	-27.07%
After Matching	Exposed	6	33.66667	20.79		
	Not Exposed	10	44.90000	30.57	-11.23333	-42.97%

.
.
.

(output reports continue for each covariate variable specified)

Matching Detail Report
Treatment = "Exposed", Control = "Not Exposed"

Match Number	Logit Propensity |Difference|	Treatment Row	Treatment Logit Propensity	Treatment ID	Matched Control Row	Matched Control Logit Propensity	Matched Control ID
1	1.20120	4	-0.34801	D	27	0.85319	AA
1	1.37805	4	-0.34801	D	9	1.03004	I
2	0.15966	6	2.50671	F	15	2.34705	O
2	0.56240	6	2.50671	F	11	1.94431	K
3	0.66941	10	0.05650	J	29	0.72590	CC
3	4.46221	10	0.05650	J	2	4.51870	B
4	1.61321	19	-1.31100	S	8	0.30221	H
4	3.92267	19	-1.31100	S	12	2.61167	L
5	0.58806	26	1.16584	Z	23	1.75390	W
6	0.08266	30	-1.36499	DD	22	-1.28232	V

Notice that only the propensity score variable was used in distance calculations, but group comparison reports were generated for each covariate variable specified. In the Matching Detail Report, you can see that not all treatments were matched (incomplete matching). Finally, notice that race and gender were both used as Forced Match variables.

If you go back to the spreadsheet and sort the data on C11 (click on **Data** > **Sort** from the ***NCSS* Home** window), you will notice that matches were only created where the race and gender were identical for both the treatment and control.

	Exposure	ID	X1	X2	X3	X4	X5	X6	Age	Race	Gender	Propensity	C11
11	Not Exposed	U	64	48	48	59	56	50	60	Hispanic	Female	0.05299638	
12	Not Exposed	X	13	25	35	8	6	9	52	Hispanic	Female	0.15013131	
13	Not Exposed	Y	30	100	105	53	60	94	72	Black	Female	0.00267398	
14	Not Exposed	BB	20	36	44	17	16	22	33	Black	Female	0.51909545	
15	Exposed	D	31	81	86	46	50	74	33	Hispanic	Female	0.58613607	1
16	Not Exposed	I	46	36	40	39	36	32	39	Hispanic	Female	0.26307673	1
17	Not Exposed	AA	31	22	28	21	17	11	60	Hispanic	Female	0.2987649	1
18	Exposed	F	22	30	39	17	15	17	29	Black	Female	0.07538899	2
19	Not Exposed	K	84	86	82	90	91	97	24	Black	Female	0.12517537	2
20	Not Exposed	O	82	79	75	86	86	89	36	Black	Female	0.08730072	2
21	Not Exposed	B	4	2	5	11	11	5	71	Hispanic	Male	0.01078557	3
22	Exposed	J	3	38	49	5	6	19	58	Hispanic	Male	0.48587995	3
23	Not Exposed	CC	15	35	44	14	13	20	65	Hispanic	Male	0.32609454	3
24	Not Exposed	H	31	91	96	50	56	85	52	Caucasian	Male	0.4250166	4
25	Not Exposed	L	73	88	86	83	85	96	63	Caucasian	Male	0.06839106	4
26	Exposed	S	64	36	36	53	49	37	47	Caucasian	Male	0.78768042	4
27	Not Exposed	W	71	5	5	45	37	7	53	Caucasian	Female	0.14755618	5
28	Exposed	Z	36	67	71	44	46	61	62	Caucasian	Female	0.23760766	5
29	Not Exposed	V	12	63	72	23	26	48	29	Caucasian	Female	0.78284523	6
30	Exposed	DD	46	60	63	49	49	57	28	Caucasian	Female	0.79656871	6
31													

Incomplete Matching Report

Incomplete Matching Report
Exposure = "Exposed"

Treatment Row	Matches (Target = 2)	Logit Propensity	ID
1	0	-1.05541	A
14	0	-1.11718	N
26	1	1.16584	Z
30	1	-1.36499	DD

This report lists the treatments that were not paired with the target number of controls (2 in this case). Rows 1 and 14 were not paired with any controls. Rows 26 and 30 were only paired with 1 control. All other treatment rows were paired with 2 treatments. Incomplete matching is usually due to the use of forced match variables, using caliper matching, or setting Matches per Treatment to 'Maximum Possible'.

Treatment Row

This is the row in the database containing the treatment subject that was not fully matched.

Matches (Target = k)

This is the number of matches that were found for each treatment. The target represents the number of Matches per Treatment specified on the input window.

Propensity Score (or first covariate variable)

This is the value of the propensity score (or logit propensity score if 'Use Logit' was selected) for the incompletely-matched treatment. If no propensity score variable was used in distance calculations, then this is the value of first covariate variable specified. The title of this column is based on the propensity score variable name (or label) or the first covariate variable name (or label).

Data Label (e.g. ID)

This is the identification label of the incompletely-matched row in the database. The title of this column is the data label variable name (or label).

Example 3 – Matching on Forced Match Variables Only

Continuing with Example 2, suppose we wanted to form matches based solely on forced match variables, i.e., we want the matches to have exactly the same values for each covariate. We could enter all of the covariates in as forced match variables, but with a database as small as we are using, we are unlikely to find any matches. We will use the greedy data matching procedure to illustrate how you can assign matches based on the gender and race forced match variables only. Random ordering is used to ensure that the treatments are randomly paired with controls (where the forced match variable values match).

In order to complete this task, you must first create a new column in the database filled with 1's. You can do this by clicking on the first cell in an empty column and selecting **Edit** > **Fill** from the ***NCSS* Home** window (for **Fill Value(s)** enter **1**, for **Increment** enter **0**, and click **OK**). A column of ones has already been created for you in the PROPENSITY dataset. This column of ones is necessary because the matching procedure requires either a propensity score variable or a covariate variable to run.

You may follow along here by making the appropriate entries or load the completed template **Example 3** from the Template tab of the Data Matching – Greedy window.

1 Open the PROPENSITY dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Propensity.s0**.
- Click **Open**.

2 Open the Data Matching - Greedy window.

- On the menus of the NCSS Data window, select **Tools**, then **Data Matching - Greedy**. The Data Matching - Greedy procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Matching - Greedy window, select the **Variables tab**.
- Enter **Exposure** in the **Grouping Variable** box.
- Enter "**Exposed**" (no quotes) in the **Treatment Group** box.
- Enter **Ones** (or the name of your variable containing all 1's) in the **Propensity Score Variable** box.
- Make sure **Use Logit** is **unchecked**.
- Enter **Race-Gender** in the **Forced Match Variable(s)** box.
- Enter **ID** in the **Data Label Variable** box.
- Enter **C11** in the **Store Match Numbers In** box.
- Choose **Propensity Score Difference** in the **Distance Calculation Method** box.
- Enter **2** in the **Matches per Treatment** box.
- Choose **Random** in the **Order for Matching** box.
- Leave all other options at their default values.

4 Specify the reports.

- On the Data Matching - Greedy window, select the **Reports tab**.
- Put a check mark next to **Matching Detail Report** and **Incomplete Matching Report**. Leave all other options at their default values.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Matching Reports

Matching Detail Report
Treatment = "Exposed", Control = "Not Exposed"

		Treatment			Matched Control		
Match Number	Logit Propensity |Difference|	Row	Logit Propensity	ID	Row	Logit Propensity	ID
1	0.00000	1	1.00000	A	29	1.00000	CC
2	0.00000	4	1.00000	D	13	1.00000	M
2	0.00000	4	1.00000	D	20	1.00000	T
3	0.00000	6	1.00000	F	15	1.00000	O
3	0.00000	6	1.00000	F	11	1.00000	K
4	0.00000	10	1.00000	J	2	1.00000	B
5	0.00000	19	1.00000	S	12	1.00000	L
5	0.00000	19	1.00000	S	8	1.00000	H
6	0.00000	26	1.00000	Z	23	1.00000	W
7	0.00000	30	1.00000	DD	22	1.00000	V

Incomplete Matching Report
Exposure = "Exposed"

Treatment Row	Matches (Target = 2)	Logit Propensity	ID
1	1	1.00000	A
10	1	1.00000	J
14	0	1.00000	N
26	1	1.00000	Z
30	1	1.00000	DD

The matching detail report is not very informative because all of the propensity scores are equal to 1. If you run the procedure several times, you will notice that the controls are randomly pairing with the treatments when the race and gender are the same. Your report may be slightly different from this report because random ordering was used. If you sort on C11, you will see that all matched pairs have the same value for race and gender.

13	Exposed	Z	36	67	71	44	46	61	62	Caucasian	Female	0.23760766	
14	Exposed	A	50	102	103	70	75	102	45	Hispanic	Male	0.74181165	1
15	Not Exposed	CC	15	35	44	14	13	20	65	Hispanic	Male	0.32609454	1
16	Exposed	N	64	1	2	38	29	0	39	Caucasian	Female	0.75346448	2
17	Not Exposed	W	71	5	5	45	37	7	53	Caucasian	Female	0.14755618	2
18	Exposed	D	31	81	86	46	50	74	33	Hispanic	Female	0.58613607	3
19	Not Exposed	I	46	36	40	39	36	32	39	Hispanic	Female	0.26307673	3
20	Not Exposed	AA	31	22	28	21	17	11	60	Hispanic	Female	0.2987649	3

Example 4 – Validation of the Optimal Data Matching Algorithm using Rosenbaum (1989)

Rosenbaum (1989) provides an example of both optimal and greedy matching using a well-known dataset from Cox and Snell (1981), which involves 26 U.S. light water nuclear power plants (six "partial turnkey" plants are excluded in the analysis). Seven of the plants were constructed on sites where a light water reactor had existed previously; these are the treatments. The 19 remaining plants serve as the controls. The sum of rank differences was used to calculate distances between treatment and control plants. Two covariate variables were used in the analysis: the date the construction permit was issued (Date), and the capacity of the plant (Capacity). Site was used as the grouping variable with "Existing" as the treatment group. Rosenbaum (1989) reports the following optimal pairings by plant number (treatment, control):

(3,2), (3,21), (5,4), (5,7), (9,7), (9,10), (18,8), (18,13), (20,14), (20,15), (22,17), (22,26), (24,23), (24,25)

The data used in this example are contained in the COXSNELL database.

You may follow along here by making the appropriate entries or load the completed template **Example 4** from the Template tab of the Data Matching – Optimal window.

1 Open the COXSNELL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **CoxSnell.s0**.
- Click **Open**.

2 Open the Data Matching - Optimal window.

- On the menus of the NCSS Data window, select **Tools**, then **Data Matching - Optimal**. The Data Matching - Optimal procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Matching - Optimal window, select the **Variables tab**.
- Enter **Site** in the **Grouping Variable** box.
- Enter "**Existing**" (no quotes) in the **Treatment Group** box.
- Make sure nothing is entered in the **Propensity Score Variable** box.
- Enter **Date-Capacity** in the **Covariate Variable(s)** box.
- Enter **Plant** in the **Data Label Variable** box.
- Choose **Sum of Rank Differences including the Propensity Score (if specified)** in the **Distance Calculation Method** box.
- Enter **2** in the **Matches per Treatment** box.
- Leave all other options at their default values.

4 Specify the reports.

- On the Data Matching - Optimal window, select the **Reports tab**.
- Put a check mark next to **Matching Detail Report**. Leave all other options at their default values.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Matching Reports

Matching Summary Report

Distance Calculation Method	Sum of Rank Differences including the Propensity Score
Order for Matching	Random
Controls Matched per Treatment	2
Sum of Match Rank Differences	74.00000
Average Match Rank Difference	5.28571

Site	N	Matched	Percent Matched	Unmatched	Percent Unmatched
Existing	7	7	100.00%	0	0.00%
New	19	14	73.68%	5	26.32%

Matching Detail Report
Treatment = "Existing", Control = "New"

		Treatment			Matched Control		
Match Number	Sum of Rank |Differences|	Row	Date	Plant	Row	Date	Plant
1	18.50000	1	2.33000	3	23	3.75000	21
1	0.00000	1	2.33000	3	9	2.33000	2
2	0.00000	2	3.00000	5	10	3.00000	4
2	10.50000	2	3.00000	5	12	3.17000	7
3	5.50000	3	3.42000	9	20	3.42000	16
3	5.50000	3	3.42000	9	14	3.33000	10
4	0.00000	4	3.42000	18	17	3.42000	13
4	2.50000	4	3.42000	18	13	3.42000	8
5	0.00000	5	3.92000	20	18	3.92000	14
5	2.50000	5	3.92000	20	19	3.92000	15
6	12.00000	6	5.92000	22	26	6.08000	26
6	5.00000	6	5.92000	22	21	4.50000	17
7	4.00000	7	5.08000	24	24	4.67000	23
7	8.00000	7	5.08000	24	25	5.42000	25

The optimal match-pairings found by ***NCSS*** match those in Rosenbaum (1989) exactly. Notice, however, that the distances (Sum of Rank |Differences|) are slightly different in some instances from those given in Table 1 of the article. This is due to the fact that Rosenbaum (1989) rounds all non-integer distances in their reports. This rounding also affects the overall sum of match rank differences; ***NCSS*** calculates the overall sum as 74, while Rosenbaum (1989) calculates the overall sum as 71, with the difference due to rounding.

Example 5 – Validation of the Greedy Data Matching Algorithm using Rosenbaum (1989)

Continuing with Example 4, Rosenbaum (1989) also reports the results from the greedy matching algorithm, where the order for matching is sorted by distance. The article reports the following greedy pairings by plant number (treatment, control):

(3,2), (3,19), (5,4), (5,21), (9,10), (9,7), (18,8), (18,13), (20,14), (20,15), (22,17), (22,26), (24,23), (24,25)

You may follow along here by making the appropriate entries or load the completed template **Example 5** from the Template tab of the Data Matching – Greedy window.

1 Open the COXSNELL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **CoxSnell.s0**.
- Click **Open**.

2 Open the Data Matching - Greedy window.

- On the menus of the NCSS Data window, select **Tools**, then **Data Matching - Greedy**. The Data Matching - Greedy procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Matching - Greedy window, select the **Variables tab**.
- Enter **Site** in the **Grouping Variable** box.
- Enter "**Existing**" (no quotes) in the **Treatment Group** box.
- Make sure nothing is entered in the **Propensity Score Variable** box.
- Enter **Date-Capacity** in the **Covariate Variable(s)** box.
- Enter **Plant** in the **Data Label Variable** box.
- Choose **Sum of Rank Differences including the Propensity Score (if specified)** in the **Distance Calculation Method** box.
- Enter **2** in the **Matches per Treatment** box.
- Leave all other options at their default values.

4 Specify the reports.

- On the Data Matching - Greedy window, select the **Reports tab**.
- Put a check mark next to **Matching Detail Report**. Leave all other options at their default values.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Output

Matching Summary Report

Distance Calculation Method	Sum of Rank Differences including the Propensity Score
Order for Matching	Sorted by Distance
Controls Matched per Treatment	2
Sum of Match Rank Differences	80.00000
Average Match Rank Difference	5.71429

Site	N	Matched	Percent Matched	Unmatched	Percent Unmatched
Existing	7	7	100.00%	0	0.00%
New	19	14	73.68%	5	26.32%

Matching Detail Report

Treatment = "Existing", Control = "New"

		Treatment			Matched Control		
Match Number	Sum of Rank |Differences|	Row	Date	Plant	Row	Date	Plant
1	0.00000	1	2.33000	3	9	2.33000	2
1	21.00000	1	2.33000	3	22	4.17000	19
2	0.00000	2	3.00000	5	10	3.00000	4
2	15.50000	2	3.00000	5	23	3.75000	21
3	4.00000	3	3.42000	9	12	3.17000	7
3	5.50000	3	3.42000	9	14	3.33000	10
4	0.00000	4	3.42000	18	17	3.42000	13
4	2.50000	4	3.42000	18	13	3.42000	8
5	0.00000	5	3.92000	20	18	3.92000	14
5	2.50000	5	3.92000	20	19	3.92000	15
6	5.00000	6	5.92000	22	21	4.50000	17
6	12.00000	6	5.92000	22	26	6.08000	26
7	4.00000	7	5.08000	24	24	4.67000	23
7	8.00000	7	5.08000	24	25	5.42000	25

The greedy match-pairings found by ***NCSS*** match those in Rosenbaum (1989) exactly. Again, some of the distances are different from those in Table 1 of the article because of rounding. ***NCSS*** calculates the overall sum of rank differences as 80, while Rosenbaum (1989) calculates the overall sum as 79 with the difference due to rounding.

Chapter 124

Data Stratification

Introduction

This procedure is used to create stratum assignments based on quantiles from a numeric stratification variable. The user is able to choose the number of strata to create and the amount of data used in the quantile calculations. Stratification is commonly used in the analysis of data from observational studies where covariates are not controlled. This procedure is based on the results given in D'Agostino, R.B., Jr. (2004), chapter 1.2.

Observational Studies

In observational studies, investigators do not control the assignment of treatments to subjects. Consequently, a difference in covariates may exist among treatment groups. Stratification (or subclassification) is often used to control for these differences in background characteristics. Strata are created by dividing subjects into groups based on observed covariates. However, as the number of covariates increases, the number of required strata grows exponentially. Propensity scores, defined as the conditional probability of treatment given a set of covariates, can be used in this situation to account for the presence of uncontrollable covariate factors. Stratification on the propensity score alone can balance the distributions of covariates among groups without the exponential increase in the number of strata. Rosenbaum and Rubin (1984) suggest that the use of five strata often removes 90% or more of the bias in each of the covariates used in the calculation of the propensity score. The propensity score is usually calculated using logistic regression or discriminant analysis with the treatment variable as the dependent (group) variable and the background covariates as the independent variables. For further information about propensity scores, their calculation, and uses, we refer you to the chapter entitled “Data Matching for Observational Studies” in this manual, or chapter 1.2 (pages 67 - 83) of D'Agostino, R.B., Jr. (2004). For more information about logistic regression or discriminant analysis, see the corresponding chapters in the ***NCSS*** manuals.

Data Structure

The data values for stratification must be entered in a single variable (column). Only numeric values are allowed. Missing values are represented by blanks. Text values are treated as missing values. Optional data label and grouping variables may also be used, with each variable representing a single column in the data file. The following is a subset of the PROPENSITY dataset, which will be used in the tutorials that follow.

PROPENSITY dataset (subset)

ID	Exposure	X1	...	Age	Race	Gender	Propensity
A	Exposed	50	...	45	Hispanic	Male	0.7418116515
B	Not Exposed	4	...	71	Hispanic	Male	0.01078557025
C	Not Exposed	81	...	70	Caucasian	Male	0.0008716385678
D	Exposed	31	...	33	Hispanic	Female	0.5861360724
E	Not Exposed	65	...	38	Black	Male	0.1174339761
F	Exposed	22	...	29	Black	Female	0.07538899371
G	Not Exposed	36	...	57	Black	Female	0.008287371892
H	Not Exposed	31	...	52	Caucasian	Male	0.4250166047
I	Not Exposed	46	...	39	Hispanic	Female	0.2630767334
J	Exposed	3	...	58	Hispanic	Male	0.4858799526
K	Not Exposed	84	...	24	Black	Female	0.1251753736

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables to be analyzed.

Data Variables

Data Stratification Variable

Specify the variable that contains the numeric data to be used for stratification. In observational studies, propensity scores are commonly used for stratification. Propensity scores are often obtained using logistic regression or discriminant analysis. This variable is required. Only numeric values are analyzed. Text values are treated as missing values in the reports.

Data Label Variable

The values in this variable contain text (or numbers) and are used to identify each row. This variable is optional.

Storage Variable

Store Stratum Numbers In

Specify a variable to store the stratum number assignments for each row. This variable is optional.

Options

Number of Strata

Specify the number of strata to create. The number of strata must be less than the number of rows with non-missing data in the database (or quantile calculation group). Rosenbaum and Rubin (1984) suggest that the use of five strata often removes 90% or more of the bias in each of the covariates used in the calculation of the propensity score.

Calculate Quantiles Using

Select the data that will be used in quantile calculations for stratification. All rows with non-missing data will be assigned to a stratum based on the calculated quantiles. The options are:

- **All Data**

  Use all data for quantile calculations.

- **Data from Quantile Calculation Group**

  Use only the data from the Quantile Calculation Group in quantile calculations. A Grouping Variable and a Quantile Calculation Group must also be specified.

Options – Group Options

Grouping Variable

Specify the variable that contains the quantile calculation group information. The response variable that was used in logistic regression or discriminant analysis to produce the propensity scores is often used as the grouping variable. This variable is only used if Calculate Quantiles Using is set to 'Data from Quantile Calculation Group'. The Quantile Calculation Group must also be specified.

Quantile Calculation Group

Specify the group that is to be used in quantile calculations. The propensity scores in this group only will be used to calculate the quantiles for stratification of the entire database. This option is only used if Determine Quantiles Using is set to 'Data from Quantile Calculation Group'. The Grouping Variable must also be specified.

Reports Tab

The following options control the format of the reports that are displayed.

Select Reports

Run Summary Report ... Strata Detail Report - Sorted by Stratum

Indicate whether to display the indicated reports.

Report Options

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimals

Quantiles and Data Values

Specify the number of digits after the decimal point to be displayed on output values of the type indicated.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating Strata Assignments

This section presents an example of how to create a column of stratum assignment numbers from a set of propensity scores. The data used in this example are contained in the PROPENSITY database. The propensity scores were created using logistic regression with Exposure as the dependent variable, X1 – Age as numeric independent variables, and Race and Gender as categorical independent variables. The propensity score represent the probability of being exposed given the observed covariate values.

You may follow along here by making the appropriate entries or load the completed template **Example 1** from the Template tab.

1 Open the PROPENSITY dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Propensity.s0**.
- Click **Open**.

2 Open the Data Stratification window.

- On the menus of the NCSS Data window, select **Tools**, then **Data Stratification**. The Data Stratification procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Stratification window, select the **Variables tab**.
- Enter **Propensity** in the **Data Stratification Variable** box.
- Enter **ID** in the **Data Label Variable** box.
- Enter **C11** in the **Store Stratum Numbers In** box.
- Leave all other options at their default values.

4 **Specify the reports.**

- On the Data Stratification window, select the **Reports tab**.
- Put a check mark next to **Strata Detail Report - Sorted by Row** and **Strata Detail Report - Sorted by Stratum**. Leave all other options at their default values.

5 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Run Summary Report

Run Summary Report

Data Stratification Variable	Propensity
Data Label Variable	ID
Stratum Number Storage Variable	C11
Quantiles Calculated Using	All Data
Total Number of Rows Read	30
Rows with Non-Missing Data	30
Rows with Missing Data	0
Rows Used in Quantile Calculations	30
Number of Strata Created	5

This report gives a summary of the variables and parameters used in the creation of the strata.

Quantile Report

Quantile	Value
0.20	0.02939
0.40	0.08015
0.60	0.25289
0.80	0.57273

This report shows the values of the four quantiles necessary to create five strata. The length of this report depends on the number of strata desired.

Quantile

This is the quantile calculated. The number of quantiles required is equal to the number of strata minus one.

Value

This is the value of the q^{th} quantile. The $100q^{th}$ quantile is computed as

$$Z_q = (1-g)X[k_1] + gX[k_2]$$

where

Z_q is the value of the quantile,

q is the fractional value of the quantile (for example, for the 75th quantile, q = .75),

$X[k]$ is the k^{th} observation when the data are sorted from lowest to highest,

k_1 is the integer part of $q(n+1)$,

$k_2 = k_1 + 1$,

g is the fractional part of $q(n+1)$ (for example, if $q(n+1) = 23.42$, then $g = .42$),

n is the total sample size.

Strata Summary Report

Stratum Number	Size	Range
1	6	Propensity <= 0.02939
2	6	0.02939 < Propensity <= 0.08015
3	6	0.08015 < Propensity <= 0.25289
4	6	0.25289 < Propensity <= 0.57273
5	6	Propensity > 0.57273

This report provides a summary of the strata created.

Stratum Number

This is the number assigned to the stratum. These represent the values stored on the database in the stratum storage variable (if specified).

Size

This is the number of rows (or subjects) in each stratum.

Range

This is the propensity score interval associated with each stratum.

Strata Detail Report – Sorted by Row

Strata Detail Report - Sorted by Row

Row	Stratum Number	Propensity	ID
1	5	0.74181	A
2	1	0.01079	B
3	1	0.00087	C
4	5	0.58614	D
5	3	0.11743	E
.	.	.	.
.	.	.	.
.	.	.	.

This report provides a row-by-row list of the assigned stratum numbers, sorted by row.

Row

This is the row on the database.

Stratum Number

This is the number of the stratum to which the observation was assigned. This represents the value stored on the database in the stratum storage variable (if specified).

Data Value (e.g. Propensity)

This is the data value for this row. The title of this column depends on the name (or label) of the Data Stratification Variable.

Data Label (e.g. ID)

This is the data label value for this row. The title of this column depends on the name (or label) of the Data Label Variable.

Strata Detail Report – Sorted by Stratum

Strata Detail Report - Sorted by Stratum

Row	Stratum Number	Propensity	ID
3	1	0.00087	C
25	1	0.00267	Y
13	1	0.00379	M
7	1	0.00829	G
2	1	0.01079	B
17	1	0.02861	Q
20	2	0.03253	T
16	2	0.03604	P
18	2	0.04800	R
21	2	0.05300	U
12	2	0.06839	L
6	2	0.07539	F
15	3	0.08730	O
5	3	0.11743	E
.	.	.	.
.	.	.	.
.	.	.	.

This report provides a row-by-row list of the assigned stratum numbers, sorted by stratum number and data value (e.g. “Propensity”).

Chapter 130

Macros

Introduction

This software has an interactive (point and click) user interface which makes in easy to learn and use. At times, however, it is necessary to repeat the same steps over and over. When this occurs, a batch system becomes more desirable. This chapter documents a batch language that lets you create a macro (script or program) and then run that macro. With the click of a single button, you can have the program run a series of procedures.

We begin with a discussion of how to create, modify, and run a macro. Next, we list all of the macro commands and their function.

Macro Command Center

This section describes how to create a macro, edit it, and run it. This is all accomplished from the Macro Command Center window. You can load this window by selecting *Macros* from the *Tools* menu or right-clicking on the *Macro Button* (green triangle) of the toolbar.

We will now describe each of the objects on the Macro Command Center window.

Active Macro Name

This box displays the name of the currently selected macro file. A preview of this macro is displayed in the Preview window. You can change this name by typing over it or by selecting a different macro from the *Existing Macros* window.

Existing Macros

This box displays a list of all existing macros. Click on a macro in this list to make it the active macro—the macro that is used when one of the control buttons to the right is pressed. Double-clicking a macro causes that macro to open in the window's Notepad program for editing.

The macros are stored in the MACROS subdirectory of the NCSS folder. They always have the file extension '.ncm'.

Preview of Active Macro

This box displays the beginning of the macro that is in the Active Macro Name box.

Record Macro

Pressing this button causes the commands you issue to be written to the macro file named in the *Active Macro* box. The Macro Command Center window will be replaced by a small *Macro* window that lets you stop recording the macro. When the recording starts, if there is a previous macro file with the same name as that in the Active Macro Name box, the previous macro file is erased.

The ***NCSS*** macro recorder does not record every keystroke and click that you make. Instead, it records major operations. For example, suppose you want to include running a t-test in your macro. You would load the t-test, change some options, and run it. The macro recorder saves a copy of the t-test settings as a template file and writes a single command line to the macro file that references this template. All of your settings are included in the template file—there is no reference in the macro to the individual settings changes. This makes the macro much smaller and easier to modify.

Functions from the File and Edit menus are not recorded during a macro recording. Those functions are performed using the SendKeys commands or other specific commands. The Sort, Enter Transform, Recalc Current, and Recalc All functions from the Data menu are not recorded. There are specific commands for these functions. As a general rule, the running of windows for which there is a template tab is recorded during a macro recording.

Edit Macro

Pressing this button causes the Active Macro to be loaded in the Windows NotePad program. This program lets you modify the macro and then save your changes.

Play Macro

Pressing this button causes the Active Macro to be run. Once the macro is finished, the Macro Command Center window will close and you can view the results of running your macro.

Delete Macro

Pressing this button causes the Active Macro to be deleted. The macro file is actually moved to the Recycle Bin from which it can be rescued if you decide it shouldn't have been deleted.

Close

Pressing this button closes the Macro Command Center window.

Select Button Macro

A macro button is displayed in the toolbar of several of the windows (such as the spreadsheet and procedure windows). This button causes the designated macro to be run. The macro that is associated with the macro button is controlled by this section of the Macro Command center. To change the macro that is associated with this button, simply select the desired macro from the *Existing Macros* list and then click the icon. This will associate the macro to the button. This association will remain even if you exit the program.

Syntax of a Macro Command Line

NCSS macros are line based. That is, each macro command expression is written on a separate line. The basic structure of a line is that it begins with a *command* followed by one or more options or parameters of the command, called *arguments*. For example, the following macro opens a dataset and runs the Descriptive Statistics procedure on the first five variables in that dataset.

```
DataOpen "C:/Program Files/NCSS2007/Data/Sample.s0"
LoadProc DescStat
Option DescStat 1 "1:5"
RunProc DescStat
```

In this example, DataOpen, LoadProc, Option, and RunProc are commands, while "C:/Program Files/NCSS2007/Data/Sample.s0", DescStat, 1, and "1:5" are arguments.

Comment Lines

It is often useful to add comment lines to a macro to make it easier to understand later. Comment lines begin with single quotes. When the macro processor encounters a single quote at the beginning of a line, the rest of the line is ignored. Single quotes occurring at a location other than the beginning of a line are treated as text.

Blank lines may also be added to a macro to improve readability. These are also ignored.

Macro Constants and Macro Variables

As stated above, macro commands lines consist of keywords followed by arguments. These arguments are either constants, such as '1' or 'DescStat', or macro variables. These will be discussed next.

Macro Constants

Macro constants fixed values. There are two types of macro constants: text and numeric.

Numeric Constants

Numeric constants are numbers. They may be whole or decimal numbers. They may be positive or negative. They may be enclosed in double quotes, although this is not necessary. When the macro processor expects a number but receives a text value, it sets the numeric value to zero.

Examples of numeric constants are

```
1, 3.14159, and 0.
```

Text Constants

Text constants are usually enclosed in double quotes. If a constant is a single word (made of letters and digits with no blanks or special characters), the double quotes are not necessary.

Examples of text constants are

```
Apple, "Apple Pie", and "D:/Program Files".
```

Macro Variables

Macro Variables are used to store temporary values for use in macro command lines. Some examples of assigning values to macro variables are

```
A# = 4
B# = 4 + 3
File$ = "C:/Program Files/NCSS/Data/ABC.S0"
F$ = "4" & "5"
```

In these examples, A#, B#, File$, and F$ are macro variables. The assigned values for each of the variables are 4, 7, "C:\Program Files\NCSS\Data\ABC.S0", and 45, respectively.

There are two types of macro variables: text and numeric.

Text Macro Variables

Text macro variables are used to hold text values. The rules for naming them are that the names can contain only letters and numbers (no spaces or special characters) and they must end with a '$'. The case of the letters is ignored (so 'A$' is used interchangeably with 'a$').

Examples of text macro variable names are

```
A$
Apple$
FileName$
```

Numeric Macro Variables

Numeric variables are used to hold numeric values. The rules for naming them are that the names can contain only letters and numbers (no spaces or special characters), and they must end with a '#'. The case of the letters is ignored (so 'A#' is used interchangeably with 'a#').

Examples of numeric macro variable names are

```
A#
Apple#
NRows#
```

Assigning Values to Macro Variables

One type of macro variable expression is that of assigning a value to a variable. The basic syntax for this type of expression is

{variable} = {value}

where the {variable} is text or numeric and {value} is a macro variable or (text or numeric) macro constant. If a text value contains spaces or special characters, it must be enclosed in double quotes.

A text value can be is assigned to a numeric macro variable. In this case, the text value is converted to a number. If it cannot be converted to a number (e.g., it is a letter), the numeric macro variable is set to zero.

Following are some examples of valid assignment expressions.

```
A# = 4
X$ = John
File$ = "C:\Program Files\NCSS\Data\ABC.S0"
X$ = File$
X$ = A#
A# = F$
F$ = 4
```

Macro Variable Combination Expressions

Macro variables can be combined using simple mathematical expressions. The basic syntax for this type of expression is

{variable} = {value} {operator} {value}.

The available operators are + (add), - (subtract), * (multiply), / (divide), and & (concatenate).

If a text value is involved in a mathematical expression, it is converted to a numeric value before the mathematical expression is evaluated. If it cannot be converted to a number, the text value is set to the numeric value of zero.

Following are some examples of valid assignment expressions.

Expression	Result
A# = 4 + 3	A# = 7
B# = A# * 2	B# = 14
C$ = "C:/Pgm/"	
D$ = C$ & "ABC.S0"	D$ = "C:/Pgm/ABC.S0"
E# = C$ * 4	E# = 0
F$ = "4" + "5"	F$ = "9"
F$ = "4" & "5"	F$ = "45"
A# = 1	
A# = A# + 1	A# = 2

Macro variable assignments are used while the macro is running, but are not saved when the macro has completed.

Displaying Macro Variables

The value of one or more macro variables may be displayed on the output using the PRINT or HEADING commands.

Print

This command outputs the requested values to the printout.

The syntax of this command is

PRINT {p1} {p2} {p3} …

where

{p1} {p2} ... are assigned macro variables or constants.

Following are some examples of Print commands.

Command	Printed Result
PRINT "Hi World"	Hi World
I# = 1	
J$ = "C:/NCSS/Data/ABC"	
F$ = J$ & i#	
F$ = F$ & ".s0"	
PRINT "File=" F$	File=C:/NCSS/Data/ABC1.s0
PRINT "1" "2" "3" "4"	1 2 3 4

Heading

This command adds a line to the page heading that is shown at the top of each page.

The syntax of this command is

HEADING {h1}

where

{h1} is an assigned macro variable or constant.

Following are some examples of valid HEADING commands.

Command	Heading
HEADING "Hi World"	Hi World
F$ = "Heart Study"	
HEADING F$	Heart Study

Logic and Control Commands

The lines in a macro are processed in succession. These commands allow you to alter the order in which macro lines are processed, allow user-input, or end the program.

List of Logic and Control Commands:

Flag
GOTO
IF
INPUT
END
SendKeys

Flag Statement

A flag is a reference point in the program. The GOTO command sends macro line control to a specific flag. A flag is made up of letters and numbers (no spaces) followed by a colon.

Following are some examples of valid flags. More extensive examples are shown in the description of the IF statement (below)

```
Examples
Flag1:
A:
Loop1:
```

GOTO command

This command transfers macro processing to the next statement after a flag.

The syntax of this command is

GOTO {P1}

where

{P1} is a text variable or text constant.

Following are some examples of valid GOTO commands.

Examples
```
GOTO Flag1
or
F$ = "Flag1"
GOTO F$
```

IF command

This command transfers macro processing to the next statement after a flag if a condition is met. The syntax of this command is

IF {p1} {logic} {p2} GOTO {p3}

where

{p1} is a variable or constant.

{p2} is a variable or constant.

{p3} is a flag.

{logic} is a logic operator. Possible logic operators are =, <, >, <=, >=, and <>.

Following are some examples of valid IF commands.

Examples
```
IF x1# > 5 GOTO flag1
IF y$ = "A" GOTO flag2
IF y$ <> "A" GOTO flag3
```

INPUT

This command stops macro execution, display a message window, and waits for a value to be input. This value is then stored in the indicated macro variable.

The syntax of this command is

INPUT {variable} {prompt} {title} {default}

where

{variable} is the name of the variable (text or numeric) to receive the value that is input.

{prompt} is the text phrase that is shown on the input window.

{title} is the text phrase that is displayed at the top of the input window.

{default} is the default value for the input.

Following is an example of this command.

Example
```
INPUT A# "Enter the number of items" "Macro Input Window" 1
```

END

This command closes the ***NCSS/PASS*** system.

The syntax of this command is

END

Example
END

SendKeys

This command sends one or more keystrokes to the program as if you had typed them in from the keyboard. This facility allows you to create macros to accomplish almost anything you can do interactively within the program.

To use this, run the program from the keyboard, noting exactly which keys are pressed. Then, type the appropriate commands into the sendkeys text. Note that spaces are treated as characters, so '{down} {tab}' is different from '{down}{tab}'.

The syntax of this command is

SendKeys {value}

where *{value}* is a text constant or variable.

Remarks

Each key is represented by one or more characters. To specify a single keyboard character, use the character itself. For example, to represent the letter A, use "A" for value. To represent more than one character, append each additional character to the one preceding it. To represent the letters A, B, and C, use "ABC" for string.

The plus sign (+), caret (^), percent sign (%), tilde (~), and parentheses ( ) have special meanings to SendKeys. To specify one of these characters, enclose it within braces ({}). For example, to specify the plus sign, use {+}. Brackets ([ ]) have no special meaning to SendKeys, but you must enclose them in braces. To specify brace characters, use {{} and {}}.

To specify characters that are not displayed when you press a key, such as ENTER or TAB, and keys that represent actions rather than characters, use the following codes:

Key	Code
Backspace	{bs}
Break	{break}
Caps Lock	{capslock}
Delete	{delete}
Down Arrow	{down}
End	{end}
Enter	{enter}
Esc	{esc}
Home	{home}
Insert	{insert}
Left Arrow	{left}
Num Lock	{numlock}
Page Down	{pgdn}
Page Up	{pgup}
Right Arrow	{right}

Key	Code
Tab	{tab}
Up Arrow	{up}
F1	{F1}
F2	{F2}
F3	{F3}
F4	{F4}
F5	{F5}
F6	{F6}
F7	{F7}
F8	{F8}
F9	{F9}
F10	{F10}
F11	{F11}
F12	{F12}
F13	{F13}
F14	{F14}
F15	{F15}
F16	{F16}

To specify keys combined with any combination of the SHIFT, CTRL, and ALT keys, precede the key code with one or more of the following codes:

Key	Code
Shift	+
Ctrl	^
Alt	%

To specify that any combination of SHIFT, CTRL, and ALT should be held down while several other keys are pressed, enclose the code for those keys in parentheses. For example, to specify to hold down SHIFT while E and C are pressed, use "+(EC)". To specify to hold down SHIFT while E is pressed, followed by C without SHIFT, use "+EC".

To specify repeating keys, use the form {key number}. You must put a space between key and number. For example, {LEFT 4} means press the LEFT ARROW key 4 times; {h 8} means press H 8 times.

Spreadsheet Note: When a macro is run the usual beginning location on the screen is the new page icon (just below File) in the upper left of the screen. A single tab may be entered in the macro to go to the upper left cell position on the spreadsheet.

Examples	Action
SendKeys "ABC {enter}"	(Types ABC and then enter)
SendKeys "{enter right}"	(Enter and then down arrow)
SendKeys "%H{down 2}{enter}"	(Activates the Serial Numbers from the Help menus)
SendKeys "{right}"	(Moves to the right one cell)

Window Position Commands

These commands allow the user to position or hide windows while the macro is running.

List of Window Commands:

WindowLeft
WindowTop

WindowLeft

This command sets the position of the left edge of the spreadsheet, panel, and output windows. This allows you to effectively hide these windows while a macro is running.

The syntax of this command is

WINDOWLEFT {value}

where

{value} is the value of the left edge in thousandths of an inch.

Examples	Action
WINDOWLEFT 0	(positions the left edge to zero)
WINDOWLEFT 10000	(positions the left edge ten inches to the right)
WINDOWLEFT -10000	(positions the left edge ten inches to the left)

WindowTop

This command sets the position of the top of the spreadsheet, panel, and output windows. This allows you to effectively hide these windows while a macro is running.

The syntax of this command is

WINDOWTOP {value}

where

{value} is the value of the left edge in thousandths of an inch.

Examples	Action
WINDOWTOP 0	(positions the top to zero)
WINDOWTOP -10000	(positions the top ten inches up)

Dataset Commands

The following commands open, close, and modify an ***NCSS*** dataset.

List of Dataset Commands:

DataOpen
DataNewS0
DataNewS0Z
Import
Export
DataSaveS0
DataSaveS0Z
AddASheet

List of Dataset Commands (continued):
RemoveLastSheet
ResizeRowsCols
SortBy
GetCell
SetCell
MaxRows
GetMaxRows
VarName
VarLabel
VarTrans
VarFormat
VarDataType
VarValueLabel
NumTransRows
RunTrans
NumVars
VarFromList

DataOpen

This command opens the database given by *{file name}*.

The syntax of this command is

DataOpen {filename}

where

{file name} a text variable or text phrase enclosed in double quotes that gives the name of the database to be opened.

Following are some examples of macro snippets that use this command.

```
DataOpen "C:/Program Files/NCSS/Sample.s0"
or
F1$="C:/Program Files/NCSS/"
F2$="Sample.s0"
F$=F1$ & F2$
DataOpen F$
```

DataNewS0

This command creates a new, untitled spreadsheet-type database. Note that these databases are limited to 16,384 rows of data.

The syntax of this command is

DataNewS0

Following is an example of this command.

```
DataNewS0
```

DataNewS0Z

This command creates a database-type dataset.

The syntax of this command is

DataNewS0Z {file name} {number variables} {extra text fields}{variable field length} {extra text field length}

where

{file name} a text variable or text phrase enclosed in double quotes that gives the name of the database to be opened.

{number variables}

a number or number variable that gives the total number of variables in the new database. Set this number large enough to include any calculated variables, since it cannot be increased later.

{extra text fields}

a number or number variable that gives this value. Additional space is added to the end of each record for this many extra long (greater than six characters) text variables.

{variable field length}

a number or number variable that gives this value which is the length of a data-cell's field. This value must be at least ten. You should make this larger when the database will contain several long text values. Note that a text value requires four additional bytes of storage.

{extra text field length}

a number or number variable that gives this value which is the length of the extra text fields.

Here is an example.

```
DataNewS0Z "C:/Program Files/NCSS/Sample.s0z" 50  5  10  25
```

Import

This command imports a file of a different format and translates it into the ***NCSS*** format. Importing data is discussed in detail in the Importing Data chapter (chapter 115). We refer you to that chapter for the details of data importing.

The syntax of this command is

Import {type} {file name} {variables} {p1} {p2} {p3} {p4}

where

{type} a text variable or text phrase enclosed in double quotes that gives the type of database to be imported. Some example file types are Access, Excel, ASCII (text), and Dbase.

{file name} a text variable or text phrase enclosed in double quotes that gives the name of the database to be imported.

{variables} a list of variables to be imported. If you want so import all of the variables in the file, enter two double quotes ("") here.

{p1} {p2} {p3} {p4}

are extra parameters that are used for some file types, as described next.

This section describes additional parameters that are required by some file types.

Access

{p1} contains the name of the table in the Access database that is to be imported. If this field is left blank, the first table will be imported.

Excel

{p1} contains the name of the sheet in the spreadsheet that is to be imported. If this field is left blank, the first sheet will be imported.

ASCII Delimited

{p1} gives the delimiter that is used to separate values. Common choices are *blank* and *comma*.

{p2} gives the Name Row number. This is the number of the row in the file that contains the variable names.

{p3} gives the LinesObs number. This is the number of rows in the file that are devoted to a single row of data. Usually, this value is one.

ASCII Fixed

{p1} contains the Name Row number. This is the number of the row in the file that contains the variable names.

{p2} contains the Format statement. This statement tells how a row of data from the file is to be interpreted. Details of how to create a format statement are contained in the topic on *Importing Fixed Format ASCII Files* in the Help system.

Here is an example of importing an Excel spreadsheet.

```
DataNewS0
Import "Excel"  "C:/Program Files/NCSS/abc.xls"  ""
```

Here is an example of importing a fixed ASCII file.

```
DataNewS0
Import "Ascii Delimited"  "C:/Program Files/NCSS/abc.txt"  ""  "comma"  "1"  "1"
```

Export

This command exports an NCSS dataset to a file of a specified format. Exporting data is discussed in detail in the Exporting Data chapter (chapter 116). We refer you to that chapter for the details of data exporting.

The syntax of this command is

Export {type} {file name} {variables} {p1} {p2} {p3} {p4}

where

{type} a text variable or text phrase enclosed in double quotes that gives the type of file to be created. Some example file types are Access, Excel, ASCII (text), and Dbase.

{file name} a text variable or text phrase enclosed in double quotes that gives the name of the file into which the data are to be exported.

{variables} a list of variables to be exported. If you want so export all of the variables, enter two double quotes ("") here.

{p1} {p2} {p3} {p4}

are extra parameters that are used for some file types, as described next.

This section describes additional parameters that are required by some file types.

Access

{p1} gives the value of the Maximum Text Width parameter.

{p2} contains the name of the table in the Access database that is to receive the data.

ASCII Delimited

{p1} gives the Delimiter that is used to separate values. Common choices are blank and comma.

{p2} gives the Missing Indicator. Common choices are space and NA.

{p3} gives the Enclose Text Between parameter.

{p4} gives the Characters Per Line.

{p5} gives the Export Variable Names option. This value is "1" for yes or "0" for no.

ASCII Fixed

{p1} contains the Format statement. This statement tells how a row of data from the file is to be interpreted. Details of how to create a format statement are contained in the topic on Importing Fixed Format ASCII Files in the Help system.

{p2} gives the Missing Indicator. Common choices are space and NA.

{p3} gives the Characters Per Line.

{p4} gives the Export Variable Names option. This value is "1" for yes or "0" for no.

Here is an example of exporting data to an Excel spreadsheet.

```
Export "Excel"  "C:/Program Files/NCSS/abc.xls"  ""
```

Here is an example of importing a fixed ASCII file.

```
Export "Ascii Delimited"  "C:/Program Files/NCSS/abc.txt"  ""  "comma"  "space"  "variable"
"Y"
```

DataSaveS0

This command saves the current spreadsheet-type database to a file.

The syntax of this command is

DataSaveS0 {file name}

where

{file name} is a text variable or text phrase enclosed between double quotes. Note that the extension of the file name must be "s0". If any data already exists in this database, it will be replaced.

Following are some examples of macro snippets that use this command.

```
DataSaveS0 "C:/Program Files/NCSS/Sample.s0"
or
F1$="C:/Program Files/NCSS/"
F2$="Sample.s0"
F$=F1$ & F2$
DataSaveS0 F$
```

DataSaveS0Z

This command saves the current database-type database.

The syntax of this command is

DataSaveS0Z {file name}

where

{file name} is a text variable or text phrase enclosed between double quotes. Note that the extension of the file name must be "s0z". If any data already exists in this database, it will be replaced.

Following are some examples of macro snippets that use this command.

```
DataSaveS0Z "C:/Program Files/NCSS/Sample.s0"
or
F1$="C:/Program Files/NCSS/"
F2$="Sample.s0z"
F$=F1$ & F2$
DataSaveS0Z F$
```

AddASheet

This command adds another 'sheet' to the current spreadsheet file. Note that each sheet contains 256 variables. You should only add a second sheet if you have used up the 256 variables on the first sheet.

The syntax of this command is

AddASheet

```
Example
AddASheet
```

RemoveLastSheet

This command removes the last 'sheet' of the current spreadsheet file. Note that a sheet cannot be removed if it contains data. Also, the first sheet cannot be removed.

The syntax of this command is

RemoveLastSheet

Example
RemoveLastSheet

ResizeRowsCols

This command resizes all rows and columns in the spreadsheet or database according to the user-specified resize type. All three types create columns (rows) no narrower (shorter) than the default width (height). If this command is used with a spreadsheet (.S0), the resulting row heights and column widths are saved when the spreadsheet is saved. When this operation is used with a database (.S0Z), the resulting row heights and column widths are not saved when the database is saved. The command must be used each time the database is opened.

The syntax of this command is

ResizeRowsCols {rtype}

where

{rtype} is an integer corresponding to the type of row/column resizing to be done (0 = resize using defaults, 1 = resize using data and titles, 2 = resize using data only).

Example
ResizeRowsCols 1

SortBy {v1} {o1} {v2} {o2} {v3} {o3}

This command executes the Sort command on the currently open dataset. The dataset can be sorted by up to three variables (columns). The syntax of this command is

SORTBY {v1} {o1} {v2} {o2} {v3} {o3}

where

{v1} is the name or number of the first variable to be sorted.

{o1} is the sort type (1 = ascending, 0 = descending).

{v2} is the name or number of the second variable to be sorted. This parameter is optional.

{o2} is the sort type (1 = ascending, 0 = descending). This parameter is only used of {v2} is used.

{v3} is the name or number of the third variable to be sorted. This parameter is optional.

{o3} is the sort type (1 = ascending, 0 = descending). This parameter is only used of {v3} is used.

Examples
SORTBY C1 0
SORTBY 1 0
SORTBY "HeartRate" 0

GetCell

This command obtains the value of a spreadsheet cell. The syntax of this command is

GetCell {variable} {row} {macro variable}

where

{variable} is the name or number of the variable (column) with the cell to be read.

{row} is the row number of the cell to be read.

{macro variable} is the text or numeric macro variable that holds the value of the spreadsheet cell.

Examples

```
GetCell "HeartRate" 27 H#
GetCell Name 16 Name16$
```

SetCell

This command sets a spreadsheet cell to a specified value. The syntax of this command is

SetCell {v1} {row1} {row2} {value}

where

{v1} is the name or number of the variable to receive the new value.

{row1} is the first row in a range of rows to receive the new value.

{row2} is the last row in a range of rows to receive the new value

{value} is the new value. This value may be text or numeric.

Examples

```
SetCell "HeartRate" 10 10 "100"
SetCell 1 10 20 100
```

NumRows

This command loads the number of rows used in a dataset column into a program variable. The syntax of this command is

NumRows {v1} {n}

where

{v1} is a variable name or number on the current database.

{n} is a numeric macro variable.

Examples

```
NumRows "HeartRate" n1#
NumRows 1 n#
```

GetMaxRows

This command loads the maximum number of rows used by any variable into a program variable.

The syntax of this command is

GetMaxRows {n}

where

{n} is a numeric macro variable.

Examples
GetMaxRows n1#
GetMaxRows n#

VarName

This command sets the name of the specified variable.

The syntax of this command is

VARNAME {variable} {name}

where

{variable} is the current name or number of the variable to be renamed.

{name} is the new name of the variable. This name must follow standard NCSS variable name restrictions.

Examples
VARNAME C1 "HeartRate"
VARNAME 1 "HeartRate"

VarLabel

This command sets the label of the specified variable.

The syntax of this command is

VARLABEL {variable} {label}

where

{variable} is the name or number of the variable to be labelled.

{label} is the new label of the variable.

Examples
VARLABEL C1 "Heart Rate"
VARLABEL 1 "Heart Rate"
VARLABEL "HR" "Heart Rate"

VarTrans

This command sets the transformation of the specified variable.

The syntax of this command is

VARTRANS {variable} {trans}

where

{variable} is the name or number of the variable to be transformed.

{trans} is the transformation formula.

Examples
VARTRANS C1 "Log(C1)"
VARTRANS 2 "C2*C3"

VarFormat

This command sets the display format of the specified variable.

The syntax of this command is

VARFORMAT {variable} {format}

where

{variable} is the name or number of the variable to be formatted.

{format} is the format.

Examples
VARFORMAT C1 "0.00"
VARFORMAT 2 "MM/DD/YYYY"

VarDataType

This command sets the datatype of the specified variable.

The syntax of this command is

VARDATATYPE {variable} {type}

where

{variable} is the name or number of the variable.

{type} is the number of the data type.

Examples
VARDATATYPE C1 1
VARDATATYPE 1 1

VarValueLabel

This command sets the value label of the specified variable.

The syntax of this command is

VARVALUELABEL {variable} {vlab}

where

{variable} is the name or number of the variable.

{vlab} is the variable name or number of the variable containing the value labels or the name of the file containing the value labels.

Examples
VARVALUELABEL C1 C4
VARVALUELABEL X3 "C:\Program files\NCSS97\abc.txt"

NumTransRows

This command sets the number of rows that are transformed when the transformations are run. Setting this value to zero causes all rows to be transformed.

The syntax of this command is

NUMTRANSROWS {rows}

where

{rows} is the number of rows.

Examples
NUMTRANSROWS 0
NUMTRANSROWS 100

RunTrans

This command causes all transformation on the current database to be executed.

The syntax of this command is

RUNTRANS

Example
RUNTRANS

NumVars

This command causes the number of variables contained in a list of variables to be loaded into a macro variable.

The syntax of this command is

NUMVARS {varlist} {x}

where

{varlist} is an expression containing variable names.

{x} is macro variable.

Example
NUMVARS "C2:C10, C15" nvars#

VarFromList

This command causes number of the ith variable in a list to be loaded into a macro variable.

The syntax of this command is

VarFromList {varlist} {i} {v}

where

{varlist} is an expression containing variable names.

{i} is the item number to be selected from the variable list.

{v} is the macro variable that receives the number of the ith item in the list.

Example
VARFROMLIST "C1,C2,C3,C10,C20" 4 V1# (V1 becomes 10)
VARFROMLIST "C1,C2,C3,C10,C20" 5 V2# (V2 becomes 20)

Procedure Commands

The following commands open, modify, run and close procedures.

List of Procedure Commands:

LoadProc
RunProc
SaveTemplate
UnloadProc
Option

LoadProc

This command loads the designated procedure window. Once loaded, the options of the procedure may be modified and then the procedure can be executed.

The syntax of this command is

LoadProc {proc} {template}

where

{proc} is a variable or constant that gives the name or number of the procedure to be loaded. Each procedure's name and number is displayed near the bottom of the window under the Template tab.

{template} is an optional text variable or text constant that gives the name of a template file that is loaded with this procedure. If this value is omitted, the default (last) template for this procedure is loaded. Note that the text value does not include the extension or the folder information for the template file.

Following are some examples of valid LOADPROC commands.

```
Example
LOADPROC 24 "macro 1"
LOADPROC DescStat "macro 1"
LOADPROC DescStat
LOADPROC 24
```

RunProc

This command executes the indicated procedure. The syntax of this command is

RunProc {proc} {template}

where

{proc} is a variable or constant that gives the name or number of the procedure to be run. Each procedure's name and number is displayed near the bottom of the window under the Template tab.

{template} is a required variable or text constant that gives the name of the resulting template file. Note that the text value does not include the extension or the folder information for the template file.

SaveTemplate

This command saves the settings in the last procedure loaded to a template file. Once loaded, the options of the procedure may be modified and then the procedure can be executed.

The syntax of this command is

SaveTemplate {proc} {template} {id}

where

{proc} is a variable or constant that gives the name or number of the procedure whose template is to be saved. Each procedure's name and number is displayed near the bottom of the window under the Template tab.

{template} is a required variable or text constant that gives the name of the resulting template file. Note that the text value does not include the extension or the folder information for the template file.

{id} is an optional text variable or text constant that is stored with the file. This text is displayed with the file name under the Template tab.

Following are some examples of valid SaveTemplate commands.

Example
```
SaveTemplate DescStat "Template1"
SaveTemplate DescStat "Template1" "This template was created by a macro on January 1"
```

UnloadProc

This command closes the indicated procedure window. The syntax of this command is

UnloadProc {proc}

where

{proc} is a variable or constant that gives the name or number of the procedure. Each procedure's name and number is displayed near the bottom of the window under the Template tab.

Option

This command lets you set the values of the individual options of a procedure. For example, you may want to change the name of the variable that is to be processed.

The syntax of this command is

Option {proc} {number} {value1} {value2} {value3} …

where

{proc} is a variable or constant that gives the name or number of the procedure. Each procedure's name and number is displayed near the bottom of the window under the Template tab.

{number} is the number of the option that is to be set. This number is displayed at the lower, left corner of the procedure window when the mouse is positioned over that option.

If the 'Opt' value is not displayed, activate it by doing the following: from the spreadsheet menus select Edit, Options, and View. Check the option labeled 'Show Option Numbers'.

{value1} is the new value of the option.

{value2}… are the new values of the remaining parameters of the option. Most options only have one value, so a second value is not necessary. However, a few options, such as text properties, bring up a window for option selection. These options have two or more parameters.

Following are some examples of valid OPTION commands.

Example
```
OPTION 24 2 4
OPTION “DescStat” 4 "HeartRate"
```

Output Commands

The following commands manage the output (word processor) windows.

List of Output Commands:

SaveOutput
ClearOutput
PrintOutput
AddToLog
NewLog
SaveLog
OpenLog

SaveOutput

This command saves the current output to the designated file name.

The syntax of this command is

SaveOutput {filename}

{filename} a text constant or variable that gives the name of the file to receive the output. Note that the extension of the file name should be '.RTF'.

Example
```
SAVEOUTPUT "C:/Program Files/NCSS/Sample.rtf"
```

ClearOutput

This command clears (erases) the current output.

The syntax of this command is

ClearOutput

Example
```
CLEAROUTPUT
```

PrintOutput

This command prints the current output.

The syntax of this command is

PrintOutput

Example
```
PRINTOUTPUT
```

AddToLog

This command copies the output in the output window to the log window. Note that nothing is saved by this command.

The syntax of this command is

AddToLog

Example
```
ADDTOLOG
```

NewLog

This command clears log window.

The syntax of this command is

NewLog

Example
```
NEWLOG
```

SaveLog

This command saves the current contents of the log output window to the designated file name.

The syntax of this command is

SaveLog {filename}

{filename} a text constant or variable that gives the name of the file to receive the log. Note that the extension of the file name should be '.RTF'.

Example
```
SAVELOG "C:/Program Files/NCSS/Reports/Sample.rtf"
```

OpenLog

This command opens and displays the contents of the specified file.

The syntax of this command is

OpenLog {filename}

{filename} a text constant or variable that gives the name of the file to opened. Note that only RTF files can be opened.

Example
```
OPENLOG "C:/Program Files/NCSS/Sample.rtf"
```

Alphabetical Macro Command List

AddASheet
AddToLog
ClearOutput
DataNewS0
DataNewS0Z {file name} {number variables} {extra text fields}{variable field length} {extra text field length}
DataOpen {filename}
DataSaveS0 {file name}
DataSaveS0Z {file name}
End
Export {type} {file name} {variables} {p1} {p2} {p3} {p4} {flag}:
GetCell {variable} {row} {macro variable}
GetMaxRows {n}
Goto {P1}
Heading {h1}
If {p1} {logic} {p2} goto {p3}
Import {type} {file name} {variables} {p1} {p2} {p3} {p4}
Input {variable} {prompt} {title} {default}
LoadProc {proc} {template}
LoadTemplate {proc} {template}
NewLog
NumRows {v1} {n}
NumTransRows {rows}
NumVars {varlist} {x}
OpenLog {filename}
Option {number} {value1} {value2} {value3} …
Print {p1} {p2} {p3} …
PrintOutput
RemoveLastSheet
ResizeRowsCols
RunProc {proc}
RunTrans
SaveLog {filename}
SaveOutput {filename}
SaveTemplate {proc} {template} {id}
SendKeys {value}
SetCell {v1} {row1} {row2} {value}
SortBy {v1} {o1} {v2} {o2} {v3} {o3}
UnloadProc {proc}
VarDataType {variable} {type}
VarFormat {variable} {format}
VarFromList {varlist} {i} {v}
VarLabel {variable} {label}
VarName {variable} {name}
VarTrans {variable} {trans}
VarValueLabel {variable} {vlab}
WindowLeft {value}
WindowTop {value}

Examples

The following section provides examples of NCSS macros. Our intention is that these examples will help you learn how to write macros to accomplish various repetitive tasks with NCSS.

Example 1 – Automatically Run a Procedure

This macro opens a dataset and calculates descriptive statistics on the first five variables of that dataset.

```
'***  Open a dataset
DataOpen "C:/Program Files/NCSS2007/Data/Sample.s0"

'***  Load the Descriptive Statistics procedure
'***  Note that the default values are used.
LoadProc DescStat

'***  Specify that the first five variables are to be analyzed
Option DescStat  1 "1:5"

'***  Run the analysis
RunProc DescStat
'*** End of Macro 1
```

Example 2 – Run Several Procedures

This macro opens a dataset, calculates descriptive statistics on the first two variables, and then runs a linear regression using those two variables.

```
DataOpen "C:/Program Files/NCSS2007/Data/Sample.s0"
LoadProc DescStat
Option DescStat  1 "1:2"
RunProc DescStat
LoadProc LinReg
Option LinReg 204 1
Option LinReg 205 2
RunProc LinReg
```

Example 3 – Simple Looping

This macro opens a dataset and calculates descriptive statistics on the first five variables of that dataset. Unlike Example1, this macro uses a simple loop to step through the five variables.

Notice that the loop finishes once the variable I# is equal to '5'.

```
DataOpen "C:/Program Files/NCSS2007/Data/Sample.s0"
I#=0
Flag1:
    I#=I#+1
    LoadProc DescStat
    Option DescStat 1 I#
    RunProc DescStat
IF I#<5 GOTO Flag1
```

Example 4 – Using NumVars and VarFromList

This macro opens a dataset and calculates descriptive statistics on a set of five variables from that dataset. Note that these variables are not contiguous. Also note that the NumVars command is used to load a numeric variable n# with the number of variables.

```
DataOpen "C:/Program Files/NCSS2007/Data/Sample.s0"
LoadProc DescStat
'*** Turn off all but the Summary Section report
option DescStat 17 0
option DescStat 18 0
option DescStat 19 0
option DescStat 20 0
option DescStat 42 0
option DescStat 43 0
option DescStat 44 0
option DescStat 45 0
option DescStat 46 0
option DescStat 47 0
option DescStat 48 0
option DescStat 49 0
'*** Load the variable list into V$
V$="Height Weight YldA YldB YldC"
'*** Load the number of variables in n#
NumVars V$ n#
I#=0
Flag1:
    I#=I#+1
    '*** Load the ith variable from the list
    VarFromList V$ I# name$
    '*** Set the Variable to name$
    Option DescStat 1 name$
    RunProc DescStat
IF I#<n# GOTO Flag1
```

Example 5 – Multiple Analyses using Filters

This macro opens a dataset and runs separate simple linear regression analyses and scatter plots for three groups within the dataset. The three groups are defined be the 'Iris' variable in the dataset. The macro finds the levels of the 'Iris' variable and uses the resulting values to determine the groups. Thus, the macro does not require knowledge of the group names nor does it require knowledge of the number of groups to be analyzed.

```
'*************************************************************************
'Open the FISHER.S0 database. %p% is replaced by the folder that
'was used for installation.
'*************************************************************************
DataOpen "%p%\Data\FISHER.S0"

'*************************************************************************
'Obtain the unique values of the Iris variable and put them in C21
'*************************************************************************
VarTrans C21 "Uniques(Iris)"
RunTrans

'*************************************************************************
'Obtain the number of unique Iris values using C21
'*************************************************************************
NumRows C21 n1#

'*************************************************************************
'Loop through the unique Iris values for grouping
'*************************************************************************
I1# = 1
Flag1:
GetCell C21 I1# IrisGroup$

'*************************************************************************
'Filter the values according to the Iris variable
'*************************************************************************
LoadProc Filter
LoadTemplate Filter "Example5_Macro"
IrisFilter$ = "Iris = " & IrisGroup$
Option Filter 3 IrisFilter$
RunProc Filter
UnLoadProc Filter
```

```
'*************************************************************************
'Run the Linear Regression procedure for the filtered group
'*************************************************************************
LoadProc LinReg
LoadTemplate LinReg "Example5_Macro"
RunProc LinReg
UnLoadProc LinReg

'*************************************************************************
'Run the Scatter Plot procedure for the filtered group
'*************************************************************************
LoadProc ScatPlot
LoadTemplate ScatPlot "Example5_Macro"
RunProc ScatPlot
UnLoadProc ScatPlot

'*************************************************************************
'Loop through the unique Iris values for grouping
'*************************************************************************
I1# = I1# + 1
IF I1# > n1# GOTO Flag2
GOTO Flag1

Flag2:

'*************************************************************************
'Turn Filter Off
'*************************************************************************
LoadProc Filter
Option Filter 1 0
RunProc Filter
UnLoadProc Filter

'*************************************************************************
'End of Macro
'*************************************************************************
Print "MACRO COMPLETE"
```

A macro without the comments is included as one of the existing macros.

Chapter 135

Probability Calculator

Introduction

Most statisticians have a set of probability tables that they refer to in doing their statistical work. This procedure provides you with a set of electronic statistical tables that will let you look up values for various probability distributions.

To run this option, select Probability Calculator from the Other menu of the Analysis menu. A window will appear that will let you indicate which probability distribution you want to use along with various input parameters. Select the Calculate button to find and display the results.

Many of the probability distributions have two selection buttons to the left of them. The first (left) button selects the inverse probability distribution. An inverse probability distribution is in a form so that when you give it a probability, it calculates the associated critical value. The second (right) button selects the regular probability distribution which is formulated so that when you give it a critical value, it calculates the (left tail) probability.

Probability Distributions

Beta Distribution

The beta distribution is usually used because of its relationship to other distributions, such as the t and F distributions. The noncentral beta distribution function is formulated as follows:

$$\Pr(0 \le x \le X \mid A, B, L) = I_X(A, B, L) = \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} \sum_{k=0}^{\infty} \frac{e^{-L} L^k}{k! 2^k} \int_0^X t^{A+k-1}(1-t)^{B-1} dt$$

where

$$0 < \text{A},\ 0 < \text{B},\ 0 \le \text{L},\ \text{and}\ 0 \le \text{x} \le 1$$

When the noncentrality parameter (NCP), L, is set to zero, the above formula reduces to the *standard* beta distribution, formulated as

$$\Pr(0 \le x \le X \mid A, B) = \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} \int_0^X t^{A-1}(1-t)^{B-1} dt$$

When the inverse distribution is selected, you supply the probability value and the program solves for *X*. When the regular distribution is selected, you supply *X* and the program solves for the cumulative (left-tail) probability.

Binomial Distribution

The binomial distribution is used to model the counts of a sequence of independent binary trials in which the probability of a success, *P*, is constant. The total number of trials (sample size) is *N*. *R* represents the number of successes in *N* trials. The probability of exactly R successes is:

$$\Pr(r = R \mid N, P) = \binom{N}{R} P^R (1-P)^{N-R}$$

where

$$\binom{N}{R} = \frac{N!}{R!(N-R)!}$$

The probability of from 0 to R successes is given by:

$$\Pr(0 \le r \le R \mid N, P) = \sum_{r=0}^{R} \binom{N}{r} P^r (1-P)^{N-r}$$

When the inverse distribution is selected, you supply the probability value and the program solves for *R*. When the regular distribution is selected, you supply *R* and the program solves for the cumulative (left-tail) probability.

Bivariate Normal Distribution

The bivariate normal distribution is given by the formula

$$\Pr(x < h, y < k \mid r) = \frac{1}{2\pi\sqrt{1-r^2}} \int_{-\infty}^{h} \int_{-\infty}^{k} \exp\left\{\frac{-x^2 + 2rxy - y^2}{2(1-r^2)}\right\} dx\, dy$$

where *x* and *y* follow the bivariate normal distribution with correlation coefficient *r*.

Chi-Square Distribution

The Chi-square distribution arises often in statistics when the normally distributed random variables are squared and added together. DF is the degrees of freedom of the estimated standard error.

The noncentral Chi-square distribution function is used in power calculations. The noncentral Chi-square distribution is calculated using the formula:

$$\Pr(0 \le x \le X \mid df, L) = \sum_{k=0}^{\infty} \frac{L^k e^{-L}}{2^k k!} P(X \mid df + 2k)$$

where

$$P(X|df) = \frac{1}{2^{df/2}\Gamma\left(\frac{df}{2}\right)} \int_0^X t^{df/2-1} e^{-t/2} dt$$

When the noncentrality parameter (NCP), L, is set to zero, the above formula reduces to the (central) Chi-square distribution.

When the inverse distribution is selected, you supply the probability value and the program solves for *X*. When the regular distribution is selected, you supply *X* and the program solves for the cumulative (left-tail) probability.

Correlation Coefficient Distribution

The correlation coefficient distribution is formulated as follows:

$$\Pr(r \le R|n,\rho) = \int_{-1}^{R} \frac{2^{n-3}}{\pi(n-3)!}(1-\rho)^{(n-1)/2}(1-r)^{(n-4)/2} \sum_{i=0}^{\infty} \Gamma^2\left(\frac{n+i-1}{2}\right) \frac{(2\rho r)^i}{i!} dr$$

where

$$|r| < 1, |\rho| < 1, \text{ and } |R| < 1$$

When the inverse distribution is selected, you supply the probability value and the program solves for *R*. When the regular distribution is selected, you supply *R* and the program solves for the cumulative (left-tail) probability.

F Distribution

The F distribution is used in the analysis of variance and in other places where the distribution of the ratio of two variances is needed. The degrees of freedom of the numerator variance is DF1 and the degrees of freedom of the denominator variance is DF2.

The noncentral-F distribution function is used in power calculations. We calculate the noncentral-F distribution using the following relationship between the F and the beta distribution function.

$$\Pr(0 \le f \le F|df_1, df_2, L) = I_X\left(\frac{df_1}{2}, \frac{df_2}{2}, L\right)$$

where

$$X = \frac{F(df_1)}{F(df_1) + df_2}$$

When the noncentrality parameter (NCP), L, is set to zero, the above formula reduces to the *standard* F distribution

When the inverse distribution is selected, you supply the probability value and the program solves for *F*. When the regular distribution is selected, you supply *F* and the program solves for the cumulative (left-tail) probability.

Hotelling's T2 Distribution

Hotelling's *T*-Squared distribution is used in multivariate analysis. We calculate the distribution using the following relationship between the *F* and the *T*2 distribution function.

$$\Pr(0 \le x \le \frac{(df - k + 1)}{k(df)} T^2_{k,df} \,|k, df) = \Pr(0 \le x \le F_{k,df-k+1} \,|k, df)$$

where *k* is the number of variables and *df* is the degrees of freedom associated with the covariance matrix. When the inverse distribution is selected, you supply the probability value and the program solves for *T*2. When the regular distribution is selected, you supply *T*2 and the program solves for the cumulative (left-tail) probability.

Gamma Distribution

The Gamma distribution is formulated as follows:

$$\Pr(0 \le g \le G \,|A, B) = \frac{1}{B^A \Gamma(A)} \int_0^G x^{A-1} e^{-X/B} dx$$

where

$$\Gamma(A) = \int_0^\infty x^{A-1} e^{-x} dx$$

$$0 < \text{A}, 0 < \text{B}, \text{ and } 0 \le \text{G}$$

When the inverse distribution is selected, you supply the probability value and the program solves for *G*. When the regular distribution is selected, you supply *G* and the program solves for the cumulative (left-tail) probability.

Hypergeometric Distribution

The hypergeometric distribution is used to model the following situation. Suppose a sample of size R is selected from a population with N items, M of which have a characteristic of interest. What is the probability that X of the items in the sample have this characteristic.

The probability of exactly X successes is:

$$\Pr(x = X \,|N, M, R) = \frac{\binom{M}{X}\binom{N-M}{R-X}}{\binom{N}{R}}$$

where

$$\binom{N}{R} = \frac{N!}{R!(N-R)!}$$

$$\text{Maximum}(0, \text{R-N+M}) <= \text{X} <= \text{Minimum}(\text{M}, \text{R})$$

Negative Binomial Distribution

The negative binomial distribution is used to model the counts of a sequence of independent binary trials in which the probability of a success, P, is constant. The total number of trials (sample size) is N. R represents the number of successes in N trials. Unlike the binomial distribution, the sample size, N, is the variable of interest.

The question answered by the negative binomial distribution is: how many tosses of a coin (with probability of a head equal to P) is necessary to achieve R heads and X tails.

The probability of exactly R successes is:

$$\Pr(x = X \mid R, P) = \binom{X + R - 1}{R - 1} P^R (1 - P)^X$$

where

$$\binom{N}{R} = \frac{N!}{R!(N - R)!}$$

Normal Distribution

The normal distribution is formulated as follows:

$$\Pr(x \le X \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{X} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx$$

When the mean is 0 and the variance is 1, we have the standard normal distribution. The regular normal distribution uses the variable X. The standard normal distribution uses the variable Z. Any normal distribution may be transformed to the standard normal distribution using the relationship:

$$z = \frac{x - \mu}{\sigma}$$

When the inverse distribution is selected, you supply the probability value and the program solves for *R*. When the regular distribution is selected, you supply *R* and the program solves for the cumulative (left-tail) probability.

Poisson Distribution

The Poisson distribution is used to model the following situation. Suppose the average number of accidents at a given intersection is 13.5 per year. What is the probability of having 2 accidents during the next half year?

The probability of exactly X occurrences with a mean occurrence rate of M is:

$$\Pr(x = X \mid M) = \frac{e^{-M} M^X}{X!}$$

Studentized Range Distribution

The studentized range distribution is used whenever the distribution of the ratio of a range and an independent estimate of its standard error is needed. This distribution is used quite often in multiple comparison tests run after an analysis of variance. DF is the degrees of freedom of the estimated standard error (often the degrees of freedom of the MSE). K is the number of items (means) in the sample. The distribution function is given by:

$$\Pr(0 \le r \le R \,|\, df, k) = \int_0^{\infty} \left(\frac{2^{-df/2+1} df^{df/2} s^{df-1}}{\Gamma\left(\frac{df}{2}\right)} \exp\left(-\frac{dfs^2}{2}\right) P(Rs|n) \right) dx$$

where *P(Rs|n)* is the probability integral of the range.

When the inverse distribution is selected, you supply the probability value and the program solves for *R*. When the regular distribution is selected, you supply *R* and the program solves for the cumulative (left-tail) probability.

Student's t Distribution

The t distribution is used whenever the distribution of the ratio of a statistic and its standard error is needed. DF is the degrees of freedom of the estimated standard error.

The noncentral-t distribution function is used in power calculations. We calculate the noncentral-t distribution using the following relationship between the t and the beta distribution function.

$$\Pr(-\infty \le t \le T \,|\, df, L) = 1 - \sum_{k=0}^{\infty} e^{-L^2/2} \frac{(L^2/2)^k}{2k!} I_X\left(\frac{df}{2}, \frac{1}{2}, 0\right)$$

where

$$X = \frac{df}{df + T}$$

When the noncentrality parameter (NCP), L, is set to zero, the above formula reduces to the (central) Student's t distribution

When the inverse distribution is selected, you supply the probability value and the program solves for *T*. When the regular distribution is selected, you supply *T* and the program solves for the cumulative (left-tail) probability.

Weibull Distribution

The Weibull distribution is formulated as follows:

$$\Pr(t \le T \,|\, \lambda, \gamma) = 1 - \exp\left(-(\lambda T)^{\gamma}\right)$$

When gamma (γ) equal to one, the distribution simplifies to the exponential distribution.

When the inverse distribution is selected, you supply the probability value and the program solves for *T*. When the regular distribution is selected, you supply *T* and the program solves for the cumulative (left-tail) probability.

Converting Summary Statistics to Raw Data

Occasionally, you will have summary statistics (mean, count, and standard deviation) but not the raw data used to create these summary statistics. Since you need the original data values, you cannot run descriptive statistics, t-tests, or AOV's on these data. This routine is useful in this case since it generates a set of raw data values with a given mean, count, and standard deviation. Although the generated values are not the same as the original values, they are good enough to use as input into statistical procedures such as analysis of variance or two-sample t-tests.

Let *n* represent the sample size (count), *M* represent the sample mean, and *s* represent the sample standard deviation. (Recall that the standard error is equal to the standard deviation times the square root of *n*, so if you have been given the standard error, use *s=SQR(n)(s.e.)*.) It is possible to find values, *X1* and *X2*, so that a variable made up of one *X1* and (n-1) *X2's* will have the same values of *n*, *M*, and *s*. The formulas to do this are:

$$X1 = M - (n-1)\frac{s}{\sqrt{n}}$$

and

$$X2 = M + \frac{s}{\sqrt{n}}$$

For example, suppose you have the following two sets of summary statistics and want to run a two-sample t-test.

	Sample One	Sample Two
Sample Size	4	5
Mean	3.2	4.5
Standard Deviation	1.4	1.8

Using the *Convert Mean and S to Data Values* option of the *Probability Calculator* and plugging in these values gives the following results:

	Sample One	Sample Two
X1	1.1	1.280062
N1	1	1
X2	3.9	5.304984
N2	3	4

These data would then be entered into a datasheet as follows:

Row	C1	C2
1	1.1	1.280062
2	3.9	5.304984
3	3.9	5.304984
4	3.9	5.304984
5		5.304984

Now, a two-sample t-test or a one-way analysis of variance could be run on these two variables. Note that since nonparametric tests, tests of assumptions, box plots, and histograms require the original data values, their output would have to be ignored, but the t-test results would be accurate.

Chapter 140

Introduction to Graphics

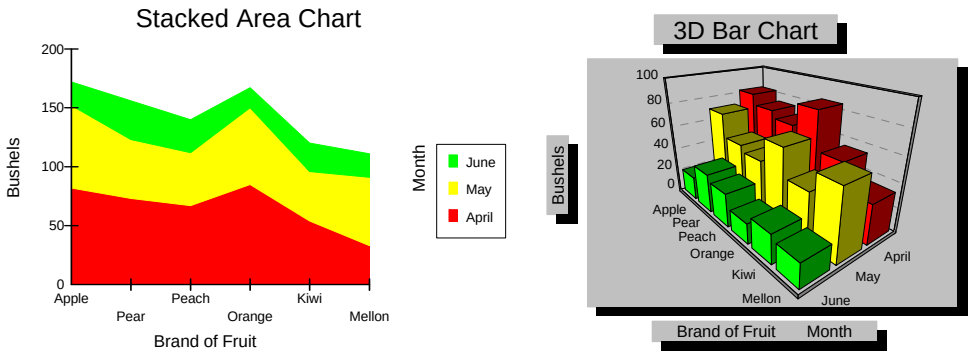
Introduction

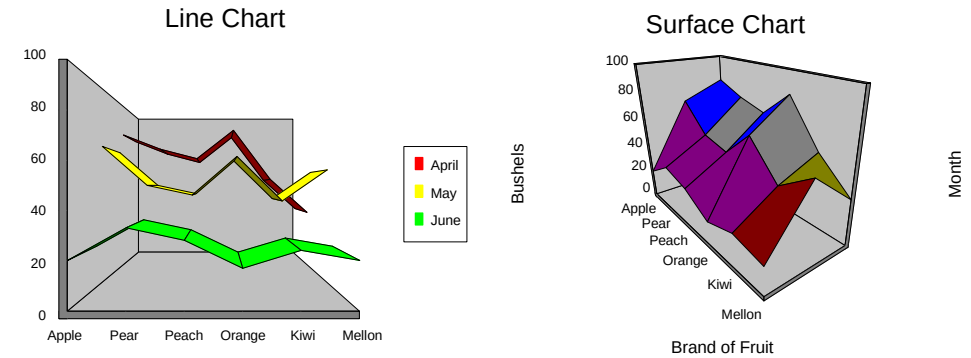
NCSS has been designed for quick and easy production of powerful graphics. This chapter gives a brief overview of the graphics available in ***NCSS***.

Single-Variable Charts

Bar Charts

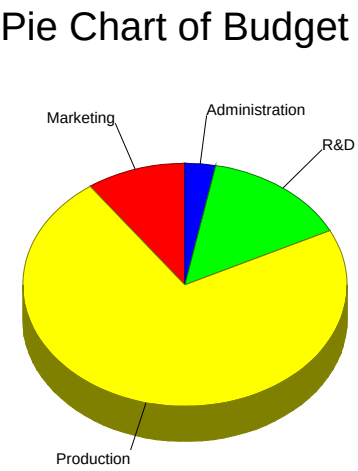
Bar charts are used to visually compare values to each other. There are several variations on the bar chart. These include the vertical bar chart, horizontal bar chart, area chart, line chart, and the surface chart. There are two and three dimensional versions of each. Below are some examples.





Pie Charts

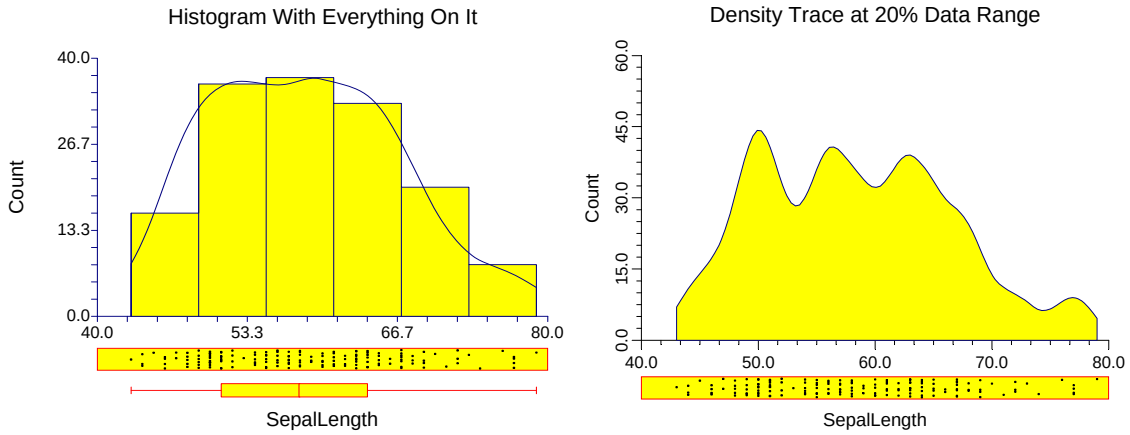
The pie chart is constructed by dividing a circle into two or more sections. The chart is used to show the proportion that each part is of the whole. Hence, it should be used when you want to compare individual categories with the whole. If you want to compare the values of categories with each other, use a bar chart or scatter plot.



Histograms

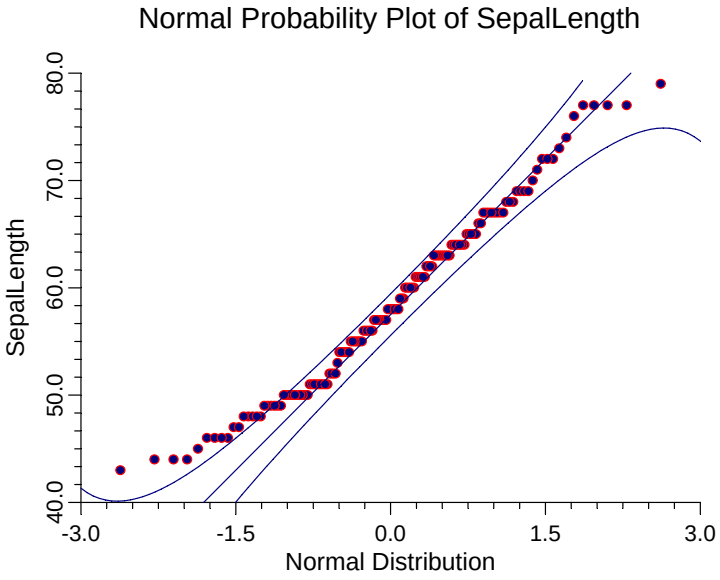
A histogram is used to display the distribution of data values along the real number line. It competes with the probability plot as a method of assessing normality. Humans cannot comprehend a large batch of observations just by reading them. To interpret the numbers, you must summarize them by sorting, grouping, and averaging. One method of doing this is to construct a frequency distribution. This involves dividing up the range of the data into a few (usually equal) intervals. The number of observations falling in each interval is counted. This gives a frequency distribution.

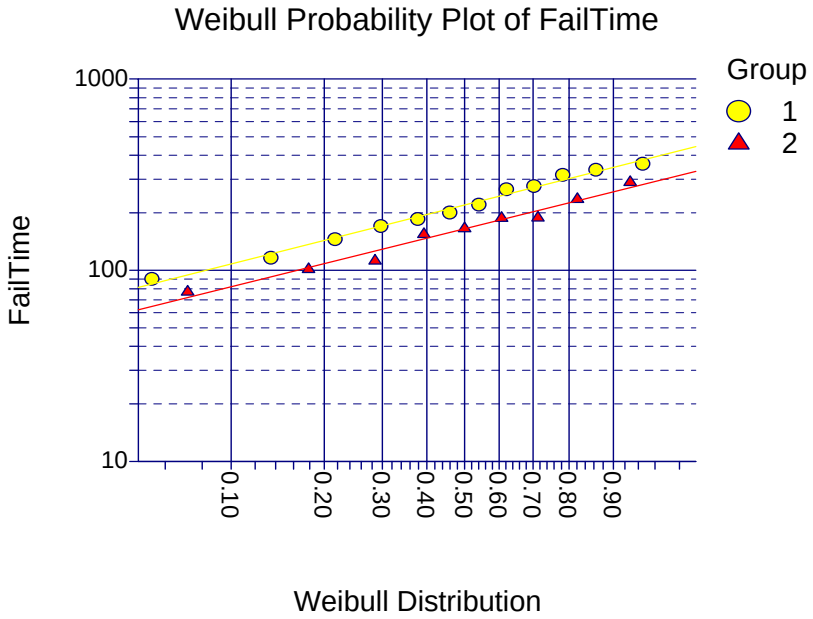
The histogram is a graph of the frequency distribution in which the vertical axis represents the count (frequency) and the horizontal axis represents the possible range of the data values.



Probability Plots

NCSS constructs probability plots for the Normal, Weibull, Chi-squared, Gamma, Uniform, Exponential, and Half-Normal distributions. It lets you try various transformations to see if one more closely fits the distribution of interest. Approximate confidence limits are drawn to help determine if a set of data follows a given distribution. If a grouping variable is specified, a separate line is drawn and displayed for each unique value of the grouping variable.

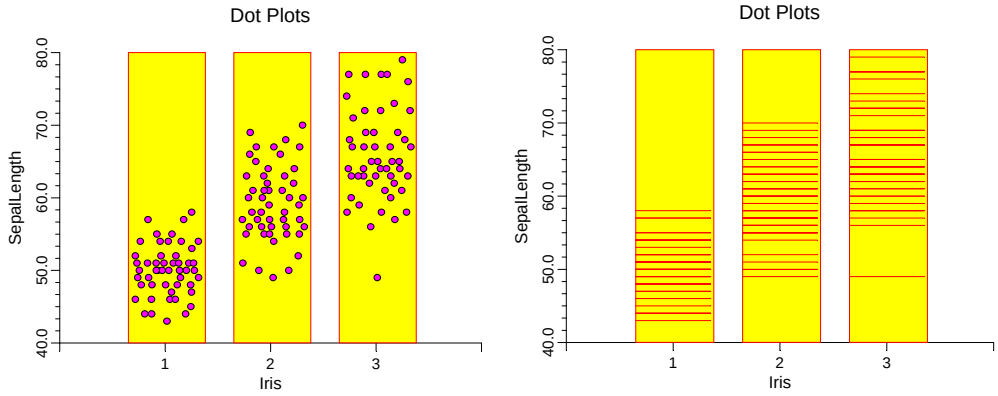




Two-Variable Charts (Discrete / Continuous)

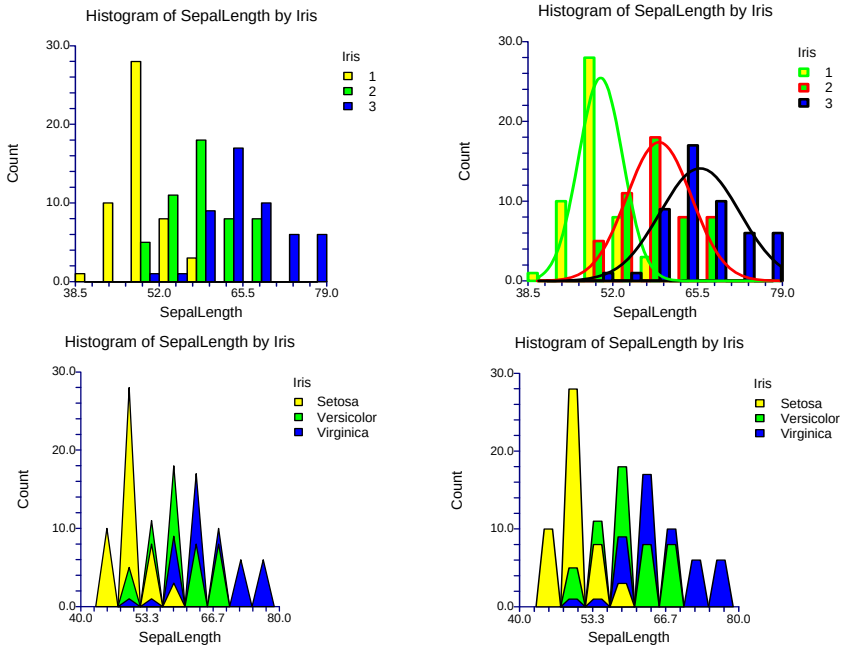
Dot Plots

The dot plot is a plot of a single batch of data. One version shows values as points. Another version shows values as lines. Dot plots are usually augmented to other plots, such as the scatter plot. The dot plot is especially useful for detecting strange patterns in your data. These patterns will show up as horizontal lines (of points).



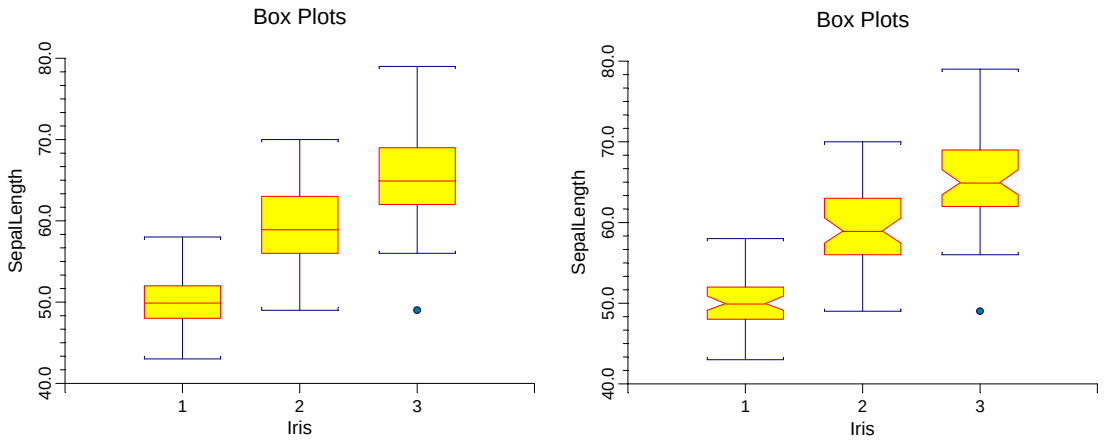
Histograms - Comparative

A histogram displays the frequency distribution of a set of data values. This procedure displays a comparative histogram created by interspersing or overlaying the individual histograms of two or more groups or variables. This allows the direct comparison of the distributions of several groups. Here are some examples.



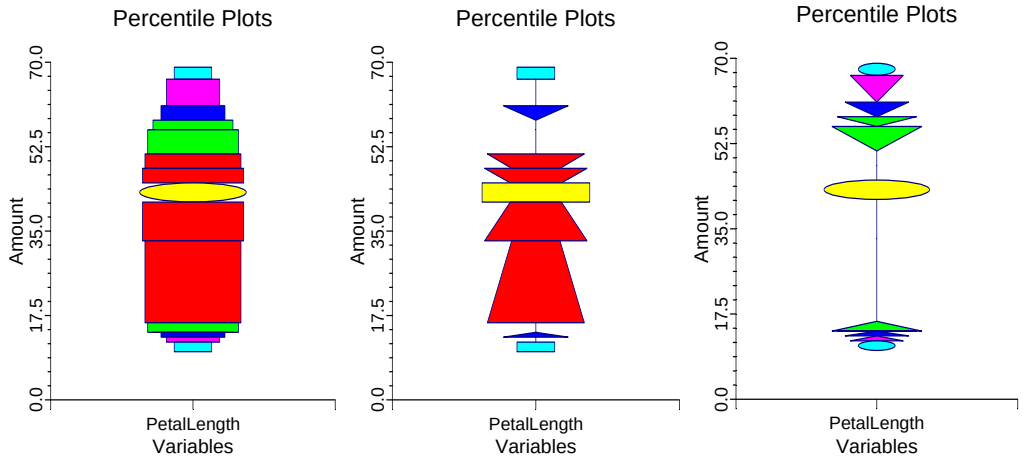
Box Plots

When analyzing data, you often need to study the characteristics of a single batch of numbers, observations, or measurements. You might want to know the center and how spread out the data are about this central value. You might want to investigate extreme values (referred to as outliers) or study the distribution of the data values (the pattern of the data values along the measurement axis). The box plot shows three main features about a variable: its center, its spread, and its outliers.



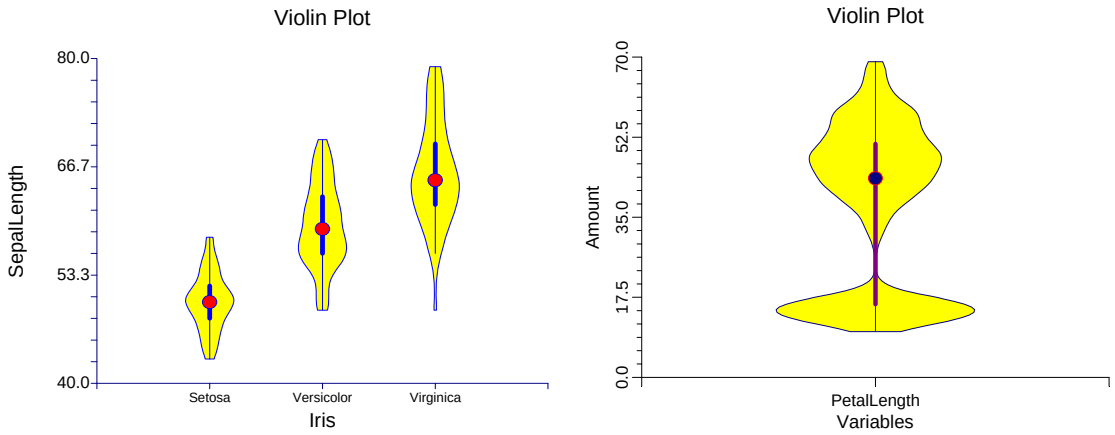
Percentile Plots

Occasionally, you need to display percentiles. ***NCSS*** calculates and displays percentiles between 0 and 100. It lets you assign different shapes, colors, and widths to each percentile group. Using various combinations, you can generate percentile plots that will be tailored to your particular need.



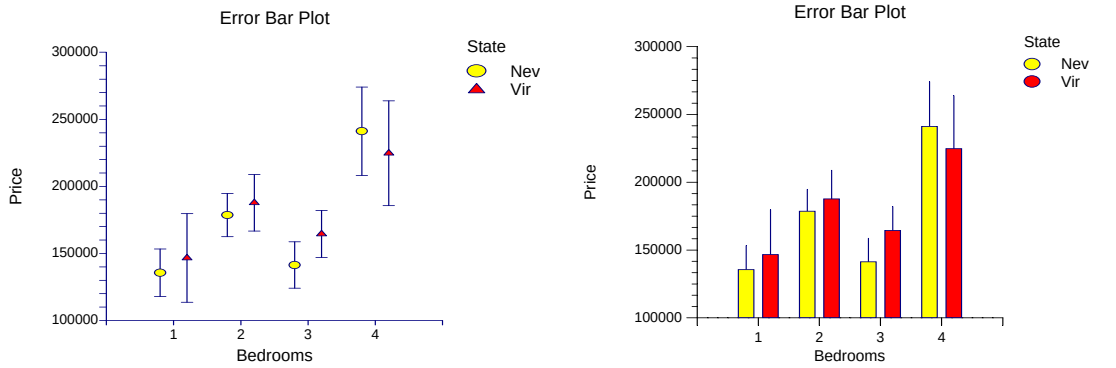
Violin Plots

The box plot is useful for displaying the mean and spread of a set of data. Several box plots may be displayed side by side to allow you to compare the average and spread of several groups. The density trace (or histogram) is useful for displaying the distribution of the data. Unfortunately, several density traces shown side by side are difficult to compare. Yet, comparing the distributions of several batches of data is a common task. We (see Hintze and Nelson 1998) have invented a new plot, which we call the Violin Plot. This plot is a hybrid of the density trace and the box plot. We think that it allows you to compare several distributions quickly.



Error-Bar Charts

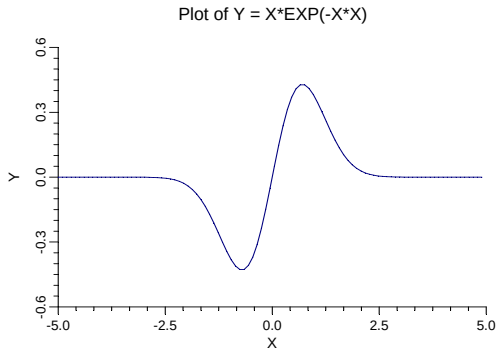
Error-Bar Charts graphically display tables of means and standard errors (or standard deviations).



Two-Variable Charts (Both Continuous)

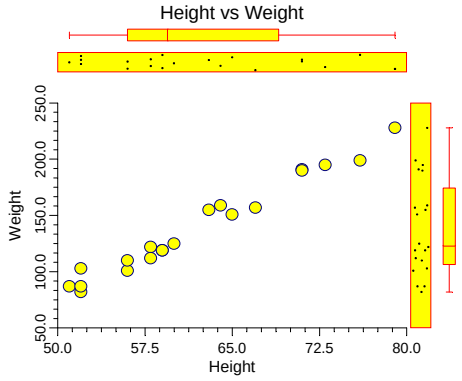
Function Plots

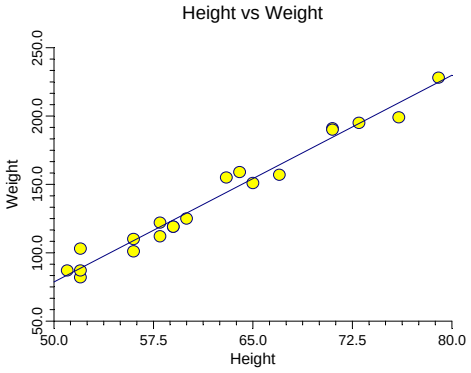
Function plots are graphs of user-specified functions. You define the function using standard mathematical syntax and set the range over which the function should be drawn. This is one of the few procedures that does not accept (or use) data.

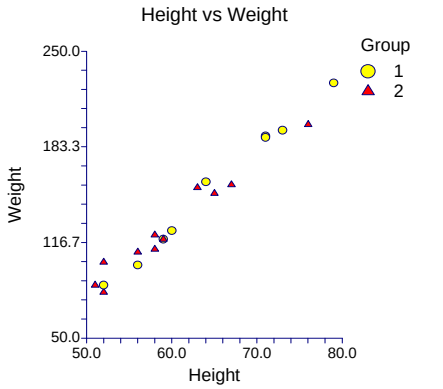


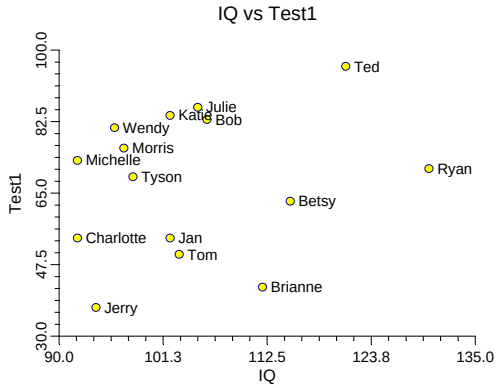
Scatter Plots

The x-y scatter plot is one of the most powerful tools for analyzing data. ***NCSS*** includes a host of features to enhance the basic scatter plot. Some of these features are trend lines (least squares) and confidence limits, polynomials, splines, lowess curves, imbedded box plots, and sunflower plots.



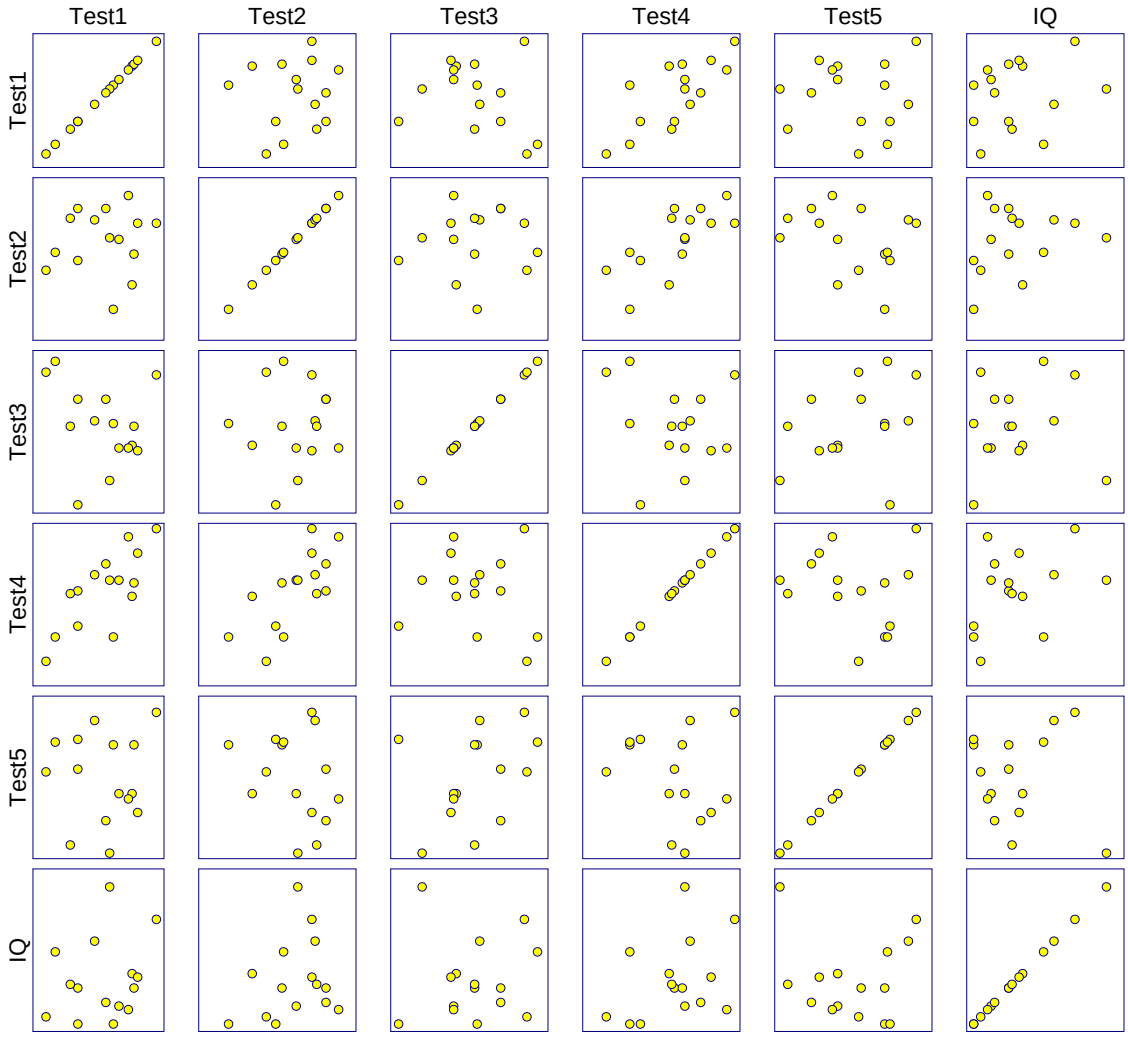






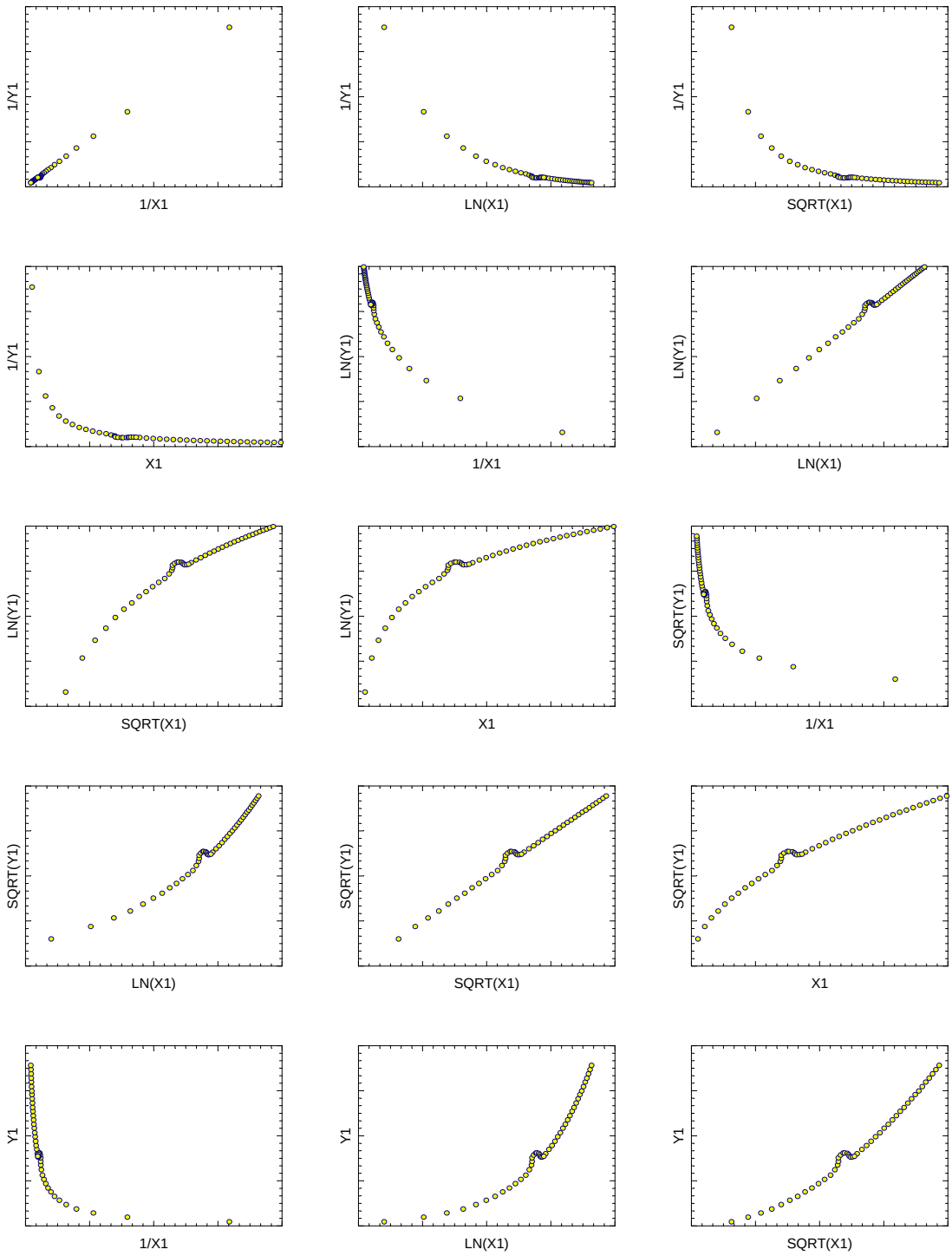
Scatter Plot Matrix

A scatter plot matrix is table of scatter plots. Each plot is small so that many plots can be fit on a page. When you need to look at a lot of plots, such as at the beginning of a multiple regression analysis, a scatter plot matrix is a very useful tool.



Scatter Plot Matrix for Curve Fitting

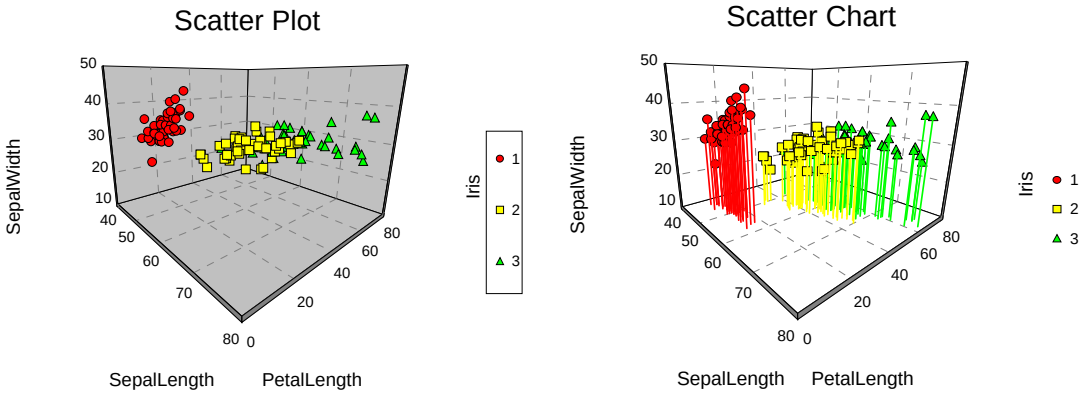
One of the first tasks in curve fitting is to graphically inspect your data. This program lets you view scatter plots of various transformations of both X and Y. These plots are shown in matrix format.



Three-Variable Charts

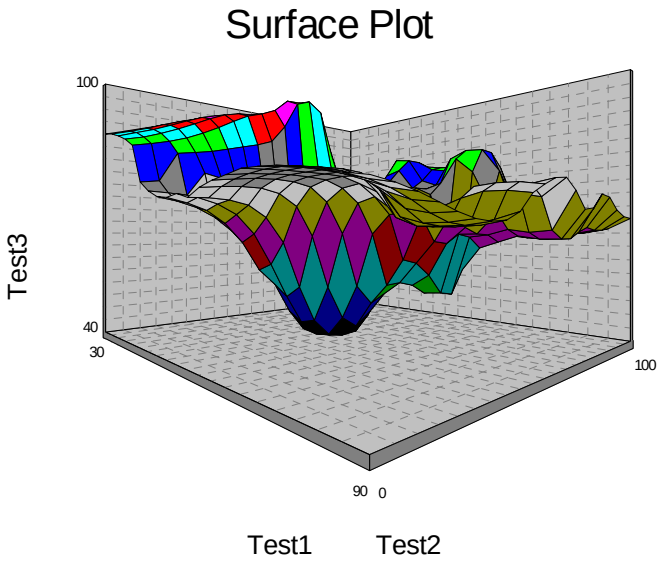
3D Scatter Plots

The 3D scatter plot displays trivariate points plotted in an X-Y-Z grid. It is particularly useful for investigating the relationships among these variables. The influence of a discrete variable may be investigated by using a different plotting symbol for each value of this variable. Hence, up to four variables (three numeric and one discrete) may be displayed on a single graph.



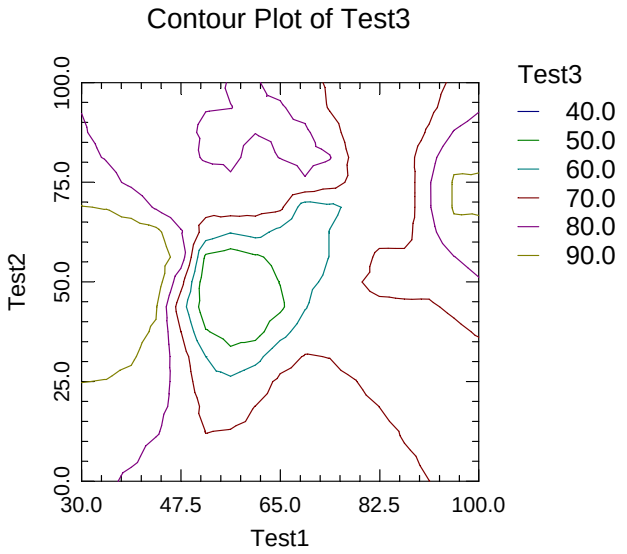
3D Surface Plots

Surface plots are diagrams of three-dimensional data. Rather than showing the individual data points, surface plots show a functional relationship between a designated dependent variable (Y), and two independent variables (X1 and X2). The plot is a companion plot to the contour plot.



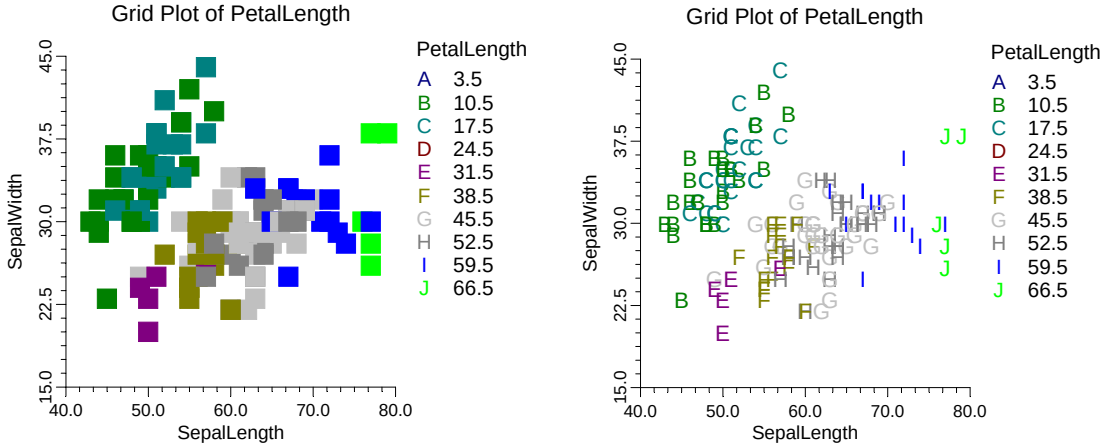
Contour Plots

Contour plots are topographical maps drawn from three-dimensional data. One variable is represented on the horizontal axis and a second variable is represented on the vertical axis. The third variable is represented by isolines (lines of constant value). These plots are often useful in data analysis, especially when you are searching for minimums and maximums in a set of trivariate data. An introduction to contour techniques is contained in Milne (1987).



Grid Plots

The grid plot is a type of contour plot developed for displaying three variables. The first two variables are displayed as in the scatter plot on the vertical and horizontal axes. The third variable is displayed either by the color of the block or by a symbol that is coded from low to high.

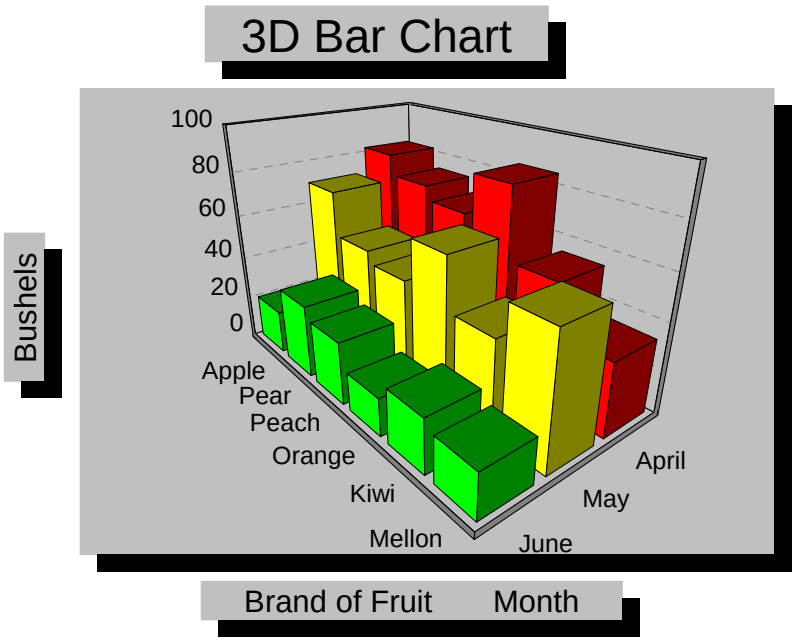


Chapter 141

Bar Charts

Introduction

Bar charts are used to visually compare values to each other. There are several variations on the bar chart. These include the vertical bar chart, horizontal bar chart, area chart, and line chart. There are two and three dimensional versions of each. Below is an example of a 3D bar chart.



Data Structure

Data for a bar chart are entered in the standard row-column format of the spreadsheet. Each numeric data value becomes a bar. Alphabetic data are used to label the rows and columns of the chart. Below is an example of data ready to be charted. These data are stored in the FRUIT database.

FRUIT dataset

Fruit	April	May	June
Apple	82	70	20
Pear	73	50	33
Peach	67	45	28
Orange	85	65	17
Kiwi	54	42	24
Mellon	33	58	20

Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Variables Tab

Specify the variables used to make a chart.

Data

Data Variables

Select the variables to be charted. The data must be numeric, but can be positive or negative. Negative values are displayed as bars going down instead of up.

Data Orientation

The orientation controls whether rows or variables are displayed across the horizontal axis. When *horizontal* orientation is selected, the variables are displayed across the horizontal axis. When *vertical* orientation is selected, each row is displayed separately across the horizontal axis.

Label Settings

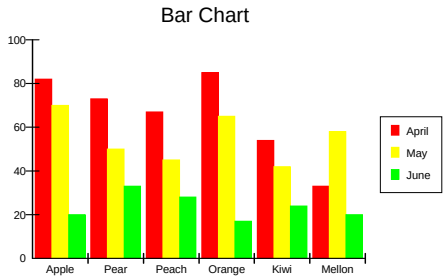
Label Variable

Specify an optional variable containing the labels for individual slices (Data Orientation = Vertical).

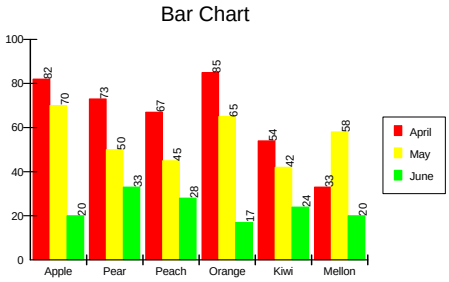
Show Data Labels

This option controls the display of the Data Labels. Note that these labels are only displayed on 2D charts. This option specifies whether labels are to be displayed above the bars.

No



Yes



Labels From First Row of Data

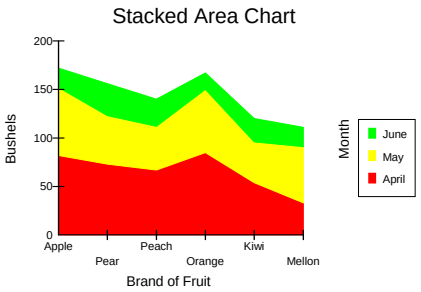
When this option is on (set to Yes) and Data Orientation is set to Horizontal, the values in the first row are used as labels. Otherwise, the variable names are used as labels. When Data Orientation is set to Vertical, the value in the first row may be used in the chart title.

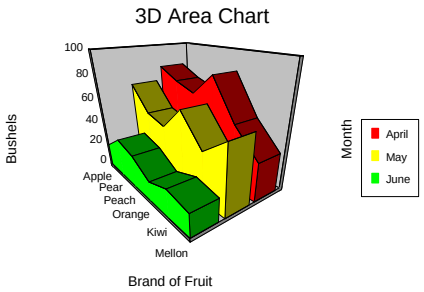
Chart Type and Style

Chart Type

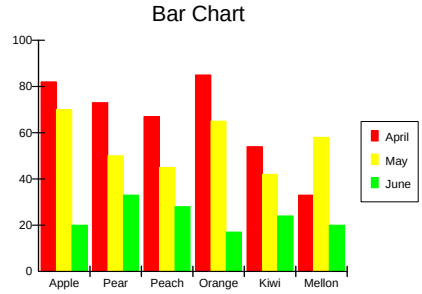
Select the type of chart you want. Possible choices are

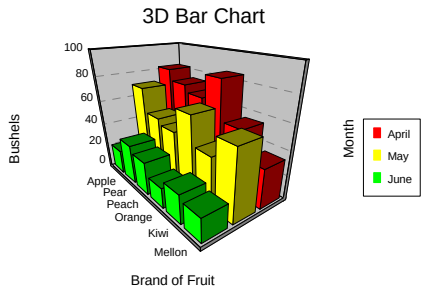
- **Area**



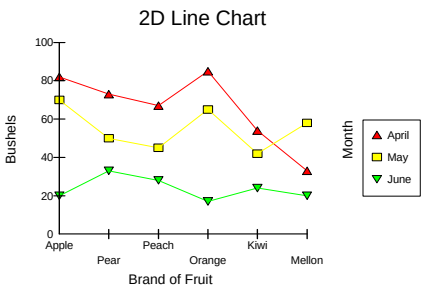


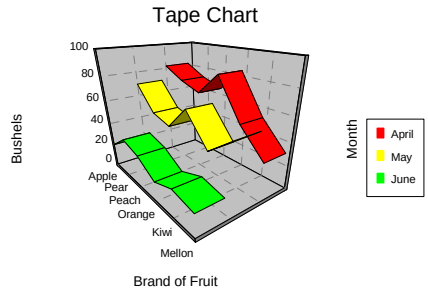
- **Bar**



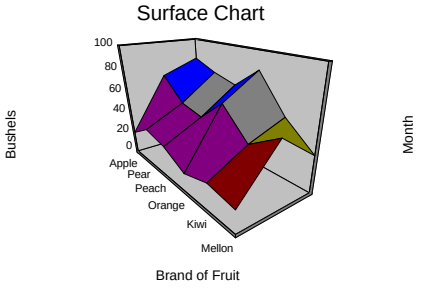


- **Line**





- **Surface**



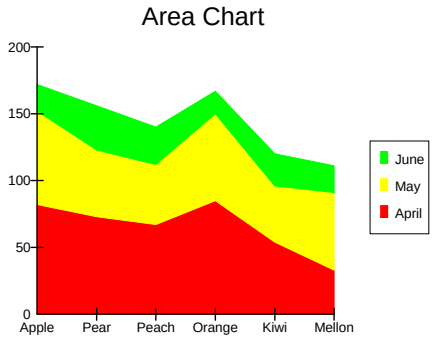
3D (not 2D)

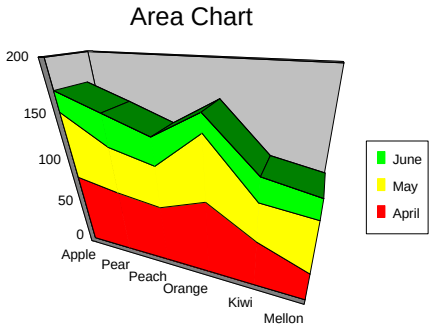
Indicate whether you want the three-dimensional version of the chart.

Style when Chart Type = Area

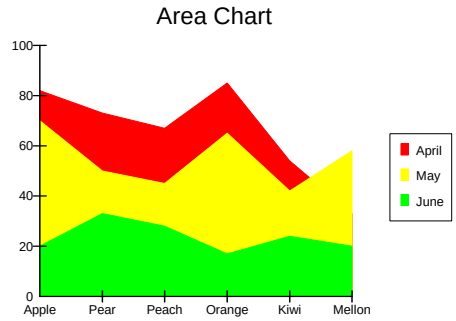
This option is only used if Chart Type is set to Area. Possible styles are

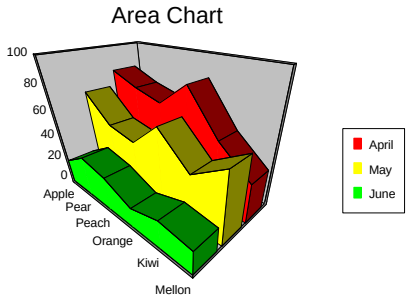
- **Stacked**



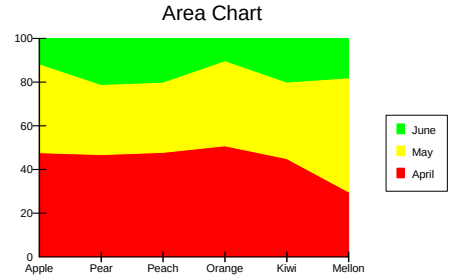


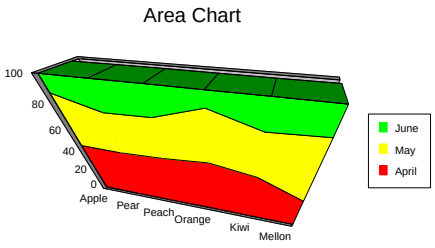
- **Absolute**





- **Stacked Percentage**

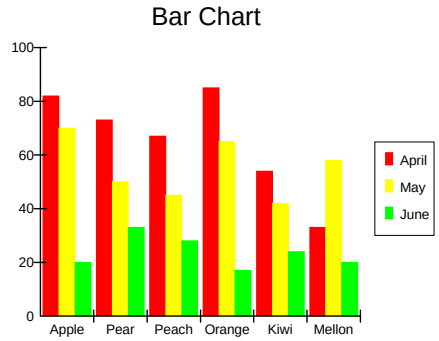


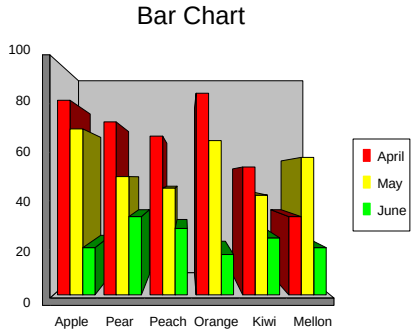


Style when Chart Type = Bar

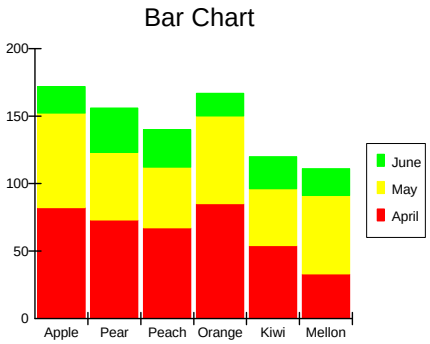
This option is only used if Chart Type is set to Bar. Possible styles are

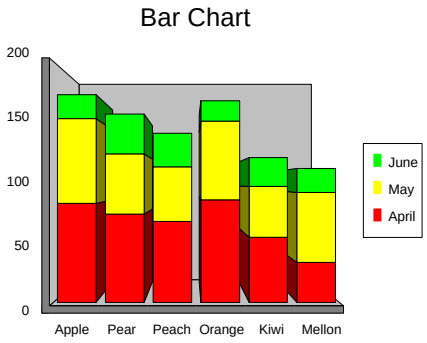
- **Bars**



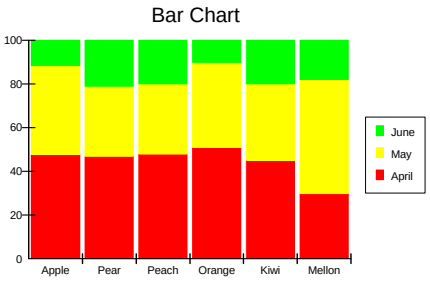


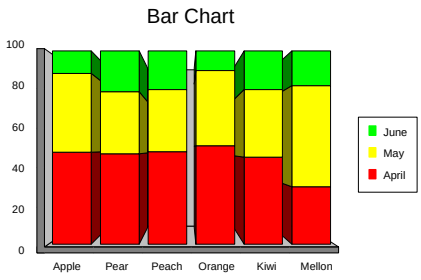
- **Stacked**



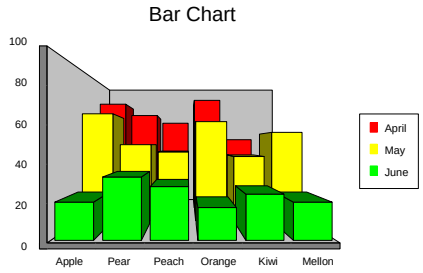


- **Stacked Percentage**





- **Z-Clustered (3D Only)**



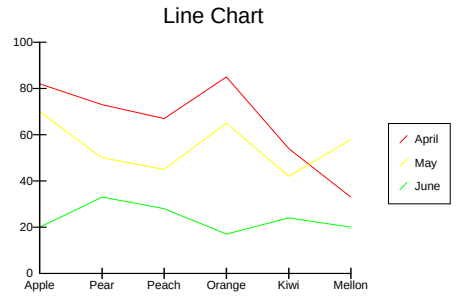
Bar Orientation

Specify whether the bars are to be displayed vertically or horizontally.

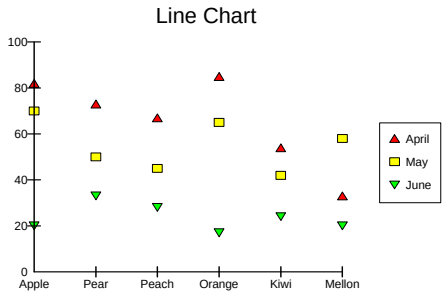
Style when Chart Type = Line

This option is only used if Chart Type is set to Line. Possible styles are

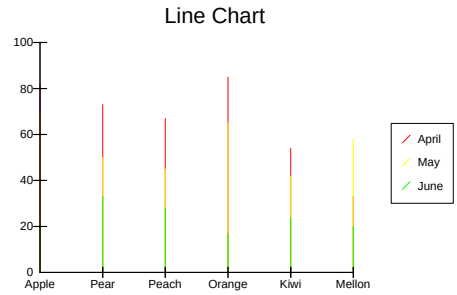
- **Lines Only**



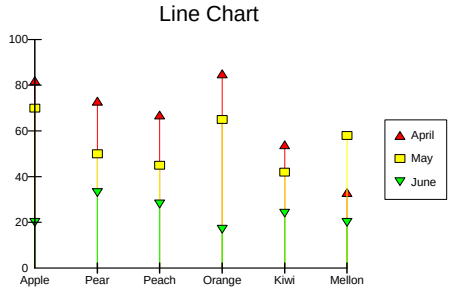
- **Symbols**



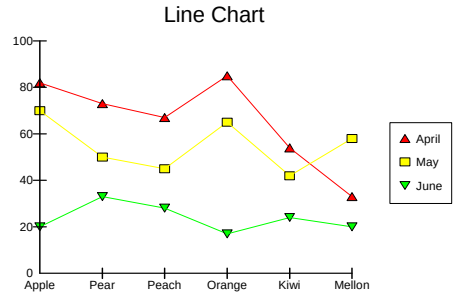
- **Sticks**



- **Sticks and Symbols**

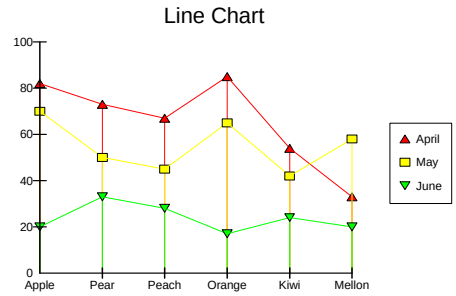


- **Lines and Symbols**

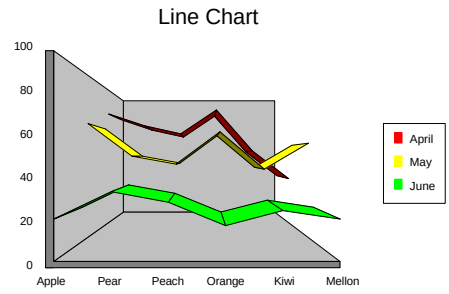


- **Lines and Sticks**

- **Lines, Sticks, and Symbols**



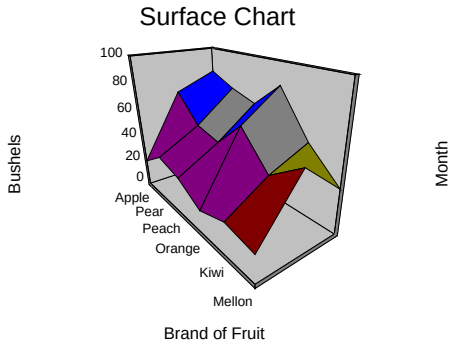
- **Tape (3D Only - Any Style)**



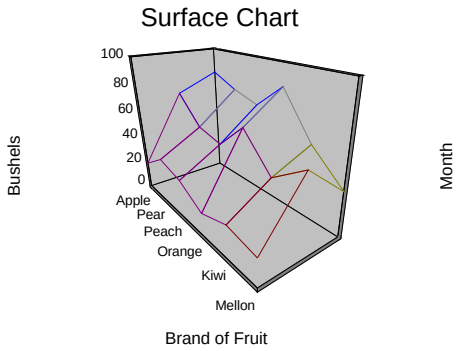
Style when Chart Type = Surface

This option is only used if Chart Type is set to Surface. Possible styles are

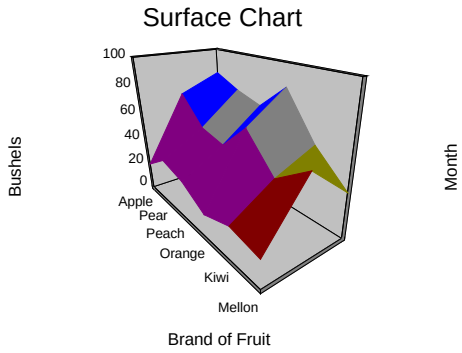
- **Painted, Lines**



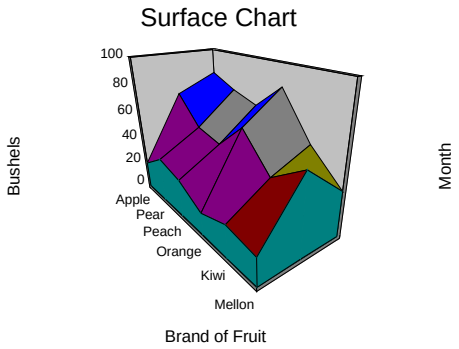
- **Lines**



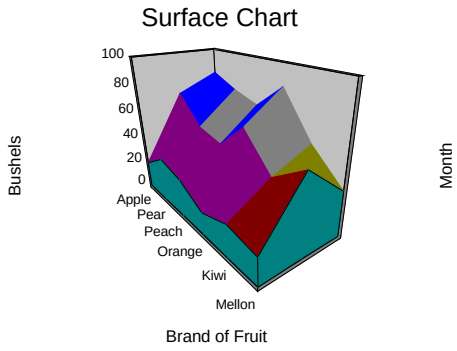
- **Painted**



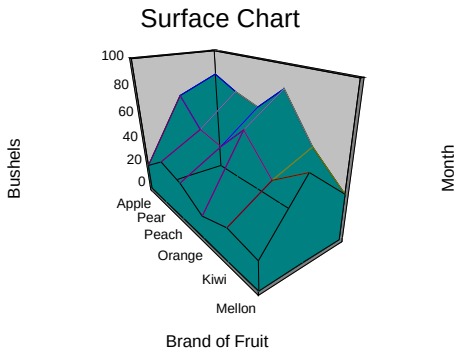
- **Painted, Lines, Wall**



- **Painted, Wall**



- **Lines, Wall**



Interactive Editing

Edit Chart Interactively

Checking this option will cause the interactive graphics editor to be displayed. This allows you to modify the graph interactively at run time. This editor is documented in the corresponding help file. Once you are through editing the bar chart, it will be displayed in the output document.

Once the graphics editor comes up, you can use the scroll-bars on the four sides of the graph to interactively rotate the chart to the viewing position and angle that you like the best.

Axes Tab

These options control the appearance and position of the axes.

Y (2D Vertical) Axis

Axis Position

This option sets the position of the vertical axis as either on the left or the right in 2D charts.

User Defined Scale (Use Min and Max)

This option specifies whether the vertical (Y) axis is scaled automatically (not checked) or from user specified maximum, minimum, and number of tick marks (checked).

Minimum

Sets the value of the vertical minimum. This value may be negative. This value must be smaller than the smallest data value. This value is only used when the User Defined Scale option is checked.

Maximum

Sets the value of the vertical maximum. This value must be greater than the largest data value. This value is only used when the User Defined Scale option is checked.

Number of Major Ticks

Sets the number of tick marks from the origin. This value is only used when the User Defined Scale option is checked.

Show Major Ticks

Indicates whether to display tick marks along the Y (vertical) axis.

Show Minor Ticks

Specify whether to display the minor tick marks. This option works with 2D graphs only.

Show Horizontal Grid

Specify whether to display the horizontal grid lines.

X (2D Horizontal) Axis

Label Frequency

This option sets the label frequency of the horizontal axis. Setting this equal to 1 displays every label. Setting this equal to 2 displays every other label.

Tick Frequency

Specify the frequency of the X (horizontal) axis tick marks.

Show X Ticks

Indicates whether to display tick marks along the X (horizontal) axis.

Show Vertical Grid (From X Axis)

Specify whether to display the vertical grid lines.

Both Axes

Grid Line Style

Specify the style (line, dots, dashes, etc.) of the grid lines.

Axis Color

Set the color of the axis lines.

Grid Color

Set the color of the grid lines.

Both Axes – 3D Only

Thin Walls

Indicate whether the axis borders of 3D charts should be thick or thin.

Cage Edge Color (Thin Walls Unchecked)

Set the color of the cage edge in 3D charts.

Cage Wall Color

Set the color of the cage wall in 3D charts.

Titles & Background Tab

These options control the titles that may be placed on all four sides of the chart.

Titles

Chart (Top) Title

This option gives the text that will appear in the title at the top of the chart. The color, font size, and style of the text is controlled by the options to the right.

Bottom Title

This option gives the text that will appear in the title at the bottom of the chart. The color, font size, and style of the text is controlled by the options to the right.

Left Title

This option gives the text that will appear in the title at the left of the chart. This may also serve as a label. The color and orientation of the text is controlled by the options to the right.

Right Title

This option gives the text that will appear in the title at the right of the chart. This may also serve as a label. The color and orientation of the text is controlled by the options to the right.

Variable Names (May be used in Titles of 3D Charts)

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Background Colors and Styles

Entire Graph

Specify the background color and style of the entire chart.

Inside Graph

Specify the background color and style of the chart itself (within the axes).

Graph Title

Specify the background color and style of the chart title.

Left Title

Specify the background color and style of the left title.

Right Title

Specify the background color and style of the right title.

Bottom Title

Specify the background color and style of the bottom title.

Tick Labels & Legend Tab

This panel controls the color and size of the reference labels.

Tick Labels

Text Color

This option sets the color of the reference items along the two axes.

Font Size

This option sets the font size of the reference items along the two axes.

Text Rotation

This option sets the display angle of the reference items along the horizontal axis.

Bold and Italics

This option sets the style of the reference items along the two axes.

Data Labels

Color

This is the color of the data labels displayed at the top of the bars.

Legend

Position

This option sets the position of the legend around the chart. Note that if you choose a position in which the full text of the legend cannot be fit, the legend will not be displayed.

Percent of Vertical Space

Specify the size of the legend as a percentage of the maximum possible. This option lets you shrink a legend that is too large.

Text Color

Specify the color of the legend text.

Background

Specify the background color of the legend.

Font Size

Specify the size of the legend text.

Bold and Italics

Specify the style of the legend text.

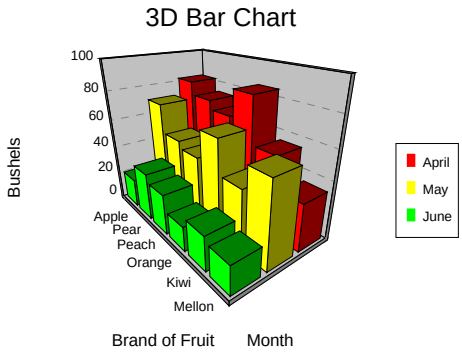
Legend Background Style

Specify the background style of the legend.

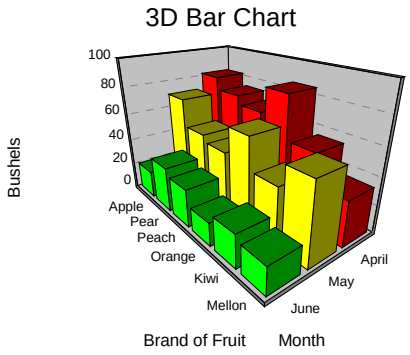
Show Legend as Labels

This option only applies to 3D charts. Selecting Yes will disable the Legend and show the legend labels along the z-axis. See below.

No



Yes



Color as Labels

Normally, text in the legend is displayed using the color selected by the Text Color option. This option indicates that each legend entry is to be displayed in the corresponding group color.

3D Options Tab

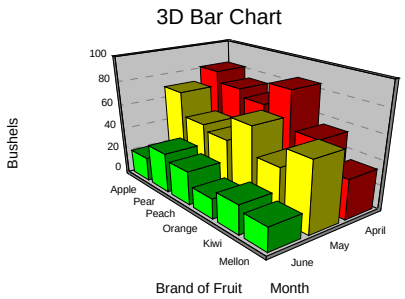
These options control the viewing position of the 3D charts. These options may be set interactively by checking the Edit Chart Interactively option and then activating the four scroll bars on the sides of the Graphics Editor window.

Whole Chart

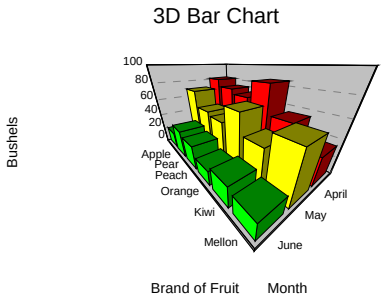
Perspective

This option specifies the perceived distance from which the graph is viewed. The range is from 0 to 100. As the value gets large, the distance gets smaller. A setting of 50 sets the viewing distance at about twice the graph's width. A setting of 100 sets the viewing distance at about equal to the graph's width.

Perspective = 10



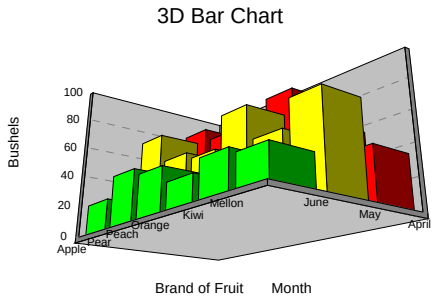
Perspective = 90



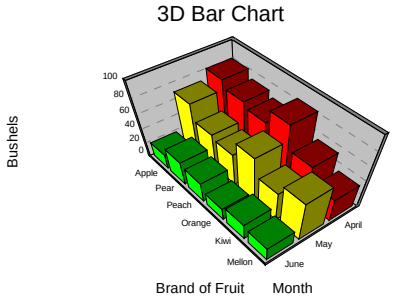
Elevation

This option sets the vertical viewing angle (in degrees). The setting represents an angle above or below a point halfway up the graph. The range is from -60 to 90.

Elevation = -20



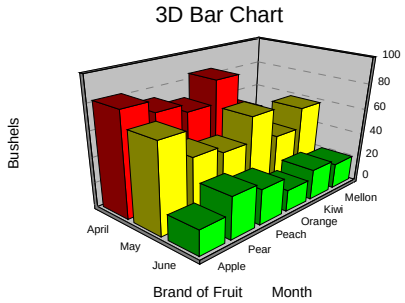
Elevation = 50



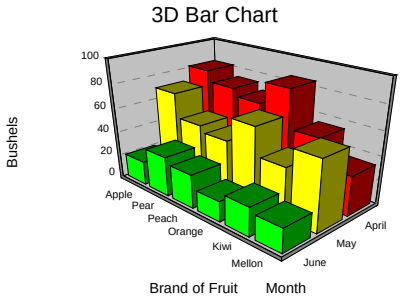
Rotation

This option sets the horizontal viewing angle (in degrees). The setting represents an angle around the base of the graph. The range is from -180 to 180.

Rotation = -45



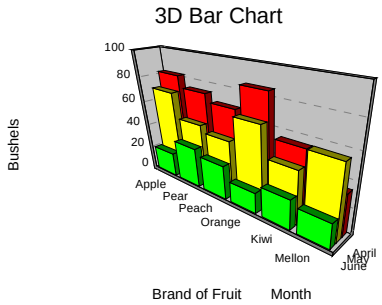
Rotation = 45



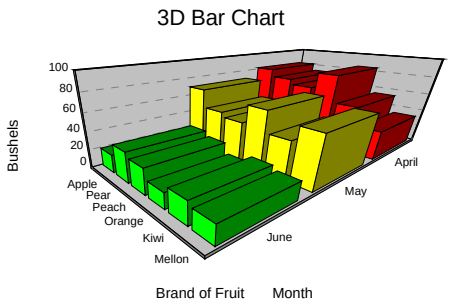
Depth

This option sets the width in the Z direction. The range is from 1 to 1,000.

Depth = 25



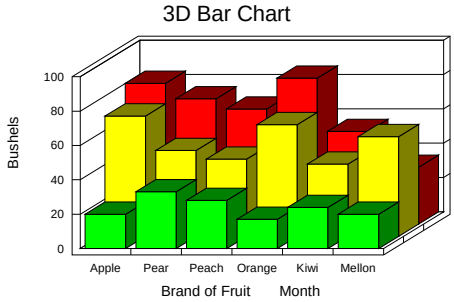
Depth = 400



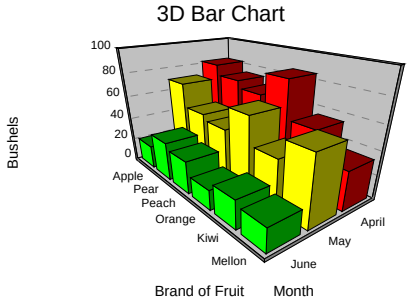
Projection Method

This option specifies the method used to determine viewer position and angle in a 3D graph.

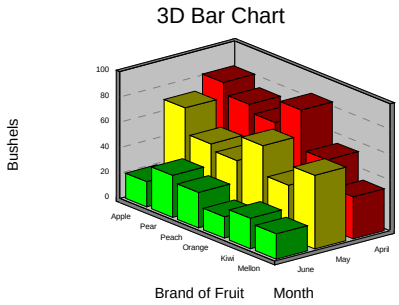
Off



Perspective



Isometric

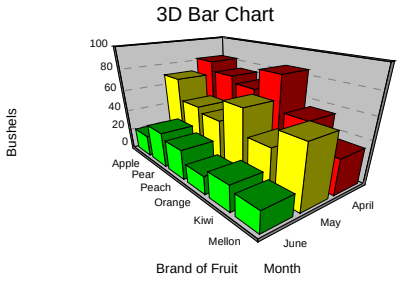


Bar Chart Type

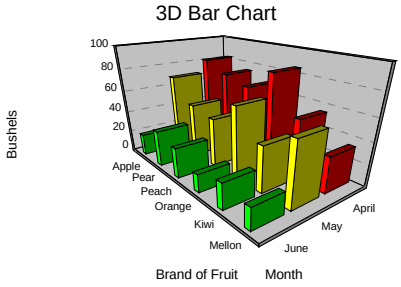
Gap Between Bars

This option specifies the gap between bars in the X (horizontal) direction in a 3D graph. The range is from 0 to 95.

Gap = 19



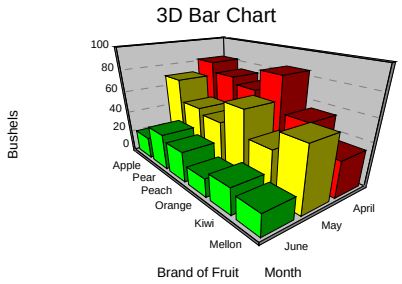
Gap = 95



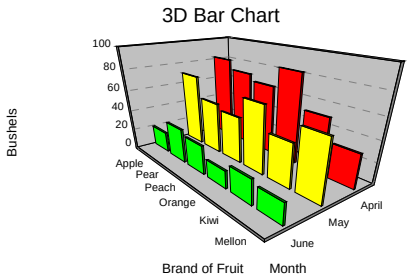
Gap Between Sets of Bars

This option specifies the gap between bars in the Z direction in a 3D graph. The range is from 0 to 95.

Gap = 19



Gap = 95



Surface Chart Type

Surface Color Palette

Specify a color palette for the surface chart. The 128-color palettes may only be used on machines with at least SuperVGA capabilities. Using a setting here of, for example, Black to Red will allow the surface plot to show a continuous array of red hues from lowest to highest.

Surface Color Min

Specifies the number of the color to be associated with the lowest numerical value. Possible values are 32 to 127. A value near 50 usually works well. Note that this option only works with 128-color palettes.

Surface Color Max

Specifies the number of the color to be associated with the highest numerical value. Possible values are 32 to 127. A value near 70 usually works well. Note that this option only works with 128-color palettes.

Surface Wall Color

Specifies the color of the border wall around the surface plot.

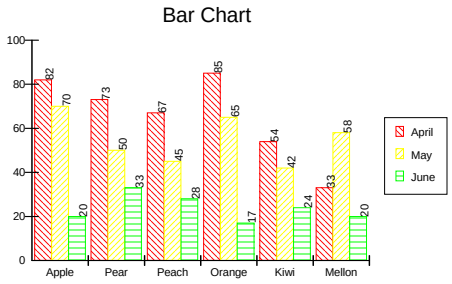
Bars Tab

These options specify the colors and patterns used to display the bars.

Bar 1 - 15

These options let you specify the color and display pattern of the bars. The first bar is associated with the first group (row or variable), the second bar with the second group, and so on. If more than fifteen bars definitions are needed, they are reused so that group 16 = group 1, group 17 = group 2, and so on.

Here is a bar chart that uses three different bar patterns:



Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Bar Chart

This section presents an example of how to create a bar chart of the data stored on the FRUIT database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Bar Charts window.

1 Open the FRUIT dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FRUIT.s0**.
- Click **Open**.

2 Open the Bar Charts window.

- On the menus, select **Graphics**, then **Other Charts and Plots**, then **Bar Charts**. The Bar Charts procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

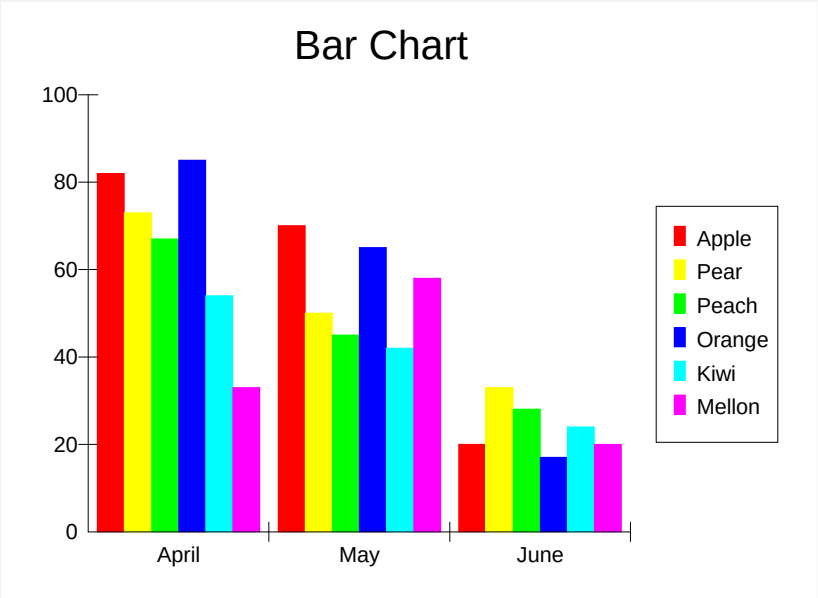
3 Specify the variables.

- On the Bar Charts window, select the **Variables tab**.
- Double-click in the Data Variables text box. This will bring up the variable selection window.
- Select **April, May, June** from the list of variables and then click **Ok**. "April-June" will appear in the Data Variables box.
- Double-click in the **Label Variable** text box. This will bring up the variable selection window.
- Select **Fruit** from the list of variables and then click **Ok**. "Fruit" will appear in the Label Variable box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Bar Chart Output



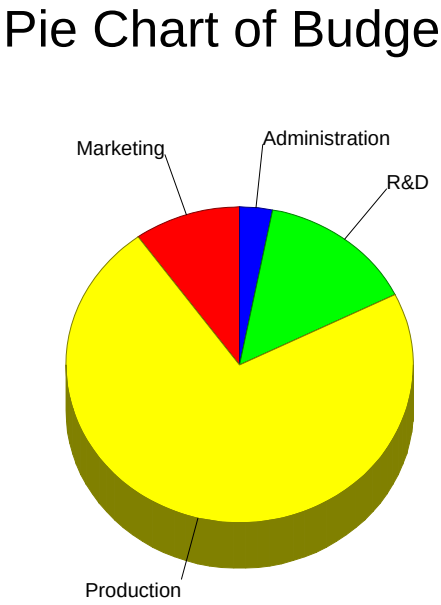
Chapter 142

Pie Charts

Introduction

The pie chart is constructed by dividing a circle into two or more sections. The chart is used to show the proportion that each part is of the whole. Hence, it should be used when you want to compare individual categories with the whole. If you want to compare the values of categories with each other, use a bar chart or scatter plot.

For example, the chart below shows the budget for each of four departments in a hypothetical company.



Pie charts are useful for displaying up to about six or seven slices.

Data Structure

Data values must be positive and numeric. Non-positive values are omitted. The data may be entered in either of two possible formats, vertical or horizontal. In the vertical format, a variable (column) contains the values to be charted. In the horizontal format, a row contains the values to be charted. These data are contained in the PIE database.

PIE dataset

Budget Data (Vertical Format)

Department	Budget
Marketing	170
Production	1239
R&D	250
Administration	52

Budget Data (Horizontal Format)

Marketing	Production	R & D	Administration
170	1239	250	52

Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Variables Tab

This panel specifies the variables that will be used in the analysis.

Data

Data Variables

Select the variables to be plotted. See Data Orientation for a full discussion of what happens with multiple rows or multiple variables.

Data Orientation

Specify either Horizontal or Vertical data input orientation. The options are

- **Horizontal**

  With horizontal orientation, each row generates a separate pie chart. The slices are made up of the variables. If you specify five variables, the pie chart will have five slices.

  If Labels From First Row is checked, the first row will be used as slice labels. The Label Variable is not used.

- **Vertical**

  With vertical orientation, each variable generates a separate pie chart. The slices are made up of the values in the rows of the variable. If the specified variable contains five rows, the pie chart will have five slices.

  If Labels From First Row is checked, the first row will be omitted. The Label Variable is used to specify slice labels.

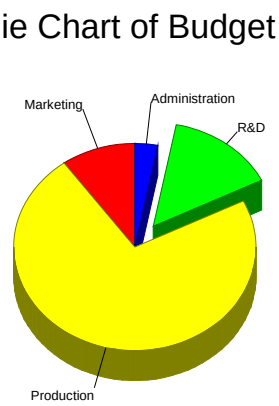
Chart Settings

3D (not 2D)

This option lets you specify whether to create a two-dimensional (flat) pie chart or a three-dimensional (raised) pie chart.

Explode (Emphasize) Slices

One or more slices may be emphasized by moving them from the rest of the pie. This may be seen from the following example.



The slice(s) to be exploded are indicated by an array of 0's and 1's. The 1's indicate the slices that are to be exploded. The entries in the array are separated by commas. For example, suppose you have a pie chart with six slices and you want to explode the second and the fourth slices. You would enter **0 1 0 1**. Note that any remaining slices not accounted for by this line are not exploded. Hence, if you just wanted to explode the first slice, you could enter only a 1.

Label Settings

Label Variable

Specify an optional variable containing the labels for individual slices (Data Orientation = Vertical).

Show Data Labels (2D Only)

Specify whether to show the slice labels.

Labels From First Row of Data

When this option is checked and Data Orientation is set to Horizontal, the values in the first row are used as labels. Otherwise, the variable names are used as labels. When Data Orientation is set to Vertical, the value in the first row is used in the chart title.

Colored Labels

The slice labels may be colored a single color (usually black) or they may be colored using the corresponding slice color.

Connect Labels to Pie

Specify whether to show the line between the slice labels and the slice.

Show Percentages with Labels

Specify whether to show the percentage amount that each slice is of the whole as part of the slice label.

Interactive Editing

Edit Chart Interactively

Checking this option will cause the interactive graphics editor to be displayed. This allows you to modify the graph interactively at run time. This editor is documented in the corresponding help file. Once you are through editing the bar chart, it will be displayed in the output document.

Once the graphics editor comes up, you can use the scroll-bars on the four sides of the graph to interactively rotate the chart to the viewing position and angle that you like the best.

Pie Slice Colors and Styles

These options specify the patterns used to display up to fifteen slices. These can be especially useful for printing on black and white printers.

Slices 1-15

These options control the color and style of each slice.

Titles and Miscellaneous Tab

These options control the title and labels that may be placed on the pie chart.

Titles

Chart (Top) Title

This option specifies the top title for the chart. Note that the {L} is replaced by an appropriate label depending on the Data Orientation. This label will be the variable name, the row label, or the entry from the first row.

Bottom Title

This option specifies an option bottom title for the chart. Note that the {L} is replaced by an appropriate label depending on the Data Orientation. This label will be the variable name, the row label, or the entry from the first row.

Color

These options set the colors of the titles.

Font Size

These options set the font size of the titles.

Bold and Italics

These options set the font style (**bold** and *italics*) of the titles.

Variable Names

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Labels

Text Color

This option sets the color of the slice labels.

Font Size

This option sets the font size of the slice labels.

Bold and Italics

This option sets the font style (**bold** and *italics*) of the slice labels.

Background Colors and Styles

Entire Chart

This option sets the background color of the graph.

Graph Title

These options set the background color and style (plain, framed, raised, etc.) of the graph title.

Bottom Title

These options set the background color and style (plain, framed, raised, etc.) of the bottom title.

Inside Chart

These options set the background color and style (plain, framed, raised, etc.) of the inside area of the chart.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Pie Chart

This section presents an example of how to create a pie chart of the data stored on the PIE database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Pie Charts window.

1 Open the PIE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **PIE.s0**.
- Click **Open**.

2 Open the Pie Charts window.

- On the menus, select **Graphics**, then **Other Charts and Plots**, then **Pie Charts**. The Pie Charts procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

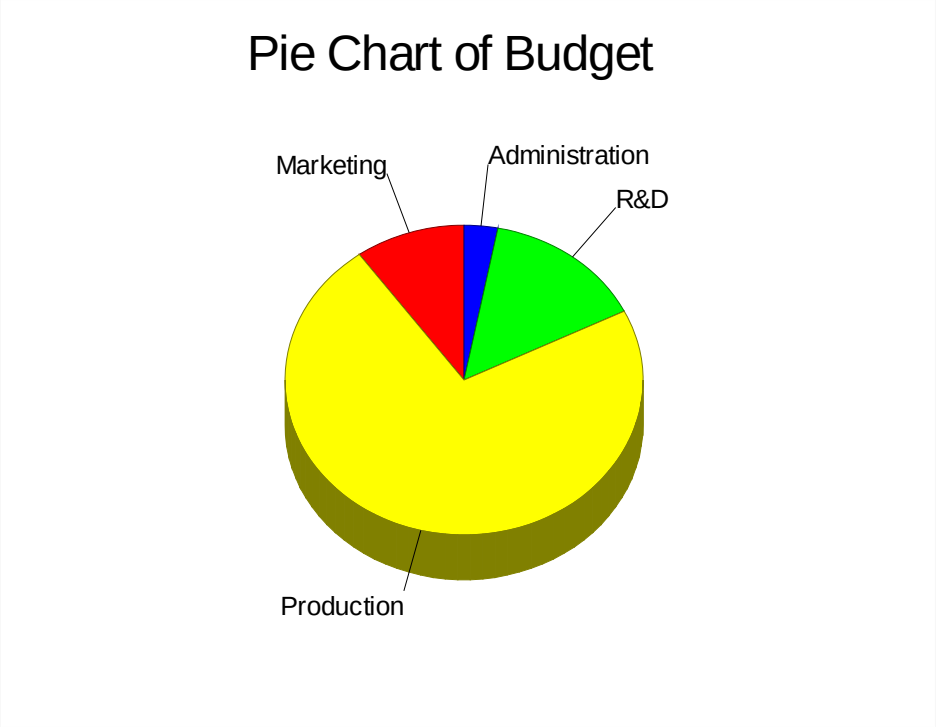
3 Specify the variables.

- On the Pie Charts window, select the **Variables tab**.
- Double-click in the **Data Variables** text box. This will bring up the variable selection window.
- Select **Budget** from the list of variables and then click **Ok**. "Budget" will appear in the Data Variables box.
- Double-click in the **Label Variable** text box. This will bring up the variable selection window.
- Select **Department** from the list of variables and then click **Ok**. "Department" will appear in the Label Variable box.
- Select **Vertical** in the **Data Orientation** list box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Pie Chart Output



Chapter 143

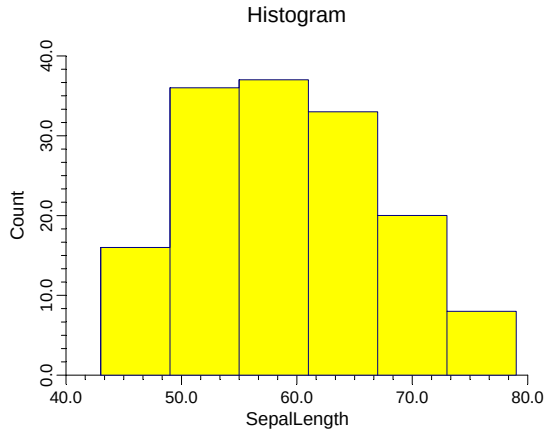
Histograms

Introduction

The word *histogram* comes from the Greek *histos*, meaning pole or mast, and *gram*, which means chart or graph. Hence, the direct definition of "histogram" is "pole chart." Perhaps this word was chosen because a histogram looks like a few poles standing side-by-side.

A histogram is used to display the distribution of data values along the real number line. It competes with the probability plot as a method of assessing normality. Humans cannot comprehend a large batch of observations just by reading them. To interpret the numbers, you must summarize them by sorting, grouping, and averaging. One method of doing this is to construct a *frequency distribution.* This involves dividing up the range of the data into a few (usually equal) intervals. The number of observations falling in each interval is counted. This gives a frequency distribution.

The *histogram* is a graph of the frequency distribution in which the vertical axis represents the count (frequency) and the horizontal axis represents the possible range of the data values.



Density Trace

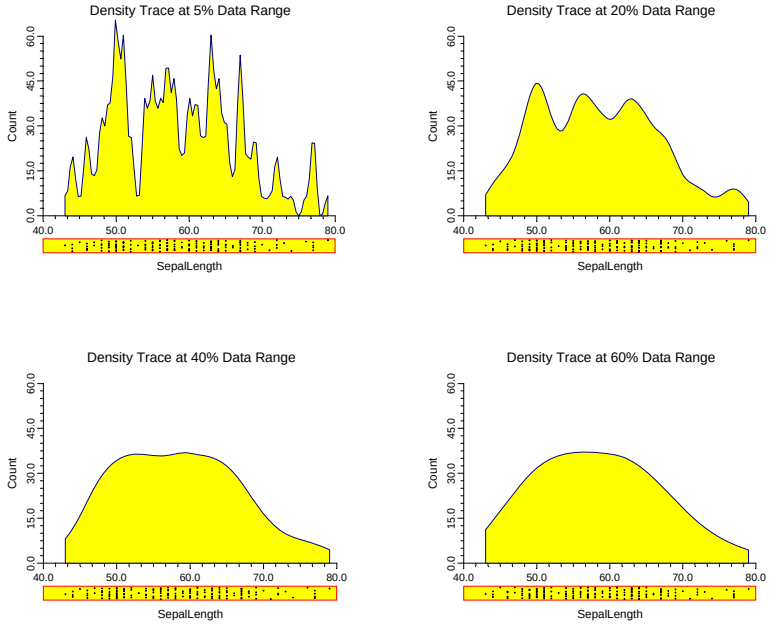
The histogram is widely used and needs little explanation. However, it does have its drawbacks. First, the number and width of the intervals are a subjective decision, yet they have a high impact on the look of the histogram. Slightly different boundary values can give dramatically different looking histograms. (You can experiment with **NCSS** to see the impact of changing the number of bins on the look of the histogram.)

Another problem with the histogram is that the rectangles make it appear that the data are spread uniformly throughout the interval. But this is often not the case. Also, the “skyscraper” look of the histogram doesn’t resemble the rather smooth nature of the data’s distribution.

These complaints against the histogram have brought many new innovations. One of the newest and most popular display techniques for showing the distribution of data is the density trace.

Density refers to the relative frequency (concentration) of data points along the data range. Mathematically, the density at a value x is defined as the fraction of data values per unit of measurement that lie in an interval centered at x. Once you pick a suitable interval width, you can calculate the density at any (and every) x value. If you calculate the density at, say, 50 values and connect them, you’ll have a density trace.

In **NCSS**, the interval width is specified as a percentage. As you increase the percentage, you increase the amount of data included in each density calculation. This increases the smoothness of the chart. The following four density traces were made of the same data at increasing percentage smoothness. Note how much more appealing these charts are than the histogram.



As the interval width is increased, data points further and further from the center value are included. In order to decrease the weight of points that are far removed from the center value, we use a weighting scheme that weights points proportionally to their distance from the center value. The weight function used is half the cosine function with its peak at the center value. It decreases symmetrically to zero, after which a weight of zero is applied. Hence, points have a smaller and smaller impact on the density trace as they are further and further from the center.

Another way to think of the density trace is to imagine that you construct 1000 histograms of the same data using slightly different boundary positions and take the average rectangle height at each of 50 values along the data range. This would give you a smoothed histogram that has many of the same properties of the density trace. Hence, the density trace should be thought of as a smoothed histogram in which interval width and number of bins do not come into play.

Data Structure

A histogram is constructed from a single variable. A second variable may be used to divide the first variable into groups (e.g., age group or gender). No other constraints are made on the input data. However, the distributions available in **NCSS** assume that the data are continuous. Note that rows with missing values in one of the selected variables are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the histogram.

Variables

Variable(s)

Select one or more variables. If more than one variable is entered, the values may be combined into one histogram or separated into separate histograms depending on the *Combine all variables as one* option.

Grouping Variable

This variable may be used to separate the observations into groups. A separate histogram is created for each unique value of this variable.

Combine all variables as one

When checked, the values for all selected variables are combined into one histogram. This option cannot be used when a Grouping Variable is specified.

Specify Number of Bars Using

The number of bars shown on the histogram is designated by either the Number of Bars option or the Bar Width option. If the Bar Width option is blank, the Number of Bars is used; otherwise, the Bar Width is used.

Number of Bars

Specifies the number of bars (bins, slices, or intervals) displayed in your histogram. This option is only used if the Bar Width option is blank.

Bar Width

This is the width of the bars in the terms of the data values. It is used in conjunction with the Data Range Minimum and Data Range Maximum to determine the number of bars.

Specify Number of Bars Using – Data Range

Minimum

This is the minimum data value displayed on the histogram. Rows with values smaller than this are omitted. If left blank, it is calculated from the data.

Maximum

This is the maximum data value of the histogram. Rows with values larger than this are omitted. If left blank, it is calculated from the data.

Histogram

Display Type

Indicates whether to fill the bars or overlay only their outline. This option is useful when you want to overlay the histogram outline on a density trace. You can also use this option to completely omit the histogram.

Cumulative Scale

Checking this option causes the program to display the cumulative frequencies instead of the individual frequencies.

Outline

Designates the color, line width, and line pattern of the outline of the histogram's bars.

Histogram – Bar Fill Color

Color

This option specifies the interior color of the histogram bars.

Histogram – Bar Outline

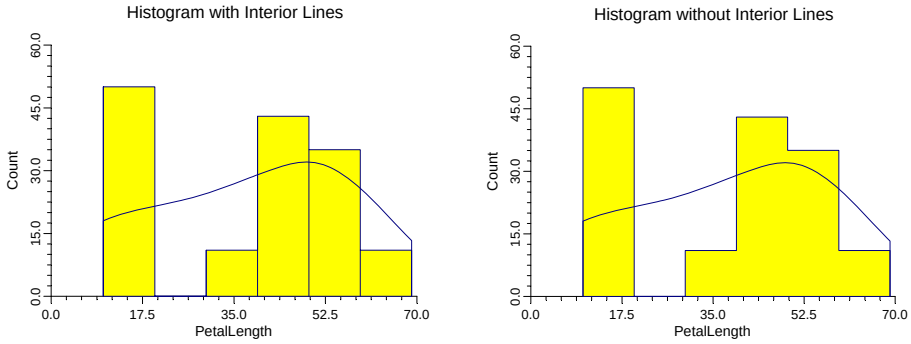
Outline Color, Width, and Pattern

Designates the line color, line width, and line pattern of the outline of the histogram's bars.

Histogram – Interior Lines

Interior Lines

Use this option to omit the internal outline lines that are normally drawn around each bar.



Axes Tab

These options specify the characteristics of the vertical and horizontal axes.

Vertical Axis

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Maximum

Specifies the largest value shown on this axis. Note that the minimum is always set to zero.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Scale Type

Specifies whether the vertical scale is displayed as a Count or a Percentage.

Vertical Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks and grid lines displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

Tick labels are displayed for the minor tickmarks. This option specifies the number of minor tickmarks and grid lines displayed between each set of major tickmarks.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the minor grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the reference numbers that are displayed along this axis. The following options are available in this window:

Decimals

Specifies the number of decimal places displayed in the reference numbers.

Font Size

Specifies the size of the reference values.

Color

Specifies the color of the reference values.

Bold, Italic, Underline

Specifies the font style of the reference values.

Text Rotation

Specifies whether the reference values are displayed vertically or horizontally.

Max Characters

The maximum length (number of characters allowed) of a reference value. This field shifts the axis label away from the axis to make room for the reference value. Hence, if your reference numbers are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

Vertical Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the right side, the left side, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the right side, the left side, on both sides, or not displayed.

Tick Labels

This option controls the position of the tick labels: whether they are placed on the right side, the left side, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Horizontal Axis

These options specifies the characteristics of the horizontal axis.

Label Text

This box supplies the horizontal axis label. The characters {Y} and {G} are replaced by the appropriate variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Location of Bar Labels

This option specifies where the tick labels are placed along the horizontal axis.

Standard

Selecting this option indicates a typical placement of the tick labels. Note that the numbers do not necessarily match the bars.

Mid Points

One tick label is placed at the middle of each bar.

End Points

One tick label is placed at the end of each bar.

Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.
- **Font Size**
  Specifies the size of the tick labels.
- **Color**
  Specifies the color of the tick labels.
- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.
- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Max Characters

The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

Horizontal Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the bottom, the top, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tick Labels

This option controls the position of the tick labels: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Variable Data Transformations

These options specify automatic transformations of the data.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

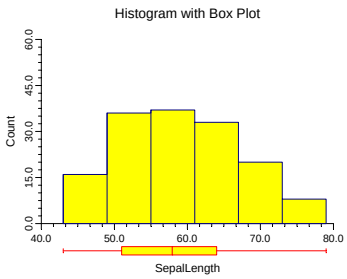
Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Box and Dot Plots Tab

Box Plot

A box plot may be placed above or below the histogram.



Shape

This option specifies the shape of the box plot. See the *Box Plot* chapter for further details.

Percentile Type

Specifies the formula used to calculate the percentile. See the *Box Plot* chapter for further details.

Reference Lines

Reference lines may be extended across the histogram at each of the three quartiles that are used to form the box.

Inner and Outer Fences

The constants used to construct the fences. See the *Box Plot* chapter for further details.

Position

This option indicates on which side of the histogram the box plot should be displayed.

Fill Color

This option specifies the interior color of the box plot.

Line Color

This option specifies the color of the box border and the lines.

Line Width

This option specifies the width of the border line.

Box Width

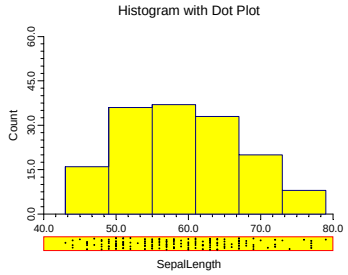
This option specifies the width of the box itself.

Show Outliers

This option indicates whether the outlying points should be displayed.

Dot Plot

A dot plot may be placed in the vertical margin of the plot. The dot plot lets you study the distribution of the data values.



Position

Specifies whether to display the dot plot and in which margin to place it in.

Box Width

Specifies the width of the dot plot.

Fill Color

The color of the data point interior (fill region). Many of the plotting symbols, such as a circle or square, have an interior region and a border. This specifies the color of the interior region.

Dot Color and Size

Specifies the color and size of the data points displayed in the dot plot.

Line Width and Color

Specifies the width and color of the dot plot’s border.

Traces and Lines Tab

These options allow certain reference lines to be specified and displayed.

Trace and Line Overlays – Density Trace

This set of options controls the appearance of the optional density trace. A density trace may be placed on, or replace, the histogram. The details of how a density trace is constructed were presented at the beginning of this chapter.

Display Type

Indicates whether to fill the density trace or overlay only its outline. This option is useful when you want to overlay the density trace outline on a histogram. You can also use this option to completely omit the density trace.

Number of Points

Specifies the number of density trace points to be calculated along the horizontal axis. This adjusts the resolution of the density trace.

Percent of Data in Calculation

The percent of the data used in the density calculation. Select 0 for an automatic value determined from your data. A low value (near 10%) will give a rough plot. A high value (near 40%) will yield a smooth plot.

Outline Color, Width, and Pattern

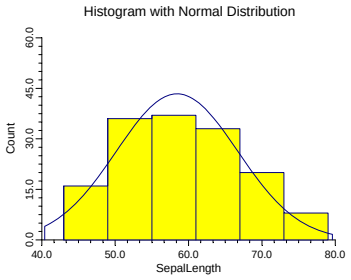
Designates the color, line width, and line pattern of the outline.

Fill Color

This option specifies the interior color of the density trace.

Trace and Line Overlays – Normal Line

This panel controls the appearance of a normal density line. This line uses the estimated mean and standard deviation to overlay a normal density function.



Display Type

Indicate whether to fill the normal density or overlay only its outline. This option is useful when you want to overlay the normal density outline on a histogram or density trace. You can also use this option to completely omit the normal density line.

Number of Calculation Points

Specifies the number of points to be calculated along the horizontal axis. This adjusts the resolution of the normal density curve.

Outline Color, Width, and Pattern

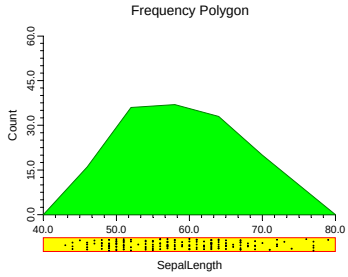
Designates the color, line width, and line pattern of the outline.

Fill Color

This option specifies the interior color of the normal density.

Trace and Line Overlays – Frequency Line

These options control the frequency polygon that may replace—or be added to—the histogram. The frequency polygon is a line that connects the top midpoints of the histogram's bars.



Display Type

Indicate whether to fill the frequency polygon or overlay only its outline. This option is useful when you want to overlay the frequency polygon outline on a histogram. You can also use this option to completely omit the frequency polygon. Note that the number of intervals (bins) is set in the Histogram Tab section.

Outline Color, Width, and Pattern

Designates the color, line width, and line pattern of the outline of the frequency polygon.

Fill Color

This option specifies the interior color of the frequency polygon.

Horizontal Lines from the Vertical Axis

Horizontal Line at Value Below

This option lets you display up to two horizontal lines at particular values. The actual value is specified to the right of the line.

Vertical Lines from the Horizontal Axis

Vertical Line at Value Below

This option lets you display up to two vertical lines at particular values. The actual value is specified to the right of the line.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Histogram

This section presents an example of how to generate a histogram. The data used are from the FISHER database. We will create a histogram of *SepalLength*.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Histograms window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Histograms window.

- On the menus, select **Graphics**, then **Histograms**. The Histograms procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Histograms window, select the **Variables tab**.
- Double-click in the **Data Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Data Variable(s) box.

4 Specify the title.

- On the Histograms window, select the **Titles and Misc. tab**.
- In the **Top Title Line 1** box, enter **Histogram With Everything On It.**

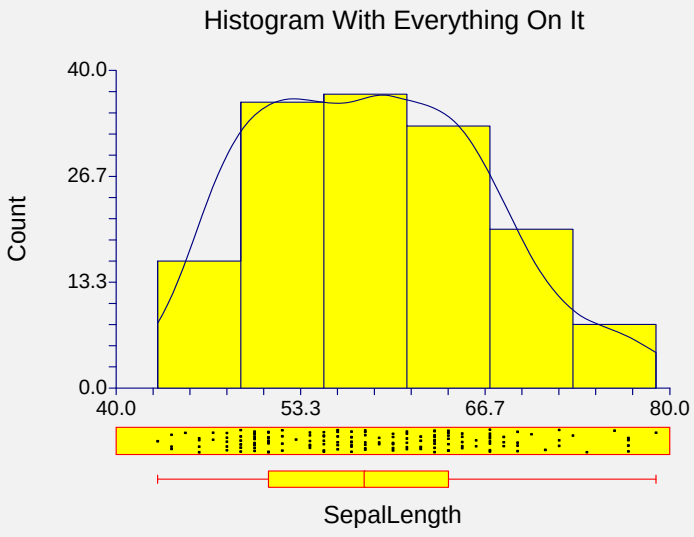
5 Specify the Box Plot and the Dot Plot.

- On the Histograms window, select the **Box and Dot Plots tab**.
- In the **Shape box under Box Plot**, select **Rectangle**.
- In the **Position box under Box Plot**, select **Bottom**.
- In the **Position box under Dot Plot**, select **Bottom**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Histogram Output



This histogram has a density trace overlay with a dot plot and box plot below it. This chart allows you to study almost all univariate features of this variable!

Creating a Histogram Style File

Many of the statistical procedures include histograms as part of their reports. Since the histogram has almost 200 options, adding it to another procedure's report greatly increases the number of options that you have to specify for that procedure. To overcome this, we let you create and save histogram style files. These files contain the current settings of all histogram options. When you use the style file in another procedure you only have to set a few of the options. Most of the options come from this style file. A default histogram style file was installed with the **NCSS** system. Other style files may be added.

We will now take you through the steps necessary to create a histogram style file.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.
- Note: You do not necessarily have to use the FISHER database. You can use whatever database is easiest for you. Just open a database with a column of numeric data.

2 Open the Histograms window.

- On the menus, select **Graphics**, then **Histograms**. The Histograms procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Histograms window, select the **Variables tab**.
- Double-click in the **Data Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Data Variable(s) box.
- Note: don't worry about the specification of this variable. The actual variable names are not stored in the style file. They are used here so that you can see what your style will look like.

4 Set your options.

- Set the various options of the histogram's appearance to the way you want them.
- Run the procedure to generate the histogram. This gives you a final check on whether it appears just how you want it. If it does not appear quite right, go back to the panel and modify the settings until it does.

5 Save the template (optional).

- Although this step is optional, it will usually save a lot of time and effort later if you store the current template. Remember, the template file is not the style file.
- To store the template, select the **Template tab** on the Histogram window.
- Enter an appropriate name in the **File Name** box.

- Enter an appropriate phrase at the bottom of the window in the **Template Id** (the long box across the bottom of the Histograms window). This phrase will be displayed in the Template Id's box to help you identify the template files.
- Select **Save Template** from the File menu. This will save the template.

6 Create and Save the Style File

- Select **Save Style File** from the File menu. The Save Style File Window will appear.
- Enter an appropriate name in the **Selected File** box. You can either reuse one of the style files that already exist or create a new name. You don't have to worry about drives, directory names, or file extensions. These are all added by the program. Just enter an appropriate file name.
- Press the **Ok** button. This will create and save the style file.

7 Using a Style File

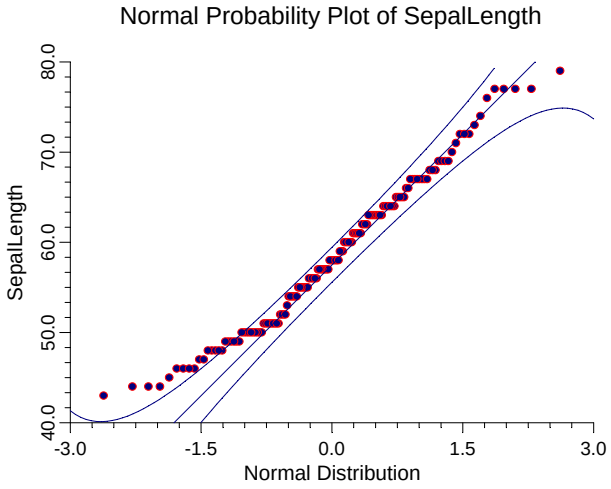
- Using the style file is easy. For example, suppose you want to use this style file to plot residuals in the Multiple Regression procedure. You do the following:
- Select the **Plot Options tab** in the Multiple Regression procedure.
- Click the **button to the right of the Histogram - Plot Style File box** (the initial file name is Default). This will bring up the Histogram Style File Selection window.
- Click on the appropriate file so that it is listed in the **Selected File** box. Click the **Ok** button.
- The new style file name will appear in the Plot Style File box of the Multiple Regression window. That's it. Your new style has been activated.

Chapter 144

Probability Plots

Introduction

This procedure constructs probability plots for the Normal, Weibull, Chi-squared, Gamma, Uniform, Exponential, and Half-Normal distributions. It lets you try various transformations to see if one more closely fits the distribution of interest. Approximate confidence limits are drawn to help determine if a set of data follows a given distribution. If a grouping variable is specified, a separate line is drawn and displayed for each unique value of the grouping variable.



We will provide a brief introduction to probability plotting techniques. A complete discussion of this topic may be found in Chambers (1983). We will try to summarize the information contained there.

Many statistical analyses assume that the data are sampled from a larger population with a specified distribution. Quite often, the distribution of this larger population is assumed to be normal (in reliability and survival work the underlying distribution is assumed to be exponential or Weibull). This is often called the *normality assumption*. (Note that the normal distribution is sometimes called the Gaussian distribution to avoid confusion with its common definition. Although “normal” implies that this is the usual distribution, it is not!) This normality assumption is made for several reasons:

1. It allows the data to be represented compactly. A thousand values that happen to come from the normal distribution may be summarized by only two numbers: the mean and variance.
2. It allows the use of several statistical procedures, such as analysis of variance, t-tests, or multiple regression.

3. It allows generalizations to be made from the sample to the population. These generalizations usually take the form of confidence intervals and hypothesis tests.
4. Understanding the distribution of a sample may provide insight into the physical process that created the data.

Obviously, Mother Nature does not automatically generate data that follows a certain probability distribution. When you assume that your data follows the normal distribution, you are really assuming that the distribution of your data is reasonably approximated by the normal distribution. The question that arises is how close to normal is close enough? This question may be studied using both numerical and graphical procedures.

Numerical hypothesis tests have been developed that allow you to determine whether your data follows a certain distribution. Tests for normality are provided in **NCSS** in the Descriptive Statistics procedure. These tests provide you with a yes or no answer.

Graphical procedures are useful because they give you a visual impression of whether the normality assumption is valid. They let you determine if the assumption is invalidated by one or two outliers (which could be removed), or if the data follow a completely different distribution. They also suggest which data transformation (square root, log, inverse, etc.) might more closely follow the normal distribution.

We feel that the best approach is to apply both numerical and graphical procedures. Since the data is available in your computer, it only takes a few keystrokes to make both checks.

Probability Plot Interpretation

This section will present some of the basics in the analysis and interpretation of probability plots. Our discussion will be brief, so we encourage you to seek further information if you find yourself interpreting these plots regularly. Also, experimentation is a very good teacher. You should make up several “training” databases that follow patterns you understand. Generate probability plots for these so you get a feel for how different data patterns show up on the plots.

If the points in the probability plot all fall along a straight line, you can assume that the data follow that probability distribution. At least, the actual distribution is well approximated by the distribution you have plotted. We will briefly discuss the types of patterns that usually coincide with departures from the straightness of this line.

Outliers

Outliers are values that do not follow the pattern of body of the data. They show up as extreme points at either end of a probability plot. Since large outliers will severely distort most statistical analyses, you should investigate them closely. If they are errors or one-time occurrences, they should be removed from your analysis. Once outliers have been removed, the probability plot should be redrawn without them.

Long Tails

Occasionally, a few points on both ends will stray from the line. These points appear to follow a pattern, just not the pattern of the rest of the data. Usually, the points at the top of the line will shoot up, while the points at the bottom of the line will fall below the line. This is caused by a data distribution with longer tails than would be expected under the theoretical distribution (e.g., normal) being considered. Data with longer tails may cause problems with some statistical procedures.

Asymmetry

If the probability has a convex or concave curve to it (rather than a straight line), the data are skewed to one side of the mean or the other. This can usually be corrected by using an appropriate power transformation.

Plateaus and Gaps

Clustering in the data shows up on the probability plot as gaps and plateaus (horizontal runs of points). This may be caused by the granularity of the data. For example, if the variable may only take on five values, the plot will exhibit these patterns. When these patterns occur, you should be sure you know the reason for them. Is it because of the discrete nature of the data, or are the clusters caused by a second variable that was not considered?

Warning / Caution

Studying probability plots is a very useful tool in data analysis. A few words of caution are in order:

1. These plots emphasize problems that may occur in the tails of the distribution, not in the middle (since there are so many points clumped together there).
2. The natural variation in the data will cause some departure from straightness.
3. Since the plot only considers one variable at a time, any relationships it might have with other variables are ignored.
4. Confidence limits displayed on the plot are only approximate. Also, they depend heavily on a reasonable sample size. For samples of under twenty points, these limits should be taken with a (large) grain of salt. Also, you can change the limits a great deal by changing the confidence level (the alpha value). Be sure that the value you are using is reasonable.

Technical Details

Let us assume that we have a set of numbers x_1, x_2, ..., x_n and we wish to visually study whether the normality assumption is reasonable. The basic method is:

1. Sort the x_i's from smallest to largest. Represent the sorted set of numbers as $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$. Hence, $x_{(1)}$ is the minimum and $x_{(n)}$ is the maximum of these data.
2. Define n *empirical quantiles*, p_1, p_2, ..., p_n, where $p_i = i/n$. These are similar to percentiles. For example, if n = 5 the p_i's would be .2, .4, .6, .8, 1.0. The p_2 value of .4 is interpreted as meaning that this is the 40th percentile.
3. Find a set of numbers, z_1, z_2, ..., z_n, that would be expected from data that exactly follows the normal distribution. For example, z_2 is the number that we would expect if we obtained 5 values from a normal distribution, sorted them, and selected the second from the lowest. These are called the *quantiles*.
4. Construct a scatter plot with the pairs $x_{(1)}$ and z_1, $x_{(2)}$ and z_2, and so on. If the x_i's came from a normal distribution, we would anticipate that the plotted points will fall along a straight line. The degree of non-normality is suggested by the amount of curvature in the plot.

There are several refinements to the procedure outlined above. The most common is the definition of the p_i's in step 2. The formula used by **NCSS** is $p_i = (i-a)/(n-2a+1)$, where "a" is a number

between 0 and 1. Many statisticians recommend a = 1/3. This is the default used by **NCSS**. (The value of a is set in the *Percentile Constant* option.)

Another modification is in the scaling used for the z_i's. If the z_i's from step 3 are used, the strict definition is the quantile plot. If the z's are converted to a probability scale, the plot is known as a probability plot. Nowadays, these definitions have weakened, and we use the term "probability plot" to represent any of these plots.

Probability plots may be constructed for any distribution, although the normal is the most common. The above four steps are used for any of the seven distribution functions that are available in **NCSS**.

Tables from Chambers, Cleveland, Kleiner, and Tukey (1983) are shown below that give technical information about these distributions. One of the most useful features of these tables is the column marked *Ordinate* in the second table. This column defines the transformation of the data that must be used in order to achieve a standard probability plot for that distribution. For example, if you wanted to generate a gamma probability plot, you should raise the data to the one-third power. Note that no special transformation is needed for the normal probability plot.

An estimate of the standard error of z_i is given by:

$$s(z_i) = \frac{\hat{\delta}}{g(q_i)}\sqrt{\frac{p_i(1-p_i)}{n}}$$

where $\hat{\delta}$ is the slope of the points, q_i is the abscissa (given in the second table below), and g(z) is given in the third table. Hence, 100(1-a)% confidence limits may be generated using the z_i as the mean and $s(z_i)$ as the standard error.

These confidence limits serve as reference bounds when you are studying a probability plot. When points fall outside these limits, you would consider them as evidence that the normality assumption (or whatever distribution you are considering) is not valid.

Distribution Functions

Name	Distribution Function	Data Range
Normal	$\Phi\left(\frac{x-\mu}{\sigma}\right)$	$-\infty \le x \le \infty$
Half-Normal	$2\Phi\left(\frac{x}{\sigma}\right) - 1$	$0 \le x$
Weibull	$1 - \exp[-(x/\lambda)^{\theta}]$	$0 \le x$
Exponential	$1 - \exp(-x/\lambda)$	$0 \le x$
Uniform	$(x - \mu)/\lambda$	$\mu \le x \le \mu + \lambda$
Gamma	$G_{\alpha}(x/\lambda)$	$0 \le x$
Chi-square	$C_{\nu}(x/2)$	$0 \le x$

Notes:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

$$G_\alpha(x) = \int_{0}^{x} \frac{z^{\alpha-1} e^{-z}}{\Gamma(\alpha)}\, dz$$

$$C_\nu(x) = G_{\nu/2}(x / 2)$$

Plotting Parameters for Probability Plotting

Name	Ordinate	Abscissa	Intercept	Slope
Normal	$x_{(i)}$	$\Phi^{-1}(p_i)$	μ	σ
Half-Normal	$x_{(i)}$	$\Phi^{-1}\left(\frac{p_i+1}{2}\right)$	0	σ
Weibull	$\log_e(x_{(i)})$	$\log_e[-\log_e(1-p_i)]$	$\log_e\lambda$	θ^{-1}
Exponential	$x_{(i)}$	$(-\log_e(1-p_i))$	0	λ
Uniform	$x_{(i)}$	p_i	μ	λ
Gamma	$x_{(i)}^{1/3}$	$[G_\alpha^{-1}(p_i)]^{1/3}$	0	$\lambda^{1/3}$
Chi-square	$x_{(i)}^{1/3}$	$[2G_{\nu/2}(p_i)]^{1/3}$	0	$\lambda^{1/3}$

Form of g(z) for Estimating Standard Deviations

Name	g(z)
Normal	$1/\sqrt{2\pi}\exp(-1/2z^2)$
Half-Normal	$2/\sqrt{2\pi}\exp(-1/2z^2)$
Weibull	$\exp(z)\exp(-\exp(z))$
Exponential	e^{-z}
Uniform	1
Gamma	$3z^{3\alpha-1}e^{-z^3}/\Gamma(\alpha)$
Chi-square	$3(2)^{-\nu/2}z^{3\nu/2-1}e^{-z^3/2}/\Gamma(\alpha)$

Data Structure

A probability plot is constructed from a single variable. A second variable may be used to divide the first variable into groups (e.g., age group or gender). No other constraints are made on the input data. However, the distributions available in **NCSS** assume that the data are continuous. Note that rows with missing values in one of the selected variables are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the probability plot.

Variables

Variable(s)

This option designates which variables are plotted. If more than one variable is designated, a separate probability plot will be generated for each (unless you have checked the Overlay option).

Overlay Plots on One Graph

This option is used when multiple variables are selected to specify whether to overlay the probability plots of each variable onto a single plot.

Grouping Variable

This variable may be used to separate the observations into groups. When a group variable is selected, the probability plots for the various groups are combined on one plot. The symbols used for each group are set in the Symbols panel.

Data Label Variable

A data label is text that is displayed beside each point. A variable containing the data labels. The values may be text or numeric. You can use dates (like Jan-23-95) as labels. Here is how. First, enter your dates using the standard date format (like 06/20/93). In the Variable Info screen, change the format of the date variable to something like *mmm-dd-yyyy* or *mm-dd-yy*. The labels will be displayed as labels. Without changing the variable format, the dates will be displayed as long integer values.

The size, style, and color of the text may be modified by pressing the second button to the right of the text box. This button brings up the text settings window.

Variables – Distribution Properties

Distribution

Specifies the probability distribution you want to use. Possible choices are normal, Weibull, uniform, exponential, gamma, Chi-square, and half normal. The horizontal axis is scaled according to this probability distribution.

Shape Parameter

This is the shape parameter used for the gamma and Chi-square distributions. This option specifies the value of α (alpha) in the gamma distribution and the degrees of freedom in the Chi-square distribution.

Variables – Options

Percentile Offset

This option supplies the value of a in the p_i formula, $p_i = (i-a)/(n-2a+1)$. The value of 1/3 is usually used because it results in the median z_i. Note that $0 < a < 1$.

X Axis Scaling

This option selects the type of scaling that you want to use on the horizontal axis. *Quantile* scaling results in equi-interval scaling (the standard on all plots) across the horizontal axis. *Probability* scaling uses a variable-interval scaling in which the middle is compressed and the edges are widened to match the probability distribution.

Symbols

Click on the arrow next to the symbol to change the attributes of the symbol.

Symbol Fill Color

The color of the symbol's interior (fill region). Many of the plotting symbols, such as a circle or square, have both an interior region and a border. This specifies the color of the interior region. The border is specified as the Symbol Outline.

Symbol Outline Color

The color of the symbol's border. This color is used when the symbols are connected by a connecting line.

Symbol Type

This option designates the shape of the plot symbol. The most popular symbols may be designated by pressing the appropriate button. About 80 symbol types are available, including letters and numbers.

Symbol Radius

The size (radius) of the symbol.

Symbol Fill Pattern

The pattern (solid, lines, etc.) of the symbol's interior (fill region).

Symbol Outline Width

The width of the symbol's border.

Symbols – Symbol Size Options

Symbol Size Variable

This variable's values are used to modify the size of the plotting symbol. Observations with larger values will be displayed as larger symbols on the plot.

This variable must be numeric.

Minimum Symbol Size

This is the size of the smallest plotting symbol—the one that represents the minimum data value. This number represents a percentage adjustment to the normal size of the plot symbol. Hence, the default value of 50 means that the diameter of the plotting symbol is one half of normal. Note that normal is represented by 100.

Maximum Symbol Size

This is the size of the largest plotting symbol—the one that represents the maximum data value. This number represents a percentage adjustment to the normal size of the plot symbol. Hence, the default value of 200 means that the diameter of the plotting symbol is twice that of normal. Note that normal is represented by 100.

Axes Tab

These options are used to specify the characteristics of the vertical and horizontal axes.

Vertical and Horizontal Axes

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Log Scale

This option lets you select logarithmic scaling for this axis.

- **No**

  Use normal scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as numbers (e.g., 1, 10, 100, 1000).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (-2, 1, 0, 1, 2, 3).

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**
  Specifies the color of the tick labels.
- **Font Size**
  Specifies the size of the tick labels.
- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.
- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.
- **Max Characters**
  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).
- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical and Horizontal Axis – Positions

Axis

This option controls the position of the axis: if and where it is displayed.

Label

This option controls the position of the label: if and where it is displayed.

Tick Labels

This option controls the position of the tick labels: if and where they are displayed.

Tickmarks

This option controls the position of the tickmarks: if and where they are displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the scatter plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. The characters {A} and {B} are replaced by the numeric values of the intercept and slope of the regression line, respectively. To display the fitted regression equation, you could use {Y} = {A} + ({B}){X}.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Legend

When data for more than one group are displayed, a legend is desirable. These options specify the legend.

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. The characters *{G}* will be replaced by the name of the group variable. Click the button on the right to specify the font size, color, and style of the legend text.

Variable Data Transformation

These options specify automatic transformations of the data for either variable.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Box and Dot Plots Tab

These options specify the attributes of the box and dot plots.

Box Plot

Box plots may be placed in the vertical margins of the plot. These box plots emphasize the univariate behavior of the variable.

Shape

This option specifies the shape of the box plot. See the *Box Plot* chapter for further details.

Percentile Type

Specifies the formula used to calculate the percentile. See the *Box Plot* chapter for further details.

Reference Lines

Reference lines may be extended across the plot at each of the three quartiles that are used to form the box.

Inner and Outer Fences

These multipliers are used to construct the fences for designating outliers. See the *Box Plot* chapter for further details.

Line Width

This option specifies the width of the border line.

Box Width

This option specifies the width of the box itself.

Position

This option indicates on which side of the plot the box plot should be displayed.

Show Outliers

This option indicates whether the outlying points should be displayed.

Fill Color

This option specifies the interior color of the box plot.

Line Color

This option specifies the color of the box border and the lines.

Dot Plots

A dot plot may be placed in the vertical margins of the plot. The dot plot lets you study the distribution of the corresponding variable by plotting the actual data values in a line plot.

Position

Specifies whether to display the dot plot and in which margin to place it in.

Box Width

Specifies the width of the dot plot.

Dot Color and Size

Specifies the color and size of the data points displayed in the dot plot.

Line Width and Color

Specifies the width and color of the dot plot’s border.

Fill Color

The color of the data point interior (fill region). Many of the plotting symbols, such as a circle or square, have an interior region and a border. This specifies the color of the interior region.

Lines Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line.

Regression

Regression Line

This option governs the display of a regression line (least-squares trend line or line of best fit) through the data points.

Residual Lines

The residual is the vertical deviation of each point from the regression line. These residuals may be displayed as vertical lines that connect each plot point to the regression line.

Regression Estimation

This option specifies the way in which the trend line is calculated.

- **L.S.**

  The standard least squares regression line is calculated. This formulation is popular but can suffer from severe distortion if one or more outliers exist in your data.

- **Median**

  A resistant least-squares algorithm is used to fit the line, and the intercept is adjusted so that the line passes through the medians of the horizontal and vertical variables. This algorithm is resistant because it removes most of the distortion caused by a few outliers (atypical points).

- **Quartiles**

  A resistant least-squares algorithm is used to fit the line, and the intercept is adjusted so that the line passes through the first and third quartiles of the horizontal and vertical variables. This algorithm is resistant because it removes most of the distortion caused by a few outliers (atypical points).

Confidence Limits

Confidence Limits

Check this box to display approximate confidence limits about the regression line.

Note: This option is available ONLY when the REGRESSION LINE option is checked.

Confidence Limit Alpha

This option gives the confidence level, α, of the $100(1-\alpha)\%$ confidence limits. For example, if α is set to 0.05, you have a 95% confidence limit constructed at each point.

Calculation Points

Number of Calculation Points

The number of positions along the horizontal axis at which the confidence limits, LOESS, and splines are calculated.

Horizontal Lines from the Vertical Axis

Horizontal Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Vertical Lines from the Horizontal Axis

Vertical Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Probability Plot

This section presents an example of how to generate a probability plot. The data used are from the FISHER database. We will create a probability plot of the *SepalLength* variable.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Probability Plots window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Probability Plots window.

- On the menus, select **Graphics**, then **Probability Plots**. The Probability Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

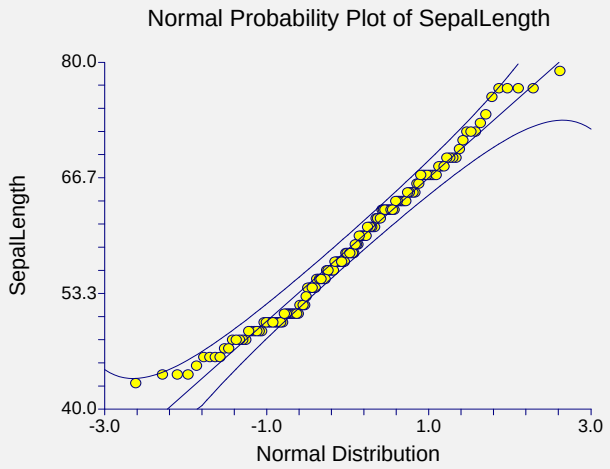
3 Specify the variables.

- On the Probability Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Variable(s) box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Probability Plot Output



If these data were normally distributed, the points would fall along a straight line (note that this line need not be at a 45-degree angle). A reference line is drawn through the points. The slope of this line is found using the 25th and 75th percentile points.

The approximate 95% confidence limits are also displayed. Using these limits as a guide, notice that points below 50.0 are “not normal.” Since only the “tail” of the curve is involved, it is obvious that these data are not normally distributed. (This should not surprise us since this data is the combination of three types of iris.)

Example 2 – Normal Probability Plot of Groups

This section presents an example of how to generate a probability plot of three groups of data. The data used are from the FISHER database. We will create a probability plot of the *SepalLength* variable for each of the three varieties of iris. To run this example, take the following steps:

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Probability Plots window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Probability Plots window.

- On the menus, select **Graphics**, then **Probability Plots**. The Probability Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Probability Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. “SepalLength” will appear in the Variable(s) box.
- Double-click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Group Variable box.

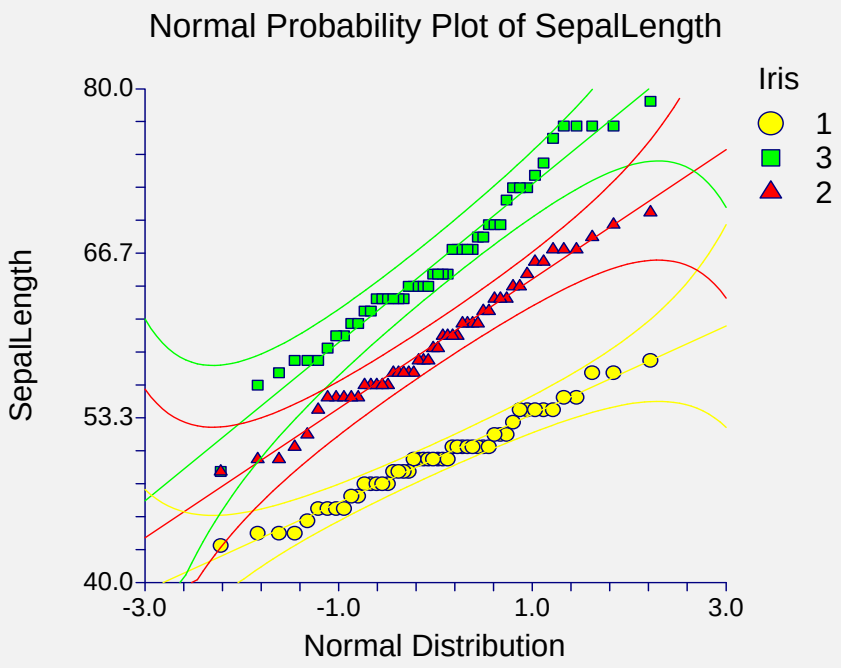
4 Activate the legend.

- On the Probability Plots window, select the **Titles and Misc. tab**.
- Check the **Show Legend** option.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Normal Probability Plot Output



This plot is of the *SepalLength* variable. However, we have separated the data according to iris variety. Note how well the data are modeled by the normal distribution, although groups 1 and 3 seem to stray from the reference line at the upper end.

Example 3 – Weibull Probability Plot

Weibull probability plotting is popular in reliability and survival analysis. This is an example of a typical Weibull plot of two groups of data. The data are contained in the WEIBULL2 database.

You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Probability Plots window.

1 Open the WEIBULL2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **WEIBULL2.s0**.
- Click **Open**.

2 Open the Probability Plots window.

- On the menus, select **Graphics**, then **Probability Plots**. The Probability Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Probability Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **FailTime** from the list of variables and then click **Ok**. "FailTime" will appear in the Variable(s) box.
- Double click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Group** from the list of variables and then click **Ok**. "Group" will appear in the Grouping Variable box.
- Select **Weibull** in the Distribution list box.
- Select **Probability** in the **X Axis Scaling** list box.

4 Setup the Vertical Axis.

- On the Probability Plots window, select the **Axes tab**.
- Under Vertical Axis, select **Yes: Numbers** in the Log Scale list box.
- Under Vertical Axis, check the **Major** and **Minor Grid Lines** options.
- Under Vertical Axis, double click in the box with the line below **Minor Ticks**. This will bring up the Settings window.
- Under Grid & Tick Settings, select **Dot** in the **Grid Pattern** list box.
- Click **OK**.

5 Setup the Horizontal Axis.

- On the Probability Plots window, select the **Axes tab**.
- Under Horizontal Axis, enter **11** beneath Major Ticks.
- Under Horizontal Axis, check the **Major Grid Lines** option.
- Under Horizontal Axis, double click in the box with the line below **Major Ticks**. This will bring up the Settings window.
- Under Grid & Tick Settings, select **Dot** in the **Grid Pattern** list box.
- Click **OK**.
- Under Horizontal Axis, click the **Tick Label Settings…** button. This will bring up the Settings of Tick Label Settings window.
- Under Text Settings set the **Decimals** list box to **2**.
- In the **Text Settings - Text Rotation** box, select **Vertical**.
- Click **OK**.

6 Omit the Confidence Limits.

- On the Probability Plots window, select the **Lines tab**.
- Click on **Confidence Limits** so that the option is no longer checked.

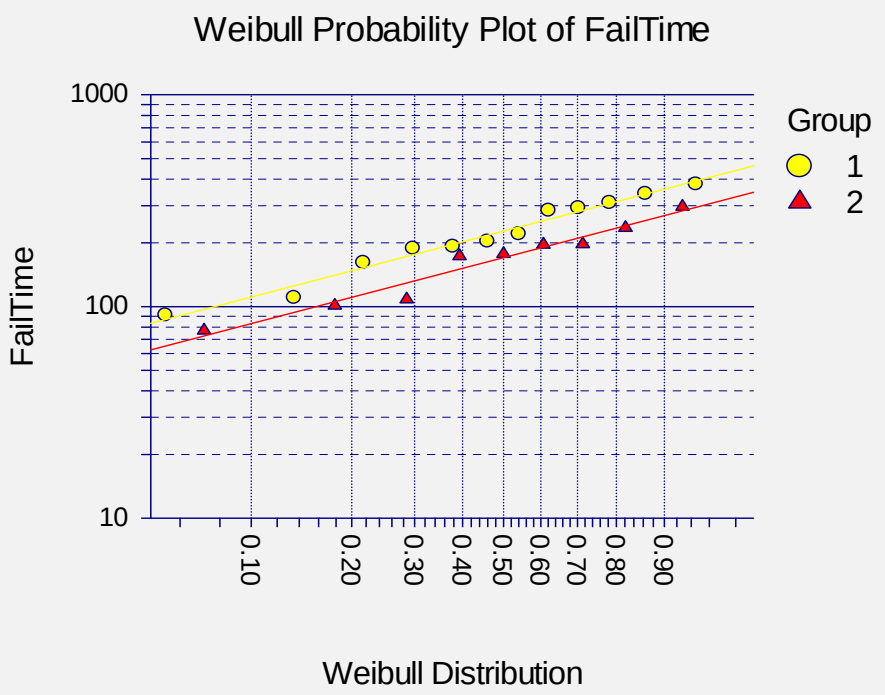
7 Activate the legend.

- On the Probability Plots window, select the **Titles and Misc. tab**.
- Check the **Show Legend** option.

8 Run the procedure.

- From the Run menu, select **Run Procedure.** Alternatively, just click the Run button (the left-most button on the button bar at the top).

Weibull Probability Plot Output



Note that the vertical axis has a logarithmic scale.

Creating a Probability Plot Style File

Many of the statistical procedures include probability plots as part of their reports. Since the probability plot has almost 200 options, adding it to another procedure's report greatly increases the number of options that you have to specify for that procedure. To overcome this, we let you create and save probability plot style files. These files contain the current settings of all probability plot options. When you use the style file in another procedure you only have to set a few of the options. The remaining options come from this style file. A default probability plot style file was installed with the **NCSS** system. Other style files may be added.

We will now take you through the steps necessary to create a probability plot style file.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.
- Note: You do not necessarily have to use the FISHER database. You can use whatever database is easiest for you. Just open a database with a column of numeric data.

2 Open the Probability Plots window.

- On the menus, select **Graphics**, then **Probability Plots**. The Probability Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Probability Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Variable(s) box.

4 Set your options.

- Set the various options of the probability plot's appearance to the way you want them.
- Run the procedure to generate the probability plot. This gives you a final check on whether it appears just how you want it. If it does not appear quite right, go back to the panel and modify the settings until it does.

5 Save the template (optional).

- Although this step is optional, it will usually save a lot of time and effort later if you store the current template. Remember, the template file is not the style file.
- To store the template, select the **Template tab** on the Probability Plots window.
- Enter an appropriate name in the **File Name** box.
- Enter an appropriate phrase at the bottom of the window in the **Template Id** (the long box across the bottom of the Probability Plots window). This phrase will be displayed in the Template Id's box to help you identify the template files.
- Select **Save Template** from the File menu. This will save the template.

6 Create and Save the Style File

- Select **Save Style File** from the File menu. The Save Style File Window will appear.
- Enter an appropriate name in the **Selected File** box. You can either reuse one of the style files that already exist or create a new name. You don't have to worry about drives, directory names, or file extensions. These are all added by the program. Just enter an appropriate file name.
- Press the **Ok** button. This will create and save the style file.

7 Using a Style File

- Using the style file is easy. For example, suppose you want to use this Probability Plot Style file in the Two-Sample T-Test procedure. You do the following:
- Select the **Probability Plot tab** in the Two-Sample T-Test procedure.
- Click the button to the right of the **Plot Style File** box (the initial file name is Default). This will bring up the Probability Plot Style File Selection window.
- Click on the appropriate file so that it is listed in the **Selected File** box. Click the **Ok** button.
- The new style file name will appear in the Plot Style File box of the Two-Sample T-Test window. Your new style has been activated.

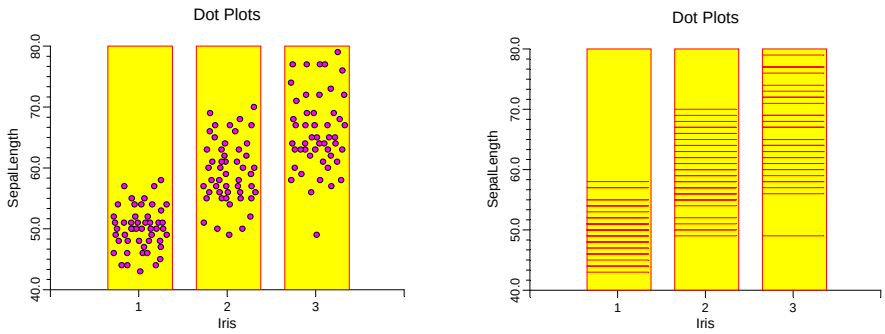
Chapter 150

Dot Plots

Introduction

The dot plot is a plot of a single batch of data. One version shows values as points. Another version shows values as lines. Dot plots are usually augmented to other plots, such as the scatter plot. The dot plot is especially useful for detecting strange patterns in your data. These patterns will show up as horizontal lines (of points).

Following are two versions of the dot plot for the SepalLength variable of Fisher's iris data.



Dot Plot Definition

The vertical axis represents the variable's values. The horizontal axis represents different groups. The width of the box in the dot plot is arbitrary. Some statisticians suggest adding a random horizontal component to avoid having several points overlie each other. This process, called *jittering*, is available as an option in **NCSS**.

Data Structure

A dot plot is constructed from one variable. A second variable may be used to divide the first variable into groups (e.g., age group or gender). In this case, a separate dot plot is displayed for each group. No other constraints are made on the input data.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the dot plot.

Variables

Variable(s)

This option lets you designate which variables are plotted. If more than one variable is designated and no Grouping Variable is selected, a set of dot plots will be displayed on a single chart, one box for each variable. If more than one variable is designated and a Grouping Variable is selected, a separate dot plot will be drawn for each variable.

Grouping Variable

Designates an optional variable used to separate the observations into groups. An individual box will be displayed for each unique value of this variable.

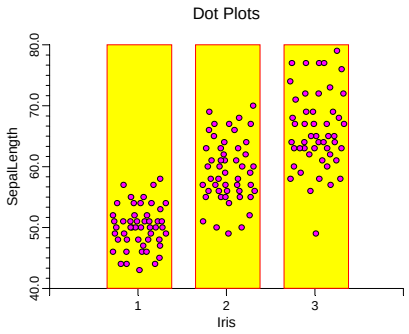
Dot Plot Dots

Type

This option specifies the type of “dots” you want plotted.

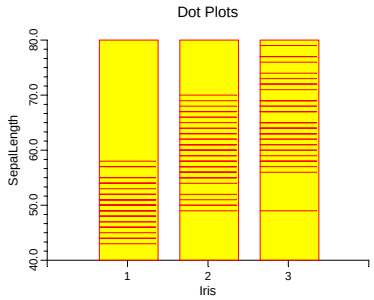
- **Textured Dots**

  A random horizontal component is added to avoid overprinting when two values are identical. This is a form of jittering.



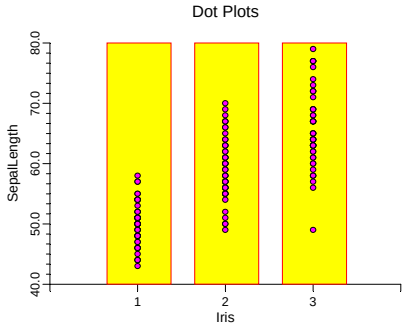
- **Lines**

  We call this the “DNA” chart for obvious reasons. This version of the dot plot is especially useful for detecting clumpiness in your data.



- **Standard Dots**

  This is the standard version in which values are represented along the vertical axis. No random horizontal component is added. As you can see, multiple values cannot be detected.



Symbol

This option define the shape, size, and color of the plotting symbol. Click the button on the right to bring up the window that lets you edit these options.

Dot Plot Boxes – Outline

Color

This option specifies the color of the box border and the lines.

Width

This option specifies the width of the border line.

Dot Plot Boxes – Fill

Color

This option specifies the interior color of the boxes.

Dot Plot Boxes - Width

Select Box Width Parameter

This option designates how you want to specify the width of boxes. You can specify an *Actual Amount* or a *Percent Space*.

Amount

Specify the exact width of the box.

Percent Empty Space

This option specifies what percent of the horizontal axis should be kept as “white space.” The smaller this value, the larger the box width.

Axes Tab

Vertical and Horizontal Axis

Label Text

This box supplies the axis label. The characters {X}, {Y}, and {G} are replaced by the horizontal, vertical, and grouping variable names, respectively. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line and its type (numeric or text).

Log Scale

This option lets you select logarithmic scaling for this axis.

- **No**

  Use normal scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as numbers (e.g., 1, 10, 100, 1000).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (-2, 1, 0, 1, 2, 3).

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical and Horizontal Axis – Positions

Axis

This option controls the position of the axis: if and where it is displayed.

Label

This option controls the position of the label: if and where it is displayed.

Tick Labels

This option controls the position of the tick labels: if and where they are displayed.

Tickmarks

This option controls the position of the tickmarks: if and where they are displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls these titles. In the text, the characters *{Y}* and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls these titles. In the text, the characters *{Y}* and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Variable Data Transformations

These options specify automatic transformations of the data.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Lines Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line.

Horizontal Lines from the Vertical Axis

Horizontal Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Vertical Lines from the Horizontal Axis

Vertical Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Dot Plot

This section presents an example of how to generate a dot plot. The data used are from the FISHER database. We will create dot plots of the *SepalLength* variable, breaking on the type of iris.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Dot Plots window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Dot Plots window.

- On the menus, select **Graphics**, then **Dot Plots**. The Dot Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Dot Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Variable(s) box.

- Double-click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Grouping Variable box.

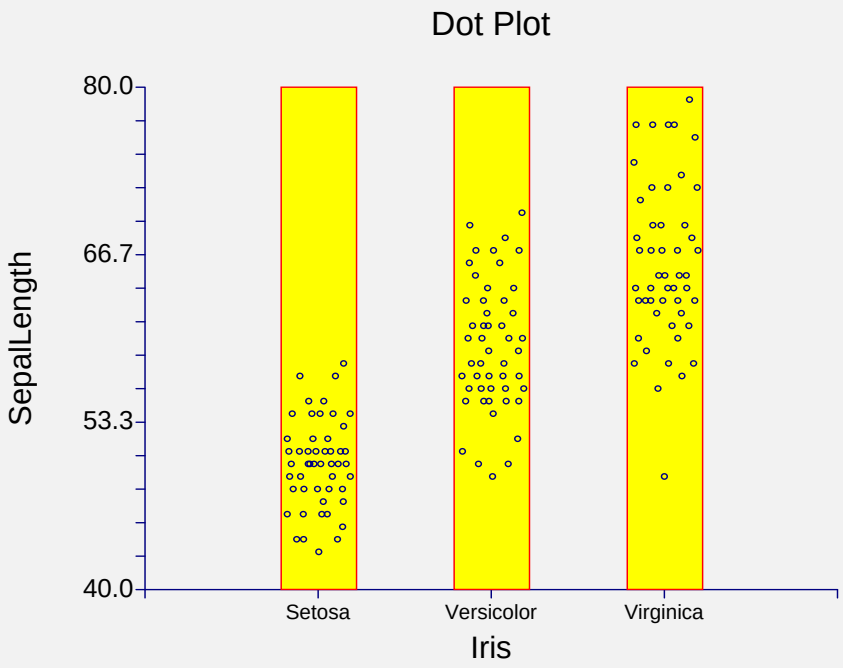
4 Specify the labels.

- On the Dot Plots window, select the **Titles and Misc tab**.
- In the **Value Labels** list box, select **Value Labels**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Dot Plot Output

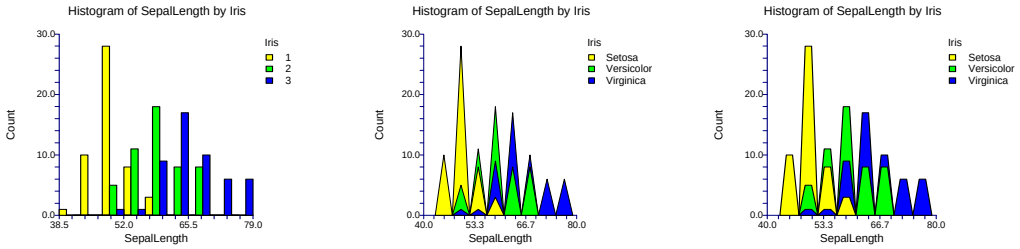


Chapter 151

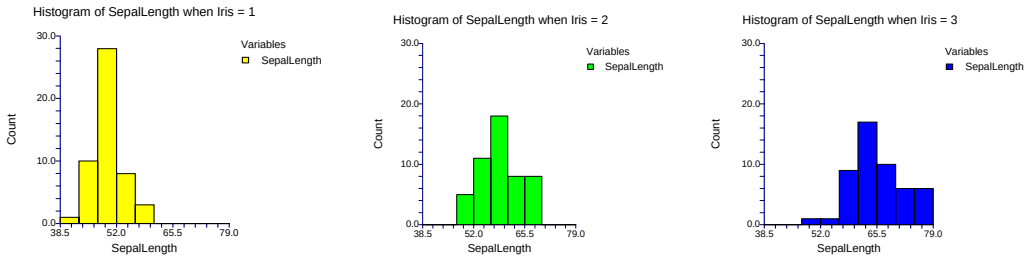
Histograms – Comparative

Introduction

A histogram displays the frequency distribution of a set of data values. This procedure displays a comparative histogram created by interspersing or overlaying the individual histograms of two or more groups or variables. This allows the direct comparison of the distributions of several groups. Here are some examples.

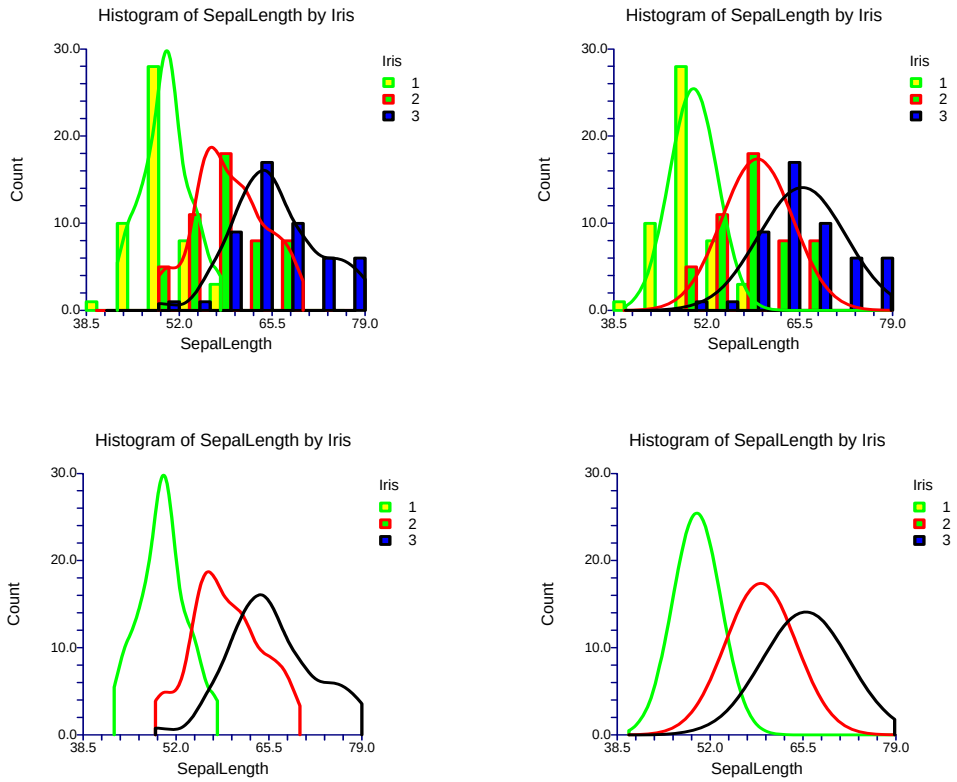


Here are the three histograms that have been combined.



In this example, the histogram on top is formed by interspersing the bars of the three individual histograms shown on the bottom. The bar width of the new plot is equal to the original bar width divided by the number of groups displayed.

To increase the usefulness of the chart, either the density trace or the normal density line may be added to the chart. These may be displayed together with the histograms, or by them selves. (Note that the color scheme used below was selected to give reasonable results when the charts are printed in black and white. These are not the colors that we would choose when printing the chart in color.)



Technical Details

The technical details of creating histograms are given in the Histogram chapter, and they will not be repeated here. We refer you to that chapter for further details. The additional options necessary to construct combined histograms are discussed in the Procedure Options section below.

Data Structure

The data for a combined histogram may be arranged in two ways. In the first arrangement, the data for each group is placed in separate variables (columns) of the database. In the second arrangement, the data is placed in a single variable and a second variable is created that contains a group identification value in each row. This arrangement is identical to that used for entering analysis of variance data.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the histogram.

Variables

Data Variable(s)

Designate the variables whose numeric values are to be arranged in a combined histogram. When two or more variables are selected, their histograms will be interspersed on the chart.

If a Group Variable is specified, the variable, or variables, displayed together on a chart are controlled by the 'If Many Groups and Many Variables' setting.

Grouping Variable

This variable may be used to separate the observations into groups. If the *If Many Groups and Many Variables* is set to *Separate plot for each GROUP*, a separate chart will be constructed for each unique value of this variable. Otherwise, a separate chart is generated for each variable and the individual histograms of the unique values of this variable are combined.

Frequency Variable

This option specifies an optional variable that contains the number of observations (cases or counts) represented by each row. If multiple *Data Variables* are specified, these frequencies are applied to all of them.

If this variable is left blank, each row of the database is assumed to represent one observation.

Since the values in this variable represent counts, they should be positive integers greater than or equal to one.

Variables – Discrete Dataset Indicator

Discrete Dataset when Uniques <=

If a discrete dataset (one with only a few unique values) is detected, the number of bins and bin dimensions are set to give only one bin per unique value (although extra bins may be necessary). Datasets with this number or fewer unique values are treated as discrete. Datasets with more unique values are considered to be continuous.

The value 10 is a reasonable value. Enter 0 to force all datasets to be continuous.

Specify Number of Bars Using

Number of Bars

Specifies the number of bars (bins, slices, or intervals). Select *0 - Automatic* to direct the program to select an appropriate number based on the number of values.

If a discrete variable (one with only a few unique values) is detected, the number of bars is equal to the number of unique values. Whether a variable is discrete is determined by the *Max Uniques* setting.

Bar Width

This option may be used to set the width of the bars. It is used with the *Axis Minimum* or *Axis Maximum* (see Horizontal tab). This value is not used when the *Axis Minimum*, the *Axis Maximum*, and the *Number of Bars* are all set.

Specify Number of Bars Using – Data Range

Minimum

This option lets you specify the smallest data value that you want included in the dataset used to construct the chart. Data below this value will be ignored.

Note: This value does not control the horizontal axis of the chart. If you want to control the horizontal axis, use the *Axis Minimum* setting under the *Horizontal* tab.

Maximum

This option lets you specify the largest data value that you want included in the dataset used to construct the chart. Data above this value will be ignored.

Note: This value does not control the horizontal axis of the chart. If you want to control the horizontal axis, use the *Axis Maximum* setting under the *Horizontal* tab.

Histograms

Display Type

Indicate whether to display the histogram as an outline, a solid, or not at all. The options are Omit, Outline Only, or Outline and Fill.

- **Omit**
  Do not display the histogram.
- **Outline Only**
  Display the histogram by overlaying an outline of it. Other objects will still be visible.
- **Outline and Fill**
  Display a solid histogram by filling the outline with the specified fill color. Other objects already displayed may be covered by the histogram.

If Many Groups and Many Variables

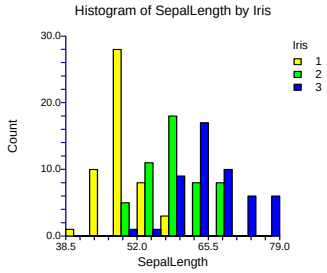
If you designate a *Group Variable* and several data variables, you can either have one chart for each group or one chart for each data variable. This option lets you make this choice.

Histograms – Bar Shape

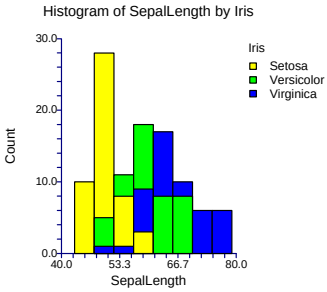
Bar Shape

This option lets you specify the type of comparative histogram that you want to display. Your choices are

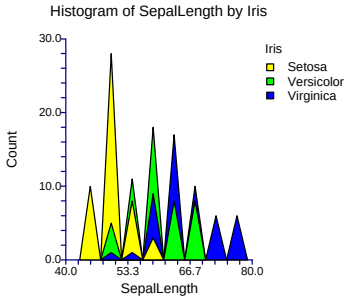
- **Side-by-Side**



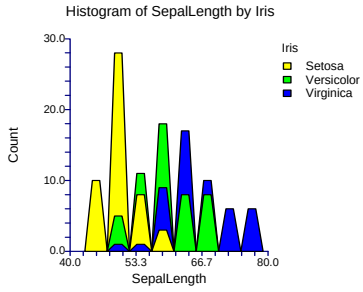
- **Direct Overlay**



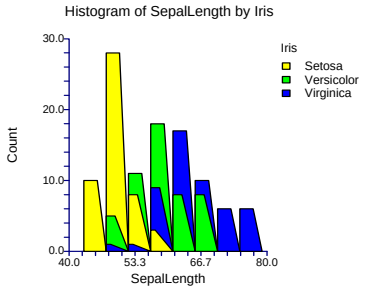
- **Triangle Overlay**



- **Trapezoid Overlay**



- **Fan Overlay**



Histograms – Bar Outline

Width and Pattern

Designates the line width and line pattern of the outline of the bars, density trace, and normal density line.

Axes Tab

These options specifies the characteristics of the vertical and horizontal axes.

Vertical Axis

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Maximum

Specifies the largest value shown on this axis. Note that the minimum is always set to zero.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Scale Type

Specifies whether the vertical scale is displayed as a Count or a Percentage.

Vertical Axis – Tick Marks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks and grid lines displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

Tick labels are displayed for the minor tickmarks. This option specifies the number of minor tickmarks and grid lines displayed between each set of major tickmarks.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the minor grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**
  Specifies the color of the tick labels.
- **Font Size**
  Specifies the size of the tick labels.
- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.
- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.
- **Max Characters**
  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).
- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the right side, the left side, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the right side, the left side, on both sides, or not displayed.

Tick Labels

This option controls the position of the tick labels: whether they are placed on the right side, the left side, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Horizontal Axis

These options specifies the characteristics of the horizontal axis.

Label Text

This box supplies the horizontal axis label. The characters {Y} and {G} are replaced by the appropriate variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Location of Bar Labels

This option specifies where the tick labels are placed along the horizontal axis.

- **Standard**

  Selecting this option indicates a typical placement of the tick labels. Note that the numbers do not necessarily match the bars.

- **Mid Points**

  One tick label is placed at the middle of each bar.

- **End Points**

  One tick label is placed at the end of each bar.

Horizontal Axis – Tick Marks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Horizontal Axis – Position

Axis

This option controls the position of the axis: whether it is placed on the bottom, the top, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tick Labels

This option controls the position of the tick labels: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Legend

Show Legend

This box determines whether the legend will be displayed.

Legend Title

This box supplies the title that will appear above the legend. If used, the characters {G} are replaced by the corresponding variable names. The font size, color, and style of the title may be modified by pressing the button on the right of the text.

Variable Data Transformation

These options specify automatic transformations of the data.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Bar and Line Colors Tab

This panel controls the colors of histogram bars and outlines.

Histogram Bar and Line Colors

Fill

This is the interior color that fills the histogram bar of the corresponding group. Double click this box to change this value.

Outline

This is the color of the bar outline, the density trace, and the normal density line of the corresponding group. Note that the width and pattern of these objects are set using the Outline Width and Outline Pattern under the Options tab. Double click this box to change this value.

Overlay Lines Tab

This panel controls the inclusion and attributes of the overlay lines.

Trace and Line Overlays – Density Trace

Display Type

Indicate whether to display the density trace as an outline, a solid, or not at all. The options are Omit, Outline Only, or Outline and Fill.

- **Omit**
  Do not display the density trace.
- **Outline Only**
  Display the density trace by overlaying an outline of it. Other objects will still be visible.
- **Outline and Fill**
  Display a solid density trace by filling the outline with the specified fill color. Other objects already displayed may be covered by the density trace.

Number of Calculation Points

Specifies the number of density trace points to be calculated along the horizontal axis. This adjusts the resolution of the density trace.

Percent of Data in Calculation

This option controls the roughness of the density trace by setting the percent of the data used in the density calculation. Select 0 for an automatic value determined from your data. A low value (near 10%) will give a rough plot. A high value (near 40%) will yield a smooth plot.

Trace and Line Overlays – Normal Line

Display Type

Indicate whether to display the normal density as an outline, a solid, or not at all. The options are Omit, Outline Only, or Outline and Fill.

- **Omit**
  Do not display the normal density.
- **Outline Only**
  Display the normal density by overlaying an outline of it. Other objects will still be visible.
- **Outline and Fill**
  Display a solid normal density by filling the outline with the specified fill color. Other objects already displayed may be covered by the normal density.

Number of Calculation Points

Specifies the number of normal density points to be calculated along the horizontal axis. This adjusts the resolution of the normal density line.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Comparative Histogram

This section presents an example of how to generate a combined histogram of the *SepalLength* variable of the FISHER database breaking on the type of iris.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Histograms – Comparative window.

1 Open the FISHER dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FISHER.S0**.
- Click **Open**.

2 Open the Histograms - Comparative window.

- On the menus, select **Graphics**, then **Histograms - Comparative**. The Histograms - Comparative procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

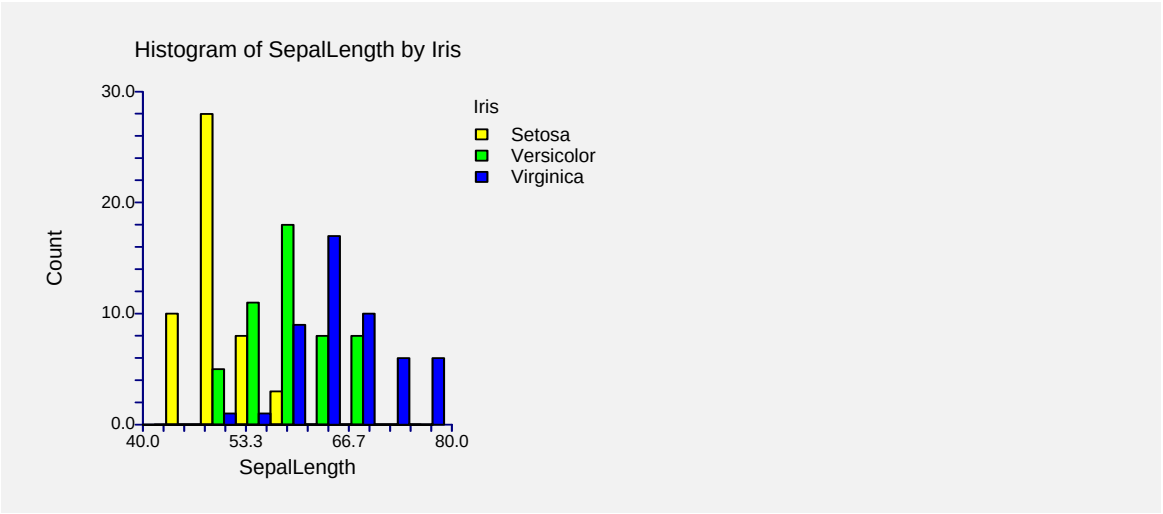
- On the Histograms - Comparative window, select the **Variables tab**.
- Set the **Data Variable(s)** box to **SepalLength**.

- Set the **Grouping Variable** box to **Iris**.
- Set the **If Many Groups and Many Variables** box to **Separate plot for each Variable**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Comparative Histogram Output



This shows a standard combined histogram with colors selected for optimum contrast when printed on black and white paper.

Example 2 – Discrete Values

This section gives an example of how to create a combined histogram on discrete data. The example will use the RESALE database. A combined histogram will be created for the variable *Bedrooms*, grouping by the variable State. Note the *Bedrooms* is a discrete variable with values from 1 to 4.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Histograms – Comparative window.

1 Open the RESALE dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **RESALE.S0**.
- Click **Open**.

2 Open the Histograms - Comparative window.

- On the menus, select **Graphics**, then **Histograms - Comparative**. The Histograms - Comparative procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Histograms - Comparative window, select the **Variables tab**.
- Set the **Data Variable(s)** box to **Bedrooms**.
- Set the **Grouping Variable** box to **State**.
- Set the **If Many Groups and Many Variables** box to **Separate plot for each Variable**.

4 Specify the text.

- On the Histograms - Comparative window, select the **Titles and Misc. tab**.
- Set the **Value Labels** box to **Value Labels**.

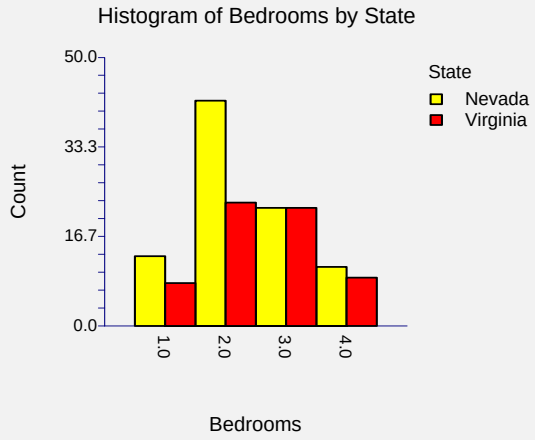
5 Specify the horizontal axis.

- On the Histograms - Comparative window, select the **Axes tab**.
- Set the **Location of Bar Labels** box to **Mid Points**.

6 Run the procedure.

- From the **Run** menu, select **Run Procedure**. Alternatively, just click the **Run** button (the left-most button on the button bar at the top).

Comparative Histogram Output



Notice how quickly we can see that on this database, Nevada tends to have a much larger percentage of two-bedroom homes than does Virginia. Note that the bin boundaries have been created so that the discrete values (1, 2, 3, 4) are centered in the bins.

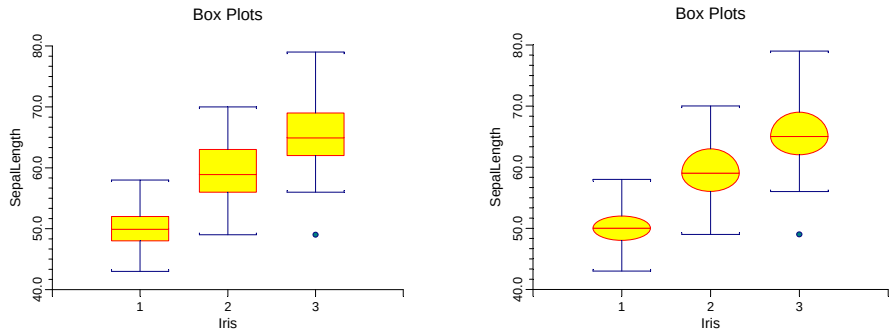
151-16 Histograms – Comparative

Chapter 152

Box Plots

Introduction

When analyzing data, you often need to study the characteristics of a single batch of numbers, observations, or measurements. You might want to know the center and how spread out the data are about this central value. You might want to investigate extreme values (referred to as outliers) or study the distribution of the data values (the pattern of the data values along the measurement axis). Several techniques are available to allow you to study the distribution. These include the stem-leaf plot, histogram, density trace, probability plot, and box plot.



Box Plot Definition

The box plot shows three main features about a variable: its center, its spread, and its outliers.

Box

A box plot is made up of a box (a rectangle) with various lines and points added to it. The width of the box is arbitrary and should be selected to make an eye-pleasing display. The top and bottom of the box are the 25^{th} and 75^{th} percentiles. The length of the box is thus the interquartile range (IQR). That is, the box represents the middle 50% of the data. The IQR is a popular measure of spread. You can represent the box as a rectangle, a diamond, an ellipse, or a special figure designed for making multiple comparisons.

A line is drawn through the middle of the box at the median (the 50^{th} percentile). The median is a popular measure of the variable's location (center or average value).

Adjacent Values

The *upper adjacent value* is the largest observation that is less than or equal to the 75th percentile plus 1.5 times IQR. The *lower adjacent value* is the smallest observation that is greater than or equal to the 25th percentile minus 1.5 times IQR.

The adjacent values are displayed as T-shaped lines that extend from each end of the box.

Outside Values

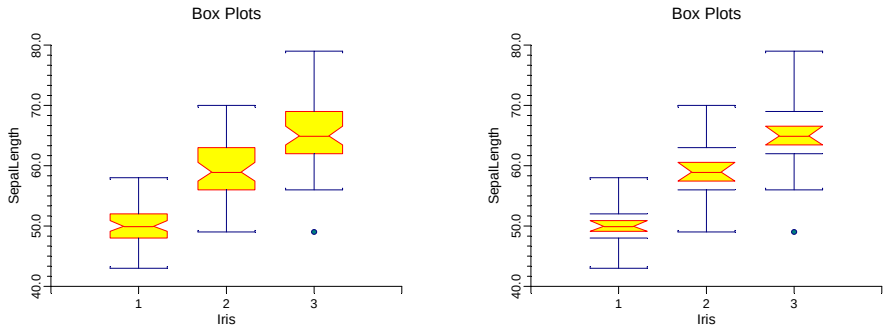
Values outside the upper and lower adjacent values are called *outside values*. Values that are under three IQRs from the 25th and 75th percentiles are called *mild* outliers. Those outside three IQRs are called *severe* outliers. Mild outliers are not unusual, but severe outliers are.

Multiple Comparisons

Box plots are often used for comparing the distributions of several batches of data, since they summarize the center and spread of the data very nicely. When making strict comparisons among the locations (medians) of various batches, a modified box plot called the *notched box plot* is useful. The notches are constructed using the formula:

$$Median \pm 1.57 \times (IQR)/\sqrt{n}$$

Notched box plots are used to make multiple comparisons among the batches. If the notches of two boxes do not overlap, we may assume that the medians are significantly different (the centers are statistically significant). The 1.57 is selected for the 95% level of significance. The box plot on the left is the classical notched box plot.



Recently, statisticians have noticed that the notched box plot does not allow you to focus on the multiple comparisons. A modern version of the notched box plot has been proposed that lets you make this comparison (see the above plot on the right). This version modifies the symbol used for the box. In fact, it leaves the box out. Two horizontal lines mark the position of the box. The part that is plotted is the notched part only. This makes it much easier to make comparisons. If two of the notches overlap, the group medians are not significantly different. Otherwise, they are.

Note that when making comparisons among several batches, the notched box plots do not make any adjustment for the multiplicity of tests being conducted. As long as the notched box plots are used informally, no technical adjustments are necessary.

Data Structure

A box plot is constructed from one variable. A second variable may be used to divide the first variable into groups (e.g., age group or gender). In this case, a separate box plot is displayed for each group. No other constraints are made on the input data.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the box plot.

Variables

Variable(s)

This option lets you designate which variables are plotted. If more than one variable is designated and no Grouping Variable is selected, a set of box plots will be displayed on a single chart, one box for each variable. If more than one variable is designated and a Grouping Variable is selected, a separate box plot will be drawn for each variable.

Grouping Variable

Designates an optional variable used to separate the observations into groups. An individual box will be displayed for each unique value of this variable.

Data Label Variable

A data label is text that is displayed beside each outside point. This option designates the variable containing the data labels. The values may be text or numeric.

Box

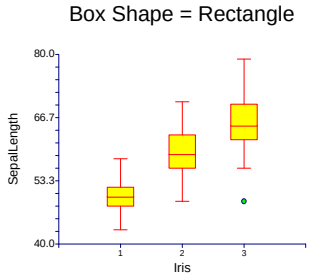
Shape

This option specifies the type (shape) of the box plot. Possible shapes are rectangle (standard), diamond, ellipse, and notched (original and modified). The special shapes are used for making comparisons among several groups (see Multiple Comparisons earlier in this chapter).

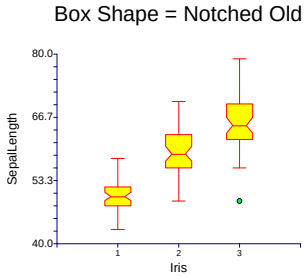
- **None**

  The box plot is not displayed.

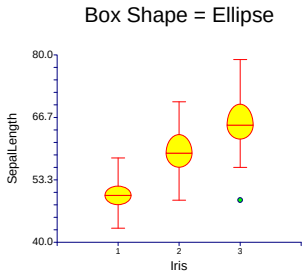
- **Rectangle**



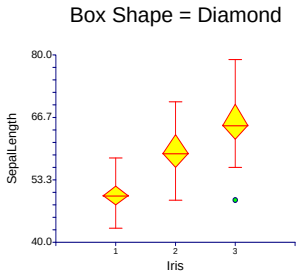
- **Notched – Old**



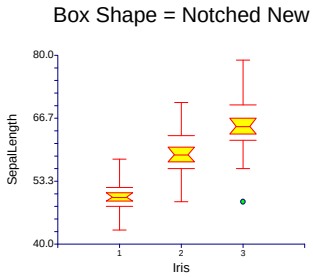
- **Ellipse**



- **Diamond**



- **Notched – New**



Median Symbol

Click in the box or on the button to its right to display a window that allows you to change the characteristics of the median symbol.

Note: If a horizontal line is selected (which is the default), it will be ignored and the outline color and width will be used instead.

Percentile Calculation Method

Specify the formula used to calculate the percentile. See the Descriptive Statistics chapter for further details.

Box – Outline

Color

This option specifies the color of the box border and the lines.

Width

This option specifies the width of the border line.

Box – Fill

Color

This option specifies the interior color of the boxes.

Box – Width

Box Width Parameter

This option designates how you want to specify the width of boxes. You can specify an actual Amount or a Percent Space.

Amount

Specify the exact width of the box. It is only used if it is specified in Use for Box Width above.

Percent Empty Space

This option specifies what percent of the horizontal axis should be kept as “white space.” The smaller this value is, the larger the box width will be. It is only used if it is specified in Use for Box Width above.

Lines (Whiskers)

The *adjacent value* is the line that extends up and down from the box. The following options concern the display of these lines.

Type

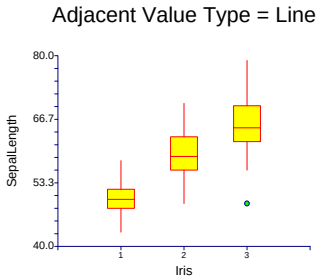
Specifies the type of pattern used to display the adjacent values. Possibilities include a simple line, a T-shaped line, a T-shaped line with backward extenders, and none (omit the adjacent values).

comparisons among several groups (see Multiple Comparisons earlier in this chapter).

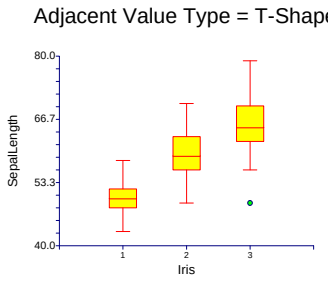
- **None**

  The adjacent values (lines) are not displayed.

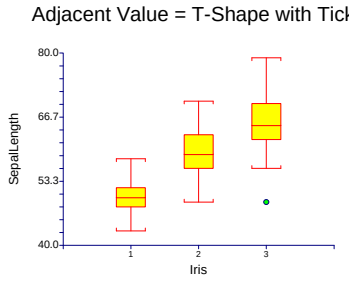
- **Line**



- **T-Shape**



- **T-Shape with Ticks**



Line

This option specifies the color, width, and pattern of the adjacent value line.

Lines (Whiskers) – Type = T-Shape with Ticks

Tick Length

The length of the T-shaped with Ticks option's backward extenders.

Outliers

Show Outliers

This option indicates whether the outlying points should be displayed. It also lets you designate whether you want to show only severe outliers or all outside points.

Outliers - Mild / Severe

Fence Multiplier

These are the constant used to construct the fences. The adjacent values are the *inner fences*. The boundary for declaring an outside value as mild or severe is the outer fence. These values are typically set at 1.5 and 3.0, respectively. These options let you manipulate these constants.

Symbol

This option designates the appearance of the symbol used to portray mild and severe outliers. Click the button on the right of the box to change features of the symbol such as the color and type.

Axes Tab

These options specify the characteristics of the vertical and horizontal axes.

Vertical Axis

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Log Scale

This option lets you select logarithmic scaling for this axis.

- **No**

  Use normal scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as numbers (e.g., 1, 10, 100, 1000).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (-2, 1, 0, 1, 2, 3).

Vertical Axis – Tickmarks and Grid Lines

Major Ticks (number)

Reference numbers are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**
  Specifies the color of the tick labels.

- **Font Size**
  Specifies the size of the tick labels.

- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.

- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**
  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the right side, the left side, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the right side, the left side, on both sides, or not displayed.

Tick Labels

This option controls the position of the reference numbers: whether they are placed on the right side, the left side, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Horizontal Axis

These options specifies the characteristics of the horizontal axis.

Label Text

This box supplies the horizontal axis label. The characters {Y} and {G} are replaced by the appropriate variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Horizontal Axis – Tickmarks

Ticks

This option sets the color, length, width, and pattern of the tickmarks along this axis. The tickmarks are positioned one per individual box plot.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Horizontal Axis – Positions

Axis

This option controls the position of the axis: whether it is placed on the bottom, the top, on both sides, or not displayed.

Label

This option controls the position of the label: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tick Labels

This option controls the position of the reference numbers: whether it is placed on the bottom, the top, on both sides, or not displayed.

Tickmarks

This option controls the position of the tickmarks: inside the axis, outside the axis, on both sides, or not displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls these titles. In the text, the characters *{Y}* and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls these titles. In the text, the characters *{Y}* and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

Background Color

The background color of the plot.

Interior Color

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display a variable’s name, its label, or both the name and label.

Value Labels

This option selects whether to display values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Variable Data Transformation

Transform Exponent

Each value of the variable is raised to this exponent automatically before it is processed. Fractional exponents require positive data values.

Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Lines Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line.

Horizontal Lines from the Vertical Axis

Horizontal Line at Value Below

This option lets you display a horizontal line at a particular value. The actual value is specified to the right of the line.

Vertical Lines from the Horizontal Axis

Vertical Line at Value Below

This option lets you display a vertical line at a particular value. The actual value is specified to the right of the line.

Horizontal Reference Lines at Median and Box Ends

Reference Line Type

Reference lines may be extended across the plot at each of the three quartiles that are used to form the box.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Box Plot

This section presents an example of how to generate a box plot. The data used are from the FISHER database. We will create box plots of the *SepalLength* variable, breaking on the type of iris.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Box Plots window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Box Plots window.

- On the menus, select **Graphics**, then **Box Plots**. The Box Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

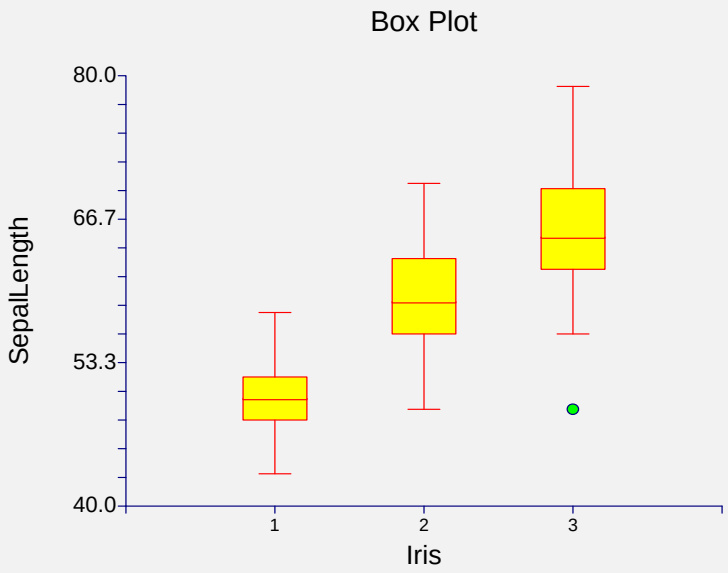
3 Specify the variables.

- On the Box Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Variable(s) box.
- Double-click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. "Iris" will appear in the Grouping Variable box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Box Plot Output



Creating a Box Plot Style File

Many of the statistical procedures include box plots as part of their reports. Since the box plot has almost 200 options, adding it to another procedure's report greatly increases the number of options that you have to specify for that procedure. To overcome this, we let you create and save box plot style files. These files contain the current settings of all box plot options. When you use the style file in another procedure you only have to set a few of the options. Most of the options come from this style file. A default box plot style file was installed with the **NCSS** system. Other style files may be added.

We will now take you through the steps necessary to create a box plot style file.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.
- Note: You do not necessarily have to use the FISHER database. You can use whatever database is easiest for you. Just open a database with a column of numeric data.

2 Open the Box Plots window.

- On the menus, select **Graphics**, then **Box Plots**. The Box Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Box Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. “SepalLength” will appear in the Variable(s) box.
- Double-click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Grouping Variable box.

4 Set your options.

- Set the various options of the box plot’s appearance to the way you want them.
- Run the procedure to generate the box plot. This gives you a final check on whether it appears just how you want it. If it does not appear quite right, go back to the panel and modify the settings until it does.

5 Save the template (optional).

- Although this step is optional, it will usually save a lot of time and effort later if you store the current template. Remember, the template file is not the style file.
- To store the template, select the **Template tab** on the Box Plot window.
- Enter an appropriate name in the File Name box.
- Enter an appropriate phrase at the bottom of the window in the Template Id (the long box across the bottom of the Box Plot’s window). This phrase will be displayed in the Template Id’s box to help you identify the template files.
- Select **Save Template** from the File menu. This will save the template.

6 Create and Save the Style File

- Select **Save Style File** from the File menu. The Save Style File Window will appear.
- Enter an appropriate name in the **Selected File** box. You can either reuse one of the style files that already exist or create a new name. You don’t have to worry about drives, directory names, or file extensions. These are all added by the program. Just enter an appropriate file name.
- Press the **Ok** button. This will create and save the style file.

7 Using a Style File

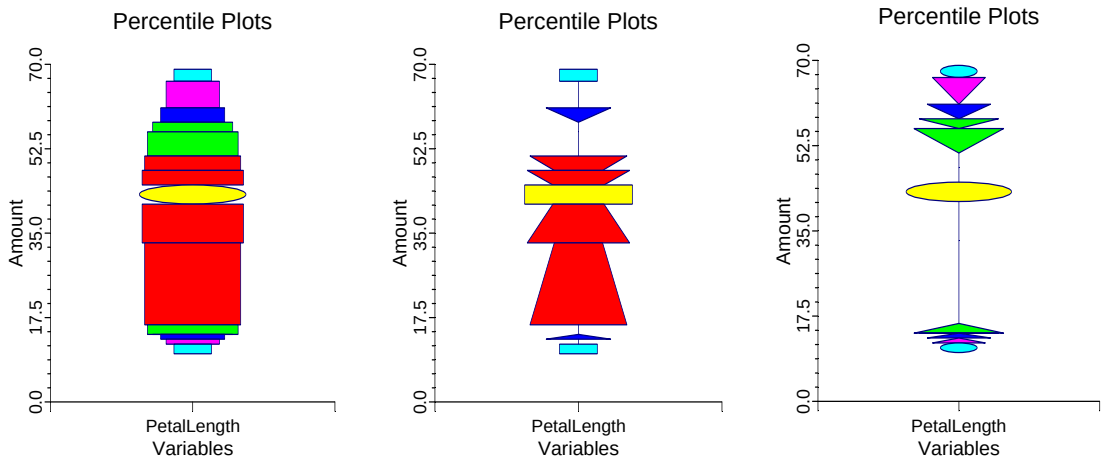
- Using the style file is easy. For example, suppose you want to use this Box Plot Style file in the Two-Sample T-Test procedure. You do the following:
- Select the **Box Plot tab** in the Two-Sample T-Test procedure.
- Click the button to the right of the **Plot Style File** box (the initial file name is Default). This will bring up the Box Plot Style File Selection window.
- Click on the appropriate file so that it is listed in the **Selected File** box. Click the **Ok** button.
- The new style file name will appear in the Plot Style File box of the Two-Sample T-Test window. That’s it. Your new style has been activated.

Chapter 153

Percentile Plots

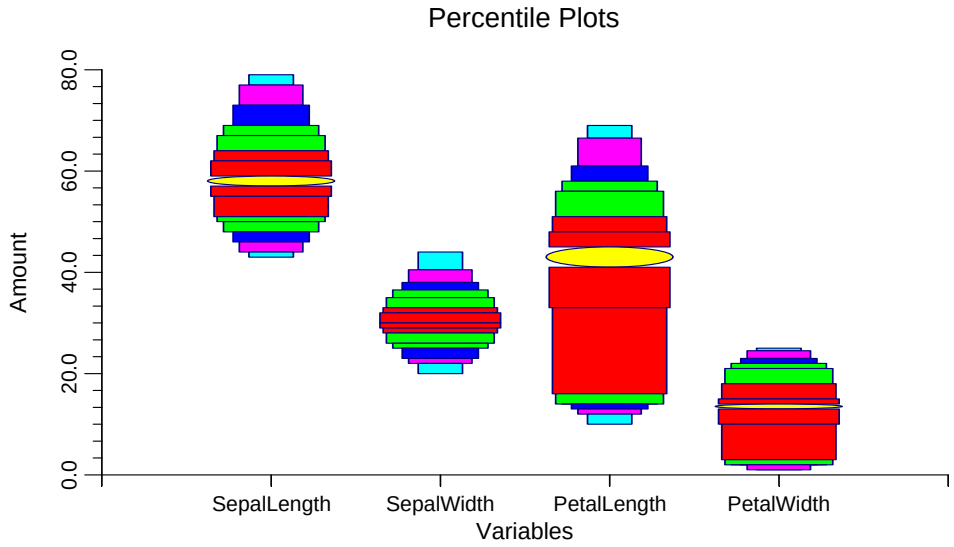
Introduction

This procedure calculates and displays percentiles between 0 and 100. It lets you assign different shapes, colors, and widths to each percentile group. Using various combinations, you can generate percentile plots that will be tailored to your particular need. Following are some examples of percentile plots that were generated on the PetalWidth variable of Fisher's iris data.



Percentile Plot

The *Percentile Plot* is constructed from a set of percentiles. These percentiles are 0-2, 2-5, 5-10, 10-15, 15-25, 25-35, 35-45, 45-55, 55-65, 65-75, 75-85, 85-90, 90-95, 95-99, 99-100. One of six possible shapes (line, rectangle, triangle, trapezoid, ellipse, nothing) is used to display the percentile values. For example, suppose the rectangle shape is selected for the percentile range 55-65. A rectangle is drawn that connects the value of 55th percentile with the value of the 65th percentile. The color and width of this rectangle are also under your control. Following is an example of a percentile plot of four variables.



Notice how easy it is to compare various percentiles across the four variables.

Data Structure

A percentile plot is constructed from one or more variables. A second variable may be specified to divide the first variable into groups (e.g., age group or gender).

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the percentile plot.

Variables

Variable(s)

This option lets you designate which variables are plotted. If more than one variable is designated and no Grouping Variable is selected, a set of percentile plots will be displayed on a single chart, one plot for each variable. If more than one variable is designated and a Grouping Variable is selected, a separate plot will be drawn for each variable.

Grouping Variable

Designates an optional variable used to separate the observations into groups. An individual plot will be displayed for each unique value of this variable.

Percentiles

Percentile

This is the percentile range whose options are designated.

Color

This option specifies the color of the corresponding percentile range. Careful choice of color is very helpful in making comparisons among the percentile charts on a percentile plot.

Shape

This option specifies the shape of the corresponding percentile range. Possible values are a line, rectangle, triangle opening down, triangle opening up, trapezoid opening down, trapezoid opening up, ellipse, and nothing. Using different shapes will produce a variety of percentile plots.

Width

This option specifies the width of each percentile range. Indicate the width as a percentage of the total chart.

Plot Options

Percentile Type

Specifies the method used to calculate percentiles. Refer to *Percentile Type* in the Descriptive Statistics chapter.

Plot Options - Lines

Median

This option designates the appearance of the line used to portray the median. Click the button on the right of the box to change features of the line such as the color and width. The check box indicates whether to display a line at the median.

Outline

This option designates the appearance of the outline around each portion of the plot. Click the button on the right of the box to change features of the line such as the color and width.

Connecting

Occasionally, you will find it useful to connect equal percentiles across the percentile plots on one percentile chart. This option lets you do this.

This option designates the appearance of the connecting lines. Click the button on the right of the box to change features of the line such as the color and width. The check box indicates whether to display the lines.

Plot Options – Box Width

Select Width Parameter

This option lets you designate whether to specify the box width using the actual amount or the percent of space between the bars.

Amount

When the Select Width Parameter is set to Amount, the option gives the width of the bars.

Percent Empty Space

When the Bar Width Method is set to Percent Space, the option gives the percent of the total space that is to be between the bars.

Axes Tab

These options are used to specify the characteristics of the vertical and horizontal axes.

Vertical and Horizontal Axis

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Log Scale

This option lets you select logarithmic scaling for this axis.

- **No**

  Use normal scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as numbers (e.g., 1, 10, 100, 1000).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (-2, 1, 0, 1, 2, 3).

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical and Horizontal Axis – Positions

Axis

This option controls the position of the axis: if and where it is displayed.

Label

This option controls the position of the label: if and where it is displayed.

Tick Labels

This option controls the position of the tick labels: if and where they are displayed.

Tickmarks

This option controls the position of the tickmarks: if and where they are displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the scatter plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. The characters {A} and {B} are replaced by the numeric values of the intercept and slope of the regression line, respectively. To display the fitted regression equation, you could use {Y} = {A} + ({B}){X}.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable’s name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Variable Data Transformation

These options specify automatic transformations of the data for either variable.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Lines Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line.

Horizontal Lines from the Vertical Axis

Horizontal Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Vertical Lines from the Horizontal Axis

Vertical Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Percentile Plot

This section presents an example of how to generate a percentile plot. The data used are from the FISHER database. We will create percentile plots of the *SepalLength* variable, breaking on the type of iris.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Percentile Plots window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Percentile Plots window.

- On the menus, select **Graphics**, then **Percentile Plots**. The Percentile Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Percentile Plots window, select the **Variables tab**.
- Double-click in the **Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Variable(s) box.
- Double-click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. "Iris" will appear in the Grouping Variable box.

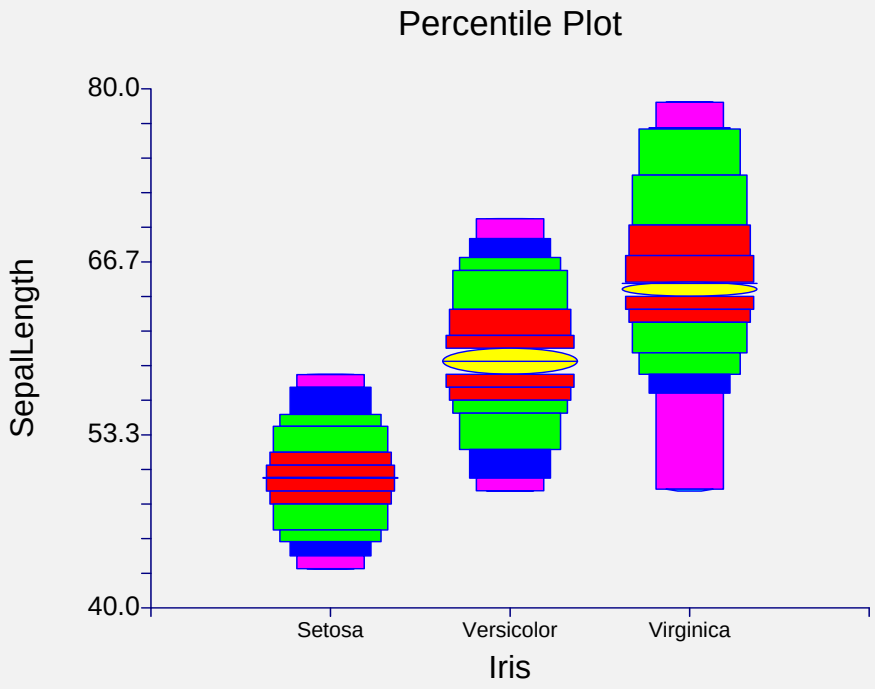
4 Specify the titles.

- On the Percentile Plots window, select the **Titles and Misc. tab**.
- In the **Value Labels** list box, select **Value Labels**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Percentile Plot Output

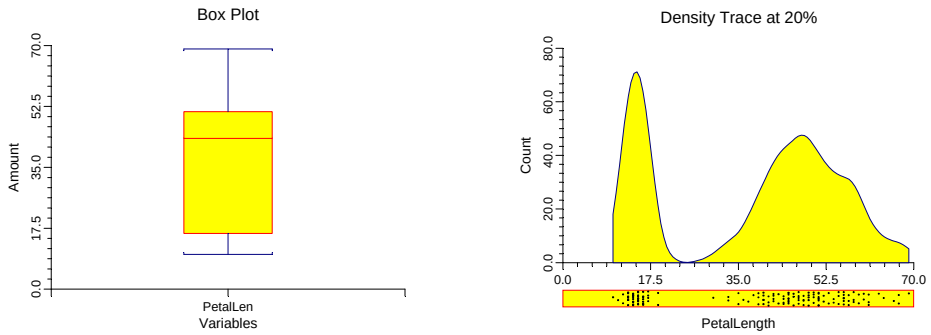


Chapter 154

Violin Plots

Introduction

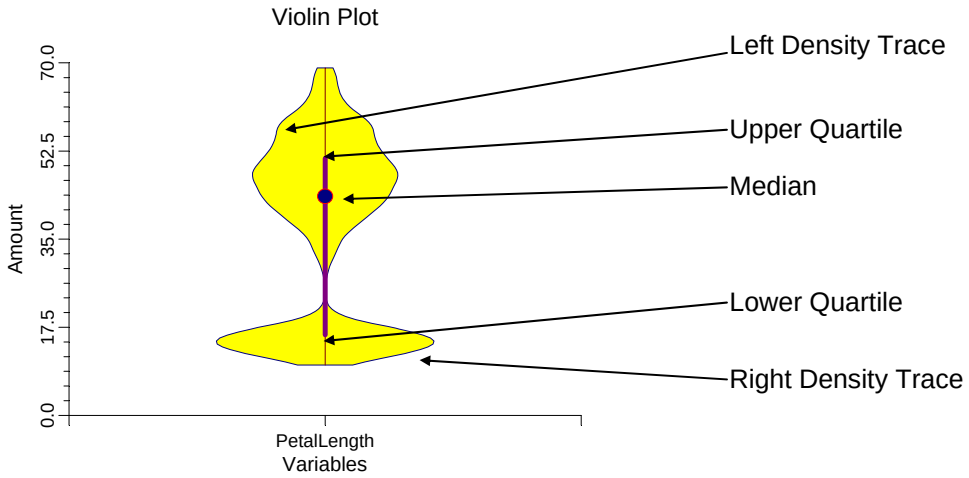
The box plot is useful for displaying the mean and spread of a set of data. Several box plots may be displayed side by side to allow you to compare the average and spread of several groups. The density trace (or histogram) is useful for displaying the distribution of the data. Unfortunately, several density traces shown side by side are difficult to compare. Yet, comparing the distributions of several batches of data is a common task. We (see Hintze and Nelson 1998) have invented a new plot, which we call the *Violin Plot.* This plot is a hybrid of the density trace and the box plot. It allows you to compare several distributions quickly.



Violin Plot

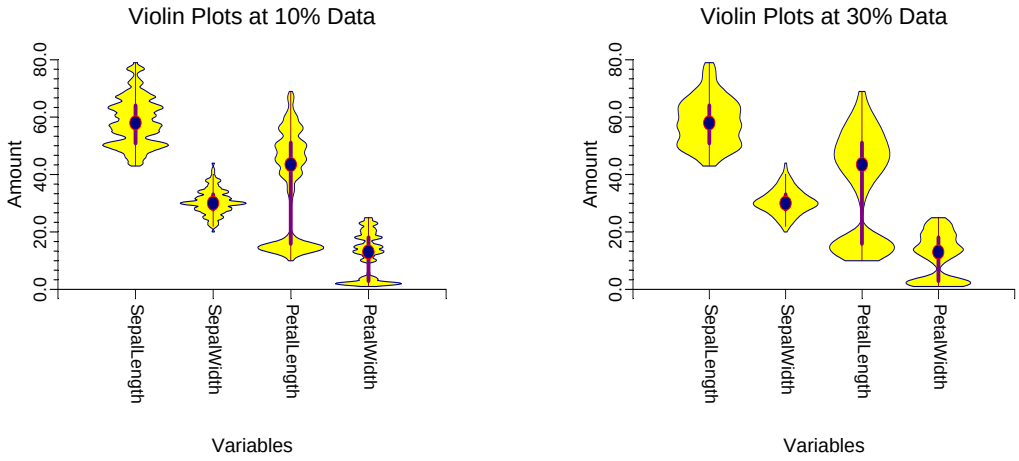
The *Violin Plot* is made by combining a form of box plot with two vertical density traces. One density trace extends to the left while the other extends to the right. There is no difference in these density traces other than the direction in which they are extended. We put two density traces on the plot to add symmetry, which makes it much easier to compare batches. The violin plot highlights the peaks and valleys of a variable's distribution.

We changed the box plot slightly by showing the median as a circle. We did this so that quick comparisons of the medians could be made. We called this a violin plot because one of the first datasets we tried had the appearance of a violin.



If you compare this plot with the box plot and density trace of the same data, you will notice that although the box plot is useful in a lot of situations, it does not represent data that are clustered (multimodal). On the other hand, although the density trace shows the distribution of the data, it is hard to see the mean and spread. The obvious answer to these shortcomings is to combine the two plots.

The following example shows violin plots of the four variables in Fisher's iris data.



Notice how easily you can compare the medians, the box lengths (the spread), and the distributional patterns in the data. In this example, notice that the two petal variables show two peaks (bimodal) while the two sepal variables are unimodal (one peak). We have provided several options that will let you adjust this plot to your needs.

Data Structure

A violin plot is constructed from one or more variables. A second variable may be used to divide the first variable into groups (e.g., age group or gender). In this case, a violin plot is displayed for each group. No other constraints are made on the input data.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the violin plot.

Variables

Variable(s)

This option lets you designate which variables are plotted. If more than one variable is designated and no Grouping Variable is selected, a set of box plots will be displayed on a single chart, one plot for each variable. If more than one variable is designated and a Grouping Variable is selected, a separate plot will be drawn for each variable.

Grouping Variable

Designates an optional variable used to separate the observations into groups. An individual plot will be displayed for each unique value of this variable.

Box Plot

Percentile Calculation Method

Specifies the method used to calculate percentiles. Refer to *Percentile Type* in the Descriptive Statistics chapter.

Median Symbol

This option designates the appearance of the symbol used to portray the median. Click the button on the right of the box to change features of the symbol such as the color and type.

Interquartile Range Line

This option designates the appearance of the line used to portray the interquartile range. The interquartile range refers to the box part of the box plot. It is a line connecting the twenty-fifth and seventy-fifth percentiles. The check box indicates whether to show this option. Click the button on the right of the box to change features of the line such as its color and thickness.

Line to Adjacent Value (Whisker)

The adjacent value is the line that extends up and down from the box part of the box plot. This option designates the appearance of the line. The check box indicates whether to show this option. Click the button on the right of the box to change features of the line such as its color and thickness.

Density Trace

Number of Calculation Points

Specifies the number of points at which the density is calculated. This adjusts the resolution of the density trace.

Percent of Data in Calculation

The percent of the data used in the density calculation. Select 0 for an automatic value determined from your data. A low value (near 10%) will give a rough plot. A high value (near 40%) will yield a smooth plot.

Outline

Designates the line width, line pattern, and color of the density trace outline.

Fill Color

This option specifies the interior color of the density trace.

Density Trace – Violin Width

Select Violin Width Parameter

This option designates how you want to specify the width of the rectangle that surrounds each "violin." You can specify an Actual Amount or a Percent Space.

Amount

Specifies the width of the rectangle containing the "violin."

Percent Empty Space

This option specifies what percent of the horizontal axis should be kept as "white space." This value determines the width of the violins. The smaller this value, the larger the violin width.

Axes Tab

These options are used to specify the characteristics of the vertical and horizontal axes.

Vertical and Horizontal Axes

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line.

Log Scale

This option lets you select logarithmic scaling for this axis.

- **No**

  Use normal scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as numbers (e.g., 1, 10, 100, 1000).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (-2, 1, 0, 1, 2, 3).

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**
  Specifies the size of the tick labels.

- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.

- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**
  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical and Horizontal Axis – Positions

Axis

This option controls the position of the axis: if and where it is displayed.

Label

This option controls the position of the label: if and where it is displayed.

Tick Labels

This option controls the position of the tick labels: if and where they are displayed.

Tickmarks

This option controls the position of the tickmarks: if and where they are displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the scatter plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. The characters {A} and {B} are replaced by the numeric values of the intercept and slope of the regression line, respectively. To display the fitted regression equation, you could use {Y} = {A} + ({B}){X}.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Variable Data Transformation

These options specify automatic transformations of the data for either variable.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Lines Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line.

Horizontal Lines from the Vertical Axis

Horizontal Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Vertical Lines from the Horizontal Axis

Vertical Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Violin Plot

This section presents an example of how to generate a violin plot. The data used are from the FISHER database. We will create violin plots of the *SepalLength* variable, breaking on the type of iris.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Violin Plots window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Violin Plots window.

- On the menus, select **Graphics**, then **Violin Plots**. The Violin Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Violin Plots window, select the **Variables tab**.
- Double-click in the **Variable(s) text** box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Variable(s) box.
- Double-click in the **Grouping Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. "Iris" will appear in the Grouping Variable box.

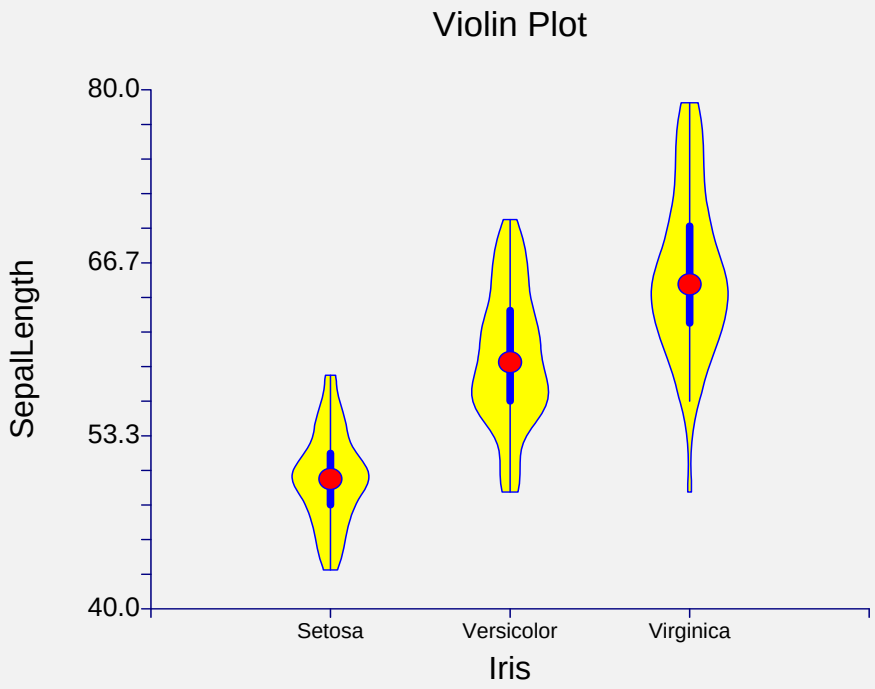
4 Specify the titles.

- On the Violin Plots window, select the **Titles and Misc. tab**.
- In the **Value Labels** list box, select **Value Labels**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Violin Plot Output

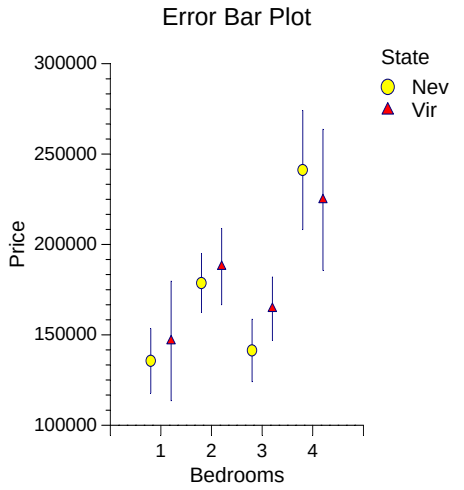


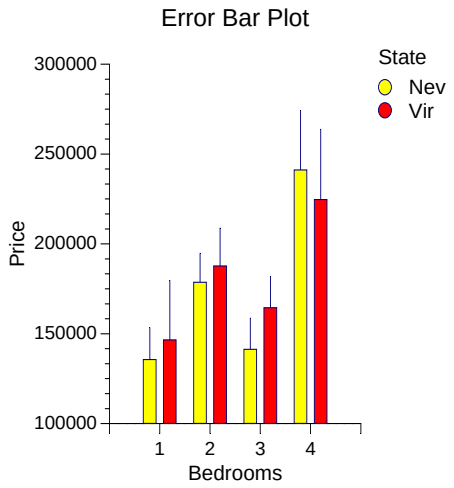
Chapter 155

Error-Bar Charts

Introduction

Error Bar Charts graphically display tables of means and standard errors (or standard deviations). Following are examples of the types of charts produced by this procedure.





Data Structure

Each row of data must contain at least one numeric response variable. Up to two categorical variables may also be designated. Also, up to five break variables may be selected. A separate chart is produced for each unique value of the break variables.

Following is an example with a single response variable and two categorical variables. The data below are a subset of the RESALE database provided with the software. This data gives the state, the selling price, and the number of bedrooms for 150 residential properties sold during four months in two states. Only the first 8 of the 150 observations are displayed.

RESALE dataset (subset)

State	Price	Bedrooms
Nev	260000	2
Nev	66900	3
Vir	127900	2
Nev	181900	3
Nev	262100	2

Missing Values

Missing values are removed on a case-by-case basis. That is, a missing value in one variable does not cause other variables in that row to be ignored.

Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Variables Tab

This panel specifies the variables that will be used in the analysis.

Variables

Response(s) Variables

Select at least one response variable. The means and standard errors generated will be for the values in these variables.

If several response variables are selected, they will be displayed across the horizontal axis, as the legend variable, or as one variable per chart depending on the other options.

Horizontal Grouping Variable

Specify a categorical variable whose categories (values) appear along the horizontal axis. If this option is left blank, the response variables are displayed across the horizontal axis (the One Plot per Response Variable option must not be check in this case).

Note that numeric values are treated as text values and will be equi-spaced across the horizontal axis. Hence, the values 1, 2, 5 will be equi-spaced just the same as the values 1, 2, 3 or A, B, C.

Legend (Subgroup) Variable

Specify a categorical variable whose categories (values) appear as separate series across the horizontal axis. These series are identified on the chart in the legend.

One Plot Per Response Variable

This option designates whether a separate plot should be produced for each response variable (checked) or the response variables should be combined to form the horizontal axis or the legend (not checked).

Frequency Variable

This optional variable specifies the number of observations that each row represents. When omitted, each row represents a single observation. If your data are the result of previous summarization, you may want certain rows to represent several observations. Note that negative values are treated as a zero frequency and are omitted. Fractional values may be used.

Break (Separate Chart) Variables (1 to 5)

You can optionally specify one or more categorical variables whose categories (values) will cause separate plots to be generated. A separate chart is generated for each unique set of values of the variable(s) specified.

Error Bar Length Settings

Variation Type

This option lets you select whether to display the *standard errors* of the means or the *standard deviations* of the data. Note that the standard error is calculated from the standard deviation using the formula:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where s is the standard deviation of an individual group of values.

Multiplier

The length of the standard error line extending up (and down) from the mean is this value times the standard error (or standard deviation). Usually, this value is 1.0.

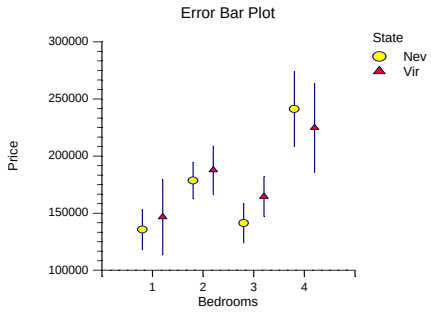
Plot Settings

Plot Type

Two plot formats are available.

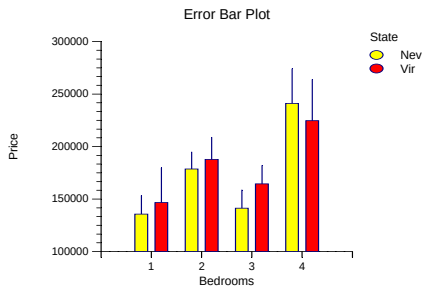
- **Regular Plot**

  In this plot, the means are shown with plot symbols and the standard errors are shown as lines extending up and down from the symbol.



- **Bar Plot**

  In this plot, the means are shown with bars and the standard errors are shown as lines extending up from the bar.



Symbol Size

Specify the size of the plot symbols.

Bar Size

Specifies the width of the bars when Plot Type is set to Bar Plot.

Connect Means

This option indicates whether to connect the means with a line.

Show Break as Title

Specifies whether the current values of any Break variables should be displayed as a second title line in the plot.

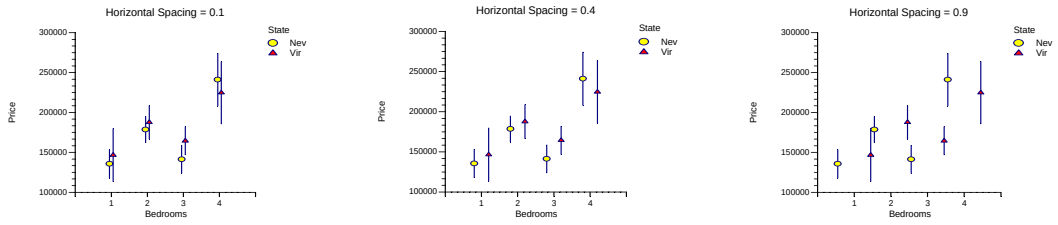
Show Cross Bar

This option indicates whether to show the horizontal “cross bar” at the end of the line(s).

Horizontal Spacing

When multiple series are shown across the horizontal axis, this value sets the spacing between groups of values. The standard distance between to groups on the horizontal axis is 1.0. Entering a value of 0.4 here causes 0.4 of this space to be empty. The possible range of values is between 0.0 and 1.0.

The following examples show this parameter at 0.1, 0.4, and 0.9.



Axes Tab

These options are used to specify the characteristics of the vertical and horizontal axes.

Vertical and Horizontal Axes

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Show Grid Lines

Checking this option causes the major grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**

  Specifies the color of the tick labels.

- **Font Size**

  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the scatter plot.

Titles

Plot Title

This is the text of the title. The characters *{Y}*, *{X}*, and *{G}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Style File

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Legend

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the appropriate default value.

Symbols Tab

Specify the symbols used to display the means.

Symbols for Means (when Plot Type = Regular Plot)

Group 1-15

Specify the symbol used to designate a particular group. Double-click on a symbol or click on the button to the right of a symbol to specify the symbol's type and color.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating an Error-Bar Plot

This section presents an example of how to create error bar plots. The data shown earlier and found in the RESALE database are used to generate the plot.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Error-Bar Charts window.

1 Open the RESALE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Resale.s0**.
- Click **Open**.

2 Open the Error-Bar Charts window.

- On the menus, select **Graphics**, then **Error-Bar Charts**. The Error-Bar Charts procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Error-Bar Charts window, select the **Variables tab**.
- Double-click in the **Response Variable(s)** text box. This will bring up the variable selection window.
- Select **Price** from the list of variables and then click **Ok**. "Price" will appear in the Response Variables box.
- Double-click in the **Horizontal Grouping Variable** text box. This will bring up the variable selection window.
- Select **Bedrooms** from the list of variables and then click **Ok**. "Bedrooms" will appear in the Horizontal Variables box.
- Double-click in the **Legend (Subgroup) Variable** text box. This will bring up the variable selection window.
- Select **State** from the list of variables and then click **Ok**. "State" will appear in the Legend Variables box.

4 Specify the vertical decimal places.

- Click on the **Axes** tab.
- Click the **Tick Label Settings…** button under **Vertical Axis**. This will bring up a window that controls the vertical reference numbers.
- In **Decimals**, select **0**.

5 Specify the legend.

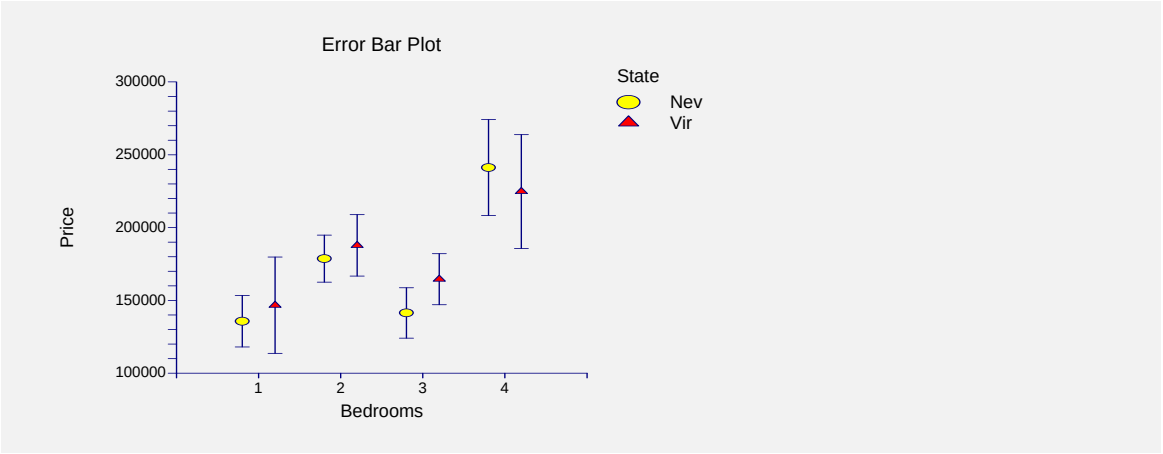
- Click on the **Titles and Misc. tab**.
- Check the **Show Legend** check box to indicate that the legend should be shown on the plot.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

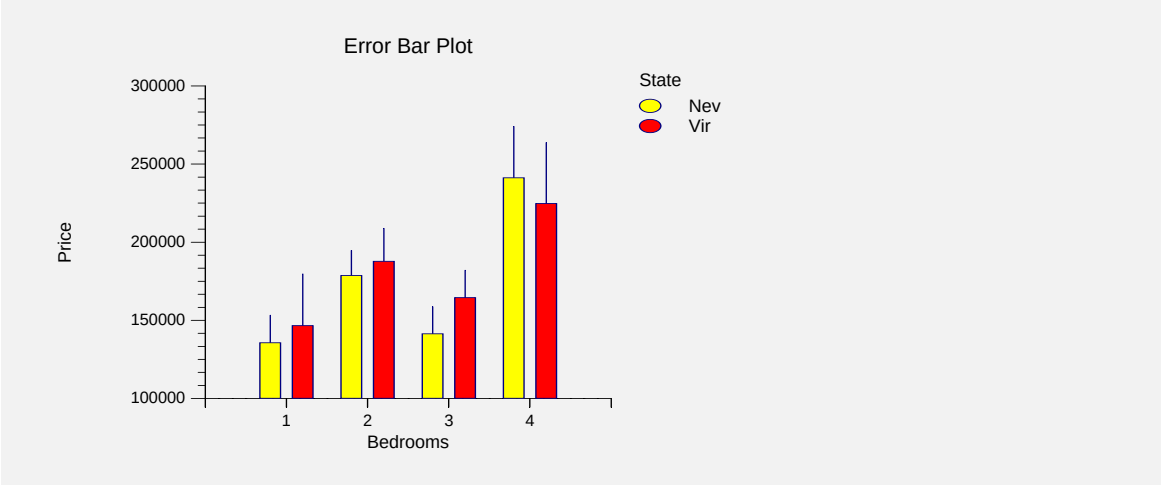
Error-Bar Plot Output

The following chart will be displayed in the Output window.



Note that the symbols mark the mean of each group. The lines extending up and down about the symbol represent the variation in the data. With default values, these lines are one standard error in either direction.

Switching the Plot Type to Bar Plot and rerunning this example produces the following plot:

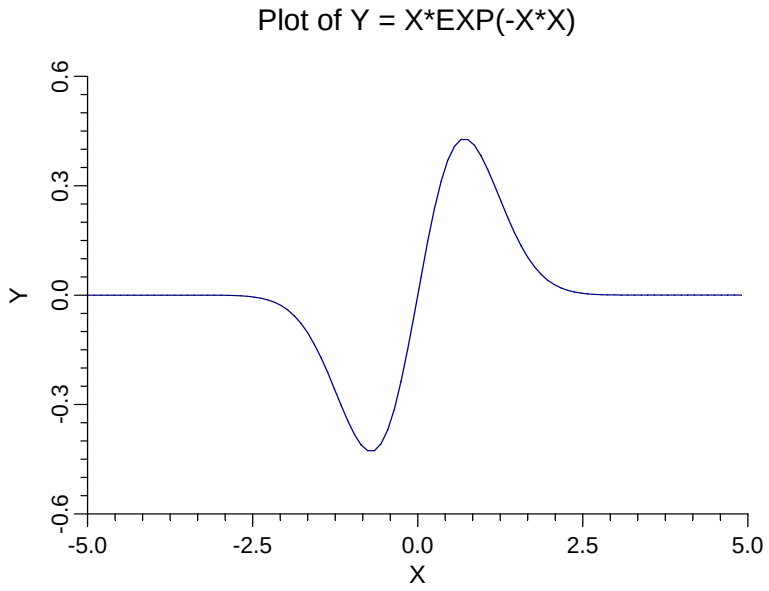


Chapter 160

Function Plots

Introduction

This program draws graphs of user specified functions. You define the function using standard mathematical syntax and set the range over which the function should be drawn. It is one of the few procedures that does not accept (or use) data.



Procedure Options

This section describes the options available in this procedure.

Model Tab

This panel specifies function that is to be plotted.

Function

Formula

This option contains the equation to be plotted. Note that you do not enter the “Y=” portion of the expression.

The expression is made up of

1. Symbols: +, -, *, /, ^, <, >, =, (, and).
2. Functions:

ABS(X)	Absolute value of X
ASN(X)	Arc sine of X
ATN (X)	Arc tangent of X
COS(X)	Cosine of X
EXP(X)	Exponential of X
INT(X)	Integer part of X
LN(X)	Log base e of X
LOG(X)	Log base 10 of X
SGN(X)	Signature of X
SIN(X)	Sine of X
SQRT(X)	Square root of X
TAN(X)	Tangent of X
TNH(X)	Hyperbolic Tangent of X

3. The horizontal variable labeled X.
4. Constants.

The syntax of the model expression follows that of the variable transformations, so we will not go into syntax here, but refer you to the Transformations chapter. Note that only a subset of the functions available as transformations are also available here.

Examples of valid models are

```
2+3*X^4
1+2*EXP(-3*X)
(5 + 3*X + 2*X) / (1 + 2*X + 2*X)
```

Function Evaluations

This option specifies at how many points the function is calculated to create the line that is displayed. A value between 100 and 200 is usually sufficient.

Line

Specify the color, width, and pattern of the function line.

Vertical and Horizontal Axis

Label

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**
  Specifies the color of the tick labels.
- **Font Size**
  Specifies the size of the tick labels.
- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.
- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.
- **Max Characters**
  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).
- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Minor Ticks (number)

Tick labels are displayed for the minor tickmarks. This option specifies the number of minor tickmarks displayed along the axis.

Show Grid Lines

Checking this option causes the major grid lines to be displayed.

Function Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Titles

Plot Title

This is the text of the title. The characters *{M}* are replaced by the formula. Press the button on the right of the field to specify the font of the text.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Function Plot

This section presents an example of how to generate a function plot. In this example, we will plot the function Y=X*EXP(-X*X). You may follow along here by making the appropriate entries or load the completed template EXAMPLE1 from the Load Template option of the File Menu.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Function Plots window.

1 Open the Function Plots window.

- On the menus, select **Graphics**, then **Other Charts and Plots**, then **Function Plots**. The Function Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

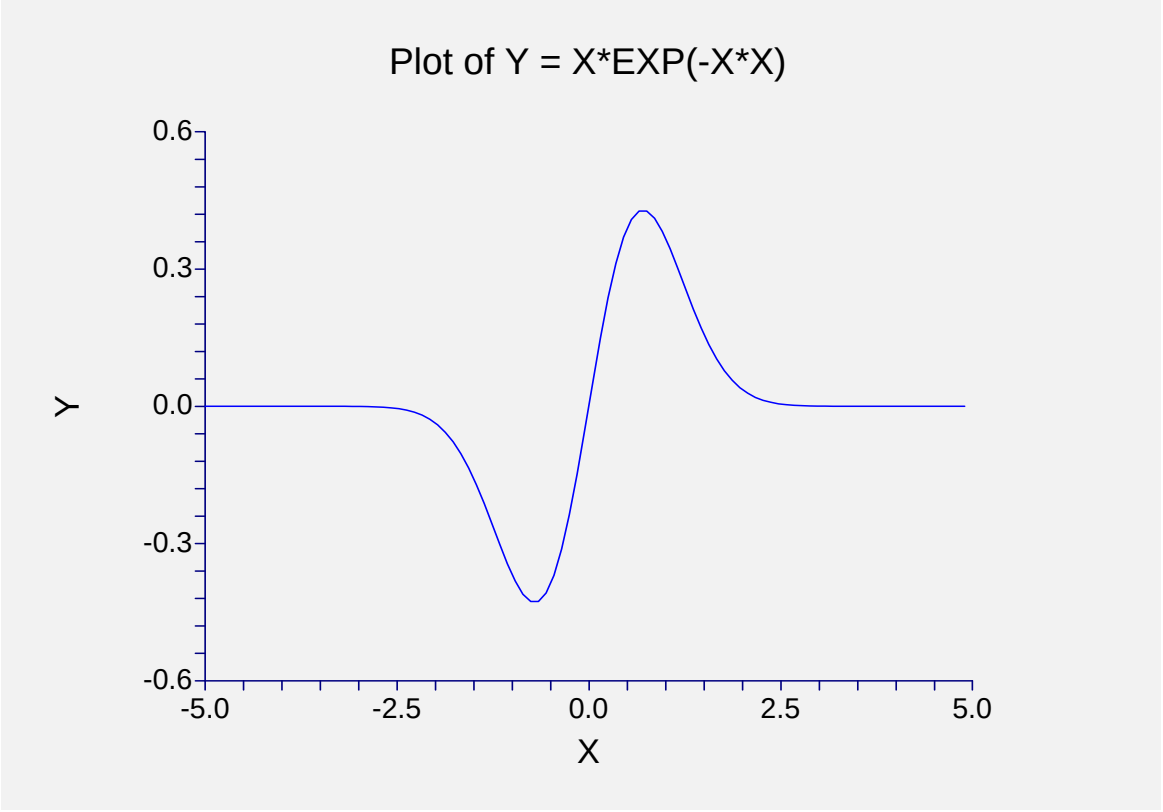
2 Specify the model.

- On the Function Plots window, select the **Model tab**.
- In the **Formula** text box, enter the function **X*Exp(-X*X)**

3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Function Plot Output

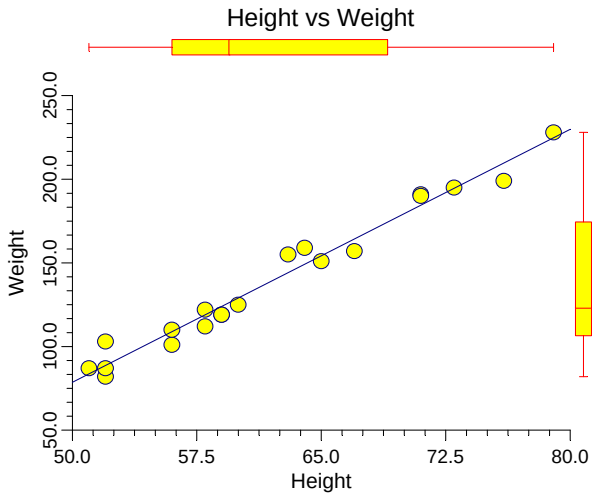


Chapter 161

Scatter Plots

Introduction

The x-y scatter plot is one of the most powerful tools for analyzing data. **NCSS** includes a host of features to enhance the basic scatter plot. Some of these features are trend lines (least squares) and confidence limits, polynomials, splines, lowess curves, imbedded box plots, and sunflower plots. Following is an example of a typical scatter plot with a trend line and imbedded box plots.



Data Structure

A scatter plot is constructed from two variables. A third variable may be used to divide the first two variables into groups (e.g., age group or gender). No other constraints are made on the input data. Note that rows with missing values in one of the selected variables are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are in the scatter plot.

Variables

Vertical Variable(s)

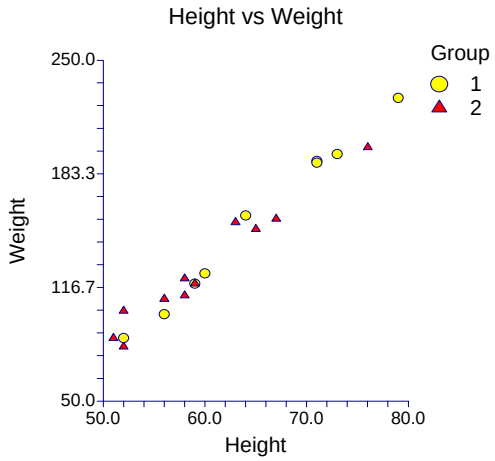
Enter one or more vertical variables. If more than one variable is entered, the number of plots is determined by the *Overlay* option.

Horizontal Variable(s)

Enter one or more horizontal variables. If more than one variable is entered, the number of plots is determined by the *Overlay* option.

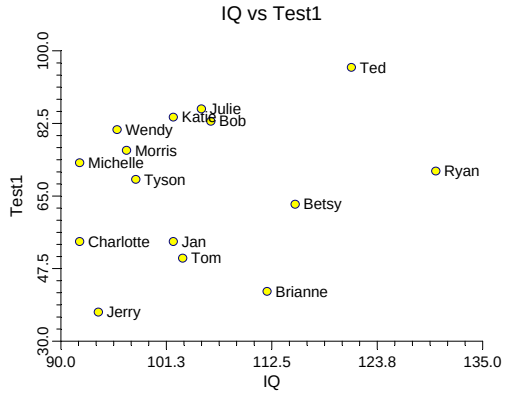
Grouping (Symbol) Variable

This variable may be used to separate the observations into groups. For example, you might want to use different plotting symbols to distinguish observations from different states. You designate the grouping variable here. The appearance of the plot symbol is designated on the Symbols Tab. Labels may be automatically substituted for values if the Value Labels option is set to Value Labels.



Data Label Variable

A data label is text that is displayed beside each point. A variable containing the data labels. The values may be text or numeric. You can use dates (like Jan-23-95) as labels. Here is how. First, enter your dates using the standard date format (like 06/20/93). In the Variable Info screen, change the format of the date variable to something like *mmm-dd-yyyy* or *mm-dd-yy*. The labels will be displayed as labels. Without changing the variable format, the dates will be displayed as long integer values.



The size, style, and color of the text may be modified by pressing the second button to the right of the text box. This button brings up the text settings window.

Plot Overlay

This option is used when multiple vertical and/or horizontal variables are entered to specify whether to overlay specified plots onto a single plot. Possible choices are:

- **None**

  No plots are overlaid. Each combination of horizontal and vertical plots produces its own separate plot.

- **Multiple Vertical**

  All vertical variables are combined onto one plot. A separate plot is drawn for each horizontal variable.

- **Multiple Horizontal**

  All horizontal variables are combined onto one plot. A separate plot is drawn for each vertical variable.

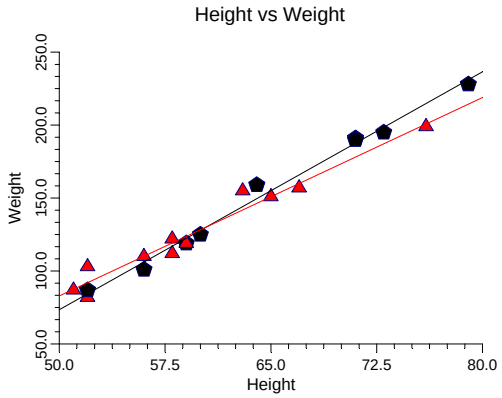
- **Series: No Overlay**

  A series of plots are drawn using each vertical and horizontal variable in sequence. The first vertical variable is matched with the first horizontal variable, the second vertical with the second horizontal, and so on.

Symbols

These options set the type, color, size, and style of the plotting symbols. Symbols for up to fifteen groups may be used. When no Group Variable is specified, the options of Symbol 1 are used to define the plot symbol.

Following is an example of possible symbol settings for two groups:



Double-clicking the symbol, or clicking the button to the right of the symbol, brings up the symbol specification window. This window lets you specify the characteristics of a symbol in detail.

- **Symbol Fill Color**

  The color of the symbol's interior (fill region). Many of the plotting symbols, such as a circle or square, have both an interior region and a border. This specifies the color of the interior region. The border is specified as the Symbol Outline.

- **Symbol Outline Color**

  The color of the symbol's border. This color is used when the symbols are connected by a connecting line.

- **Symbol Type**

  This option designates the shape of the plot symbol. The most popular symbols may be designated by pressing the appropriate button. About 80 symbol types are available, including letters and numbers.

- **Symbol Radius**

  The size (radius) of the symbol.

- **Symbol Fill Pattern**

  The pattern (solid, lines, etc.) of the symbol's interior (fill region).

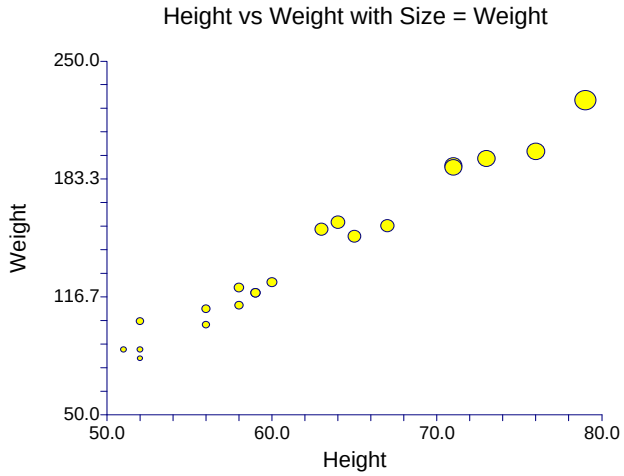
- **Symbol Border Width**

  The width of the symbol's border.

Symbols – Symbol Size Options

Symbol Size Variable

This option designates a third variable that is represented by the size the plotting symbol. Sometimes, this is referred to as a bubble chart. Here is an example of such a plot:



The minimum and maximum values of this variable are associated with the smallest and largest plot symbols. The rest of the points fall in between. The size of the smallest and largest points is controlled by the Minimum and Maximum options explained next.

Minimum Symbol Size

This is the size of the smallest plotting symbol—the one that represents the minimum data value. This number represents a percentage adjustment to the normal size of the plot symbol. Hence, the default value of 50 means that the diameter of the plotting symbol is one half of normal. Note that normal is represented by 100.

Maximum Symbol Size

This is the size of the largest plotting symbol—the one that represents the maximum data value. This number represents a percentage adjustment to the normal size of the plot symbol. Hence, the default value of 200 means that the diameter of the plotting symbol is twice that of normal. Note that normal is represented by 100.

Axes Tab

These options specify the characteristics of the vertical and horizontal axes.

Vertical and Horizontal Axis

Label Text

This box supplies the axis label. The characters {X}, {Y}, and {G} are replaced by the horizontal, vertical, and grouping variable names, respectively. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Minimum

Specifies the smallest value shown on this axis.

Maximum

Specifies the largest value shown on this axis.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line and its type (numeric or text).

Log Scale

This option lets you select logarithmic scaling for this axis.

- **No**

  Use normal scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as numbers (e.g., 1, 10, 100, 1000).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (-2, 1, 0, 1, 2, 3).

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**
  Specifies the color of the tick labels.
- **Font Size**
  Specifies the size of the tick labels.
- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.
- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.
- **Max Characters**
  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).
- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical and Horizontal Axis – Positions

Axis

This option controls the position of the axis: if and where it is displayed.

Label

This option controls the position of the label: if and where it is displayed.

Tick Labels

This option controls the position of the tick labels: if and where they are displayed.

Tickmarks

This option controls the position of the tickmarks: if and where they are displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the scatter plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. The characters {A} and {B} are replaced by the numeric values of the intercept and slope of the regression line, respectively. To display the fitted regression equation, you could use {Y} = {A} + ({B}){X}.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable’s name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Legend

When data for more than one group are displayed, a legend is desirable. These options specify the legend.

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. The characters *{G}* will be replaced by the name of the group variable. Click the button on the right to specify the font size, color, and style of the legend text.

Variable Data Transformations

These options specify automatic transformations of the data for either variable.

Transform Exponent

Each value of the variable is raised to this exponent. Note that fractional exponents require positive data values.

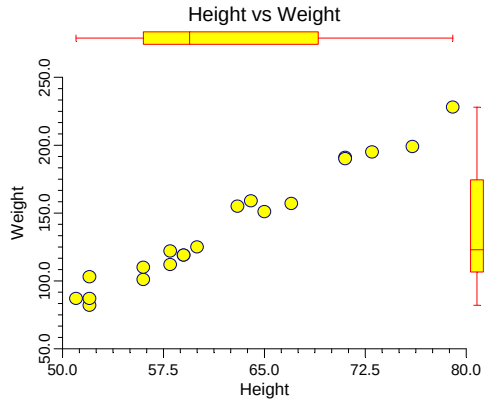
Additive Constant

This constant is added to the variable. This option is often used to make all values positive.

Box & Dot Plot Tab

Box Plots

Box plots may be placed in the vertical and horizontal margins of the scatter plot. These box plots emphasize the univariate behavior of the corresponding variable.



Shape

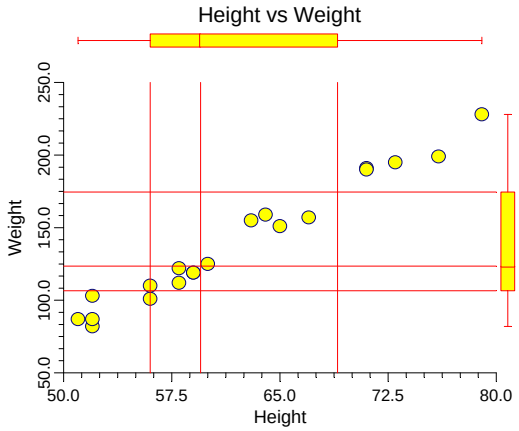
This option specifies the shape of the box plot. See the *Box Plot* chapter for further details.

Percentile Type

Specifies the formula used to calculate the percentile. See the *Box Plot* chapter for further details.

Reference Lines

Reference lines may be extended across the plot at each of the three quartiles that are used to form the box.



Inner and Outer Fences

These multipliers are used to construct the fences for designating outliers. See the *Box Plot* chapter for further details.

Line Width

This option specifies the width of the border line.

Box Width

This option specifies the width of the box itself.

Position

This option indicates on which side of the plot the box plot should be displayed.

Show Outliers

This option indicates whether the outlying points should be displayed.

Fill Color

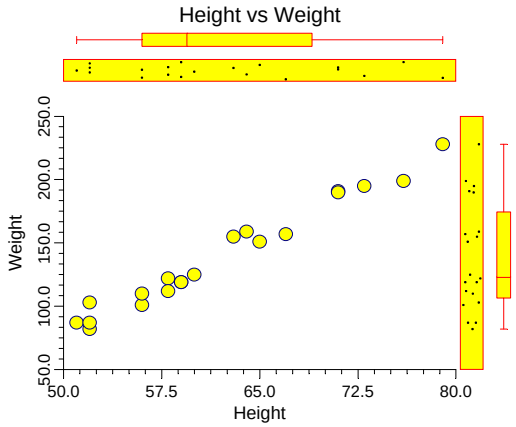
This option specifies the interior color of the box plot.

Line Color

This option specifies the color of the box border and the lines.

Dot Plots

A dot plot may be placed in the vertical and horizontal margins of the scatter plot. The dot plot lets you study the distribution of the corresponding variable by plotting the actual data values in a line plot.



Position

Specifies whether to display the dot plot and in which margin to place it in.

Box Width

Specifies the width of the dot plot.

Dot Color and Size

Specifies the color and size of the data points displayed in the dot plot.

Line Width and Color

Specifies the width and color of the dot plot's border.

Fill Color

The color of the data point interior (fill region). Many of the plotting symbols, such as a circle or square, have an interior region and a border. This specifies the color of the interior region.

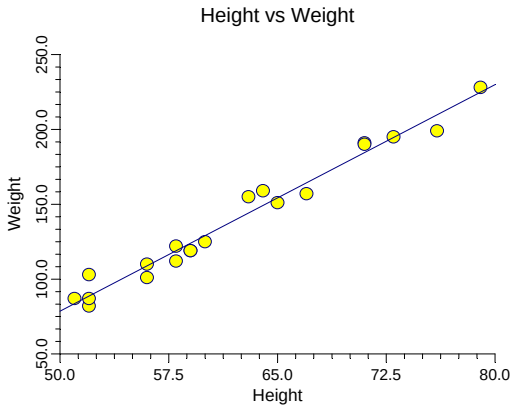
Lines 1 Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line. Several of the lines may be further defined using the pull-down box to the left of the line.

Regression

Regression Line

This option governs the display of a regression line (least-squares trend line or line of best fit) through the data points.



Regression Estimation

This option specifies the way in which the trend line is calculated.

- **L.S.**

  The standard least squares regression line is calculated. This formulation is popular but can suffer from severe distortion if one or more outliers exist in your data.

- **Median**

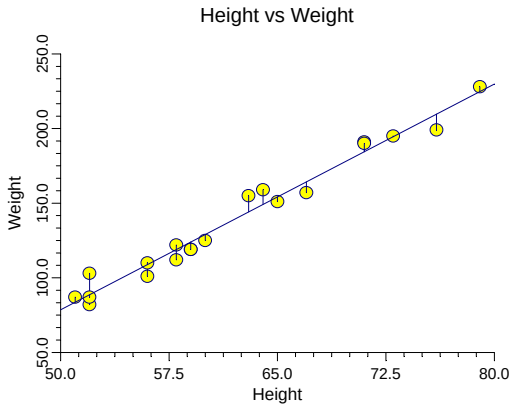
  A resistant least-squares algorithm is used to fit the line, and the intercept is adjusted so that the line passes through the medians of the horizontal and vertical variables. This algorithm is resistant because it removes most of the distortion caused by a few outliers (atypical points).

- **Quartiles**

  A resistant least-squares algorithm is used to fit the line, and the intercept is adjusted so that the line passes through the first and third quartiles of the horizontal and vertical variables. This algorithm is resistant because it removes most of the distortion caused by a few outliers (atypical points).

Residual Lines

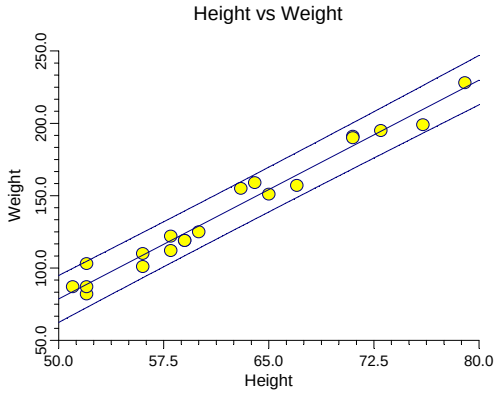
The residual is the vertical deviation of each point from the regression line. These residuals may be displayed as vertical lines that connect each plot point to the regression line.



Confidence Limits

C.L. Individuals (or Means)

Confidence limits for the regression line may be displayed. These limits are for either the mean of a future group of individuals with a common horizontal value or for a single individual. The most commonly used confidence interval is for a future individual. The confidence limits for the mean are for specialized use only. Note that these confidence lines are formed by calculating individual confidence limits at various points along the horizontal axis and connecting those points. These are not the same as confidence bands.



Confidence Limit Alpha

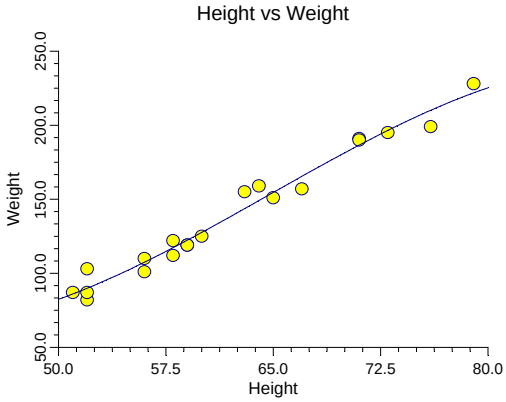
This option gives the confidence level, α, of the $100(1-\alpha)\%$ confidence limits. For example, if α is set to 0.05, you have a 95% confidence limit constructed at each point.

Polynomial Fit

Polynomial Fit

An n-degree polynomial line may be fit and displayed over the data. The number of points along the line that are calculated is controlled by the Number of Calculation Points setting.

If you wish to see the equation of this line, use the response surface regression procedure.



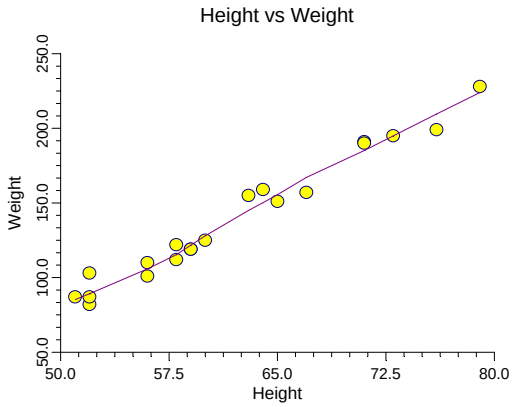
Polynomial Order

The order of the polynomial is the value of the largest exponent in the polynomial regression equation. Usually, values of two or three are used, since polynomials of larger order often exhibit strange behavior between data points.

LOESS

Loess Line

The locally weighted regression scatter plot smooth (*lowess* or *loess*) is a popular, computer-intensive technique that usually provides a reasonable smoothing of your data without being overly sensitive to outliers. A reasonable smooth is one that travels more or less through the middle of the data. The degree of smoothing is controlled by the loess options. The number of points at which the loess curve is computed is given by the *Intervals* option. The mathematical details of the loess method are given in the Linear Regression chapter.



Calculation Fraction

This is the percent of the sample points that are included in the computation for a particular value of the smooth. Most authors recommend 40% as the first value to try. This means that 40% of the data values are used in the computations for each loess smooth value.

LOESS Polynomial Order

This is the order of the polynomial fit in the loess procedure. Select ‘1’ for a linear fit or ‘2’ for a quadratic fit. The linear fit tends to be smoother. The quadratic fit tends to pick up peaks and values better.

Robust Iterations

This is the number of robust iterations used in the loess algorithm to downplay the influence of outliers. Select ‘0’ if you do not want robust iterations. This will show the impact of outliers on the loess curve. It will reduce the execution time in large datasets.

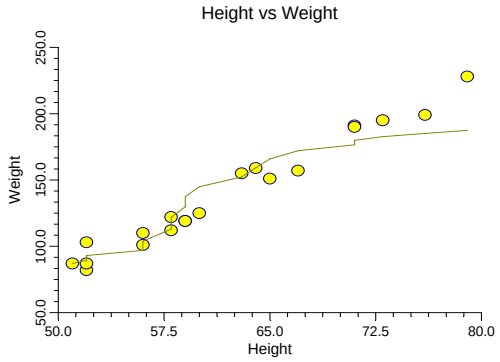
There is little reason for using more than two iterations.

Median Smooth

Median-Smooth Line

A median smooth line may be displayed over the data. This type of smooth line does well for level series (those with no vertical trend), but does not do well when a trend is apparent.

The median smooth is constructed by first ordering the observations by the horizontal variable. Next, running medians of *n* observations are found, where *n* is the Rows parameter.



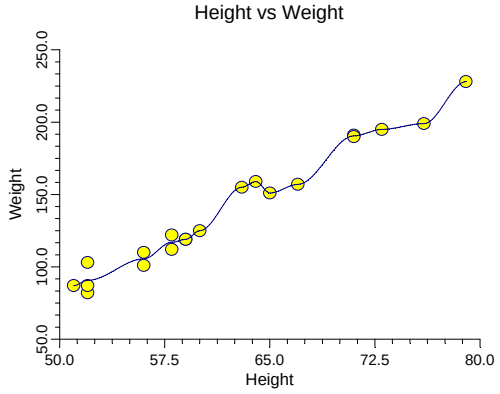
Rows in Smooth Calculation

The number of observations used to calculate the median at each step.

Spline

Spline Fit

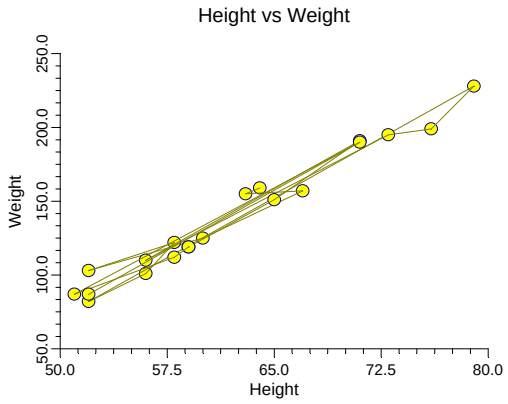
A cubic spline may be fit to the data.



Connect Points

Connect All Points

The points may be connected with a line sequentially, proceeding from the first row, to the second row, to the third row, and so on.



Calculation Points

Number of Calculation Points

The number of positions along the horizontal axis at which the confidence limits, LOESS, and splines are calculated.

Lines 2 Tab

These options allow certain reference lines to be specified and displayed. For each line, you can click the line or the button on the right of the line to bring up a window that specifies the color, width, and pattern of the line. The check box to the left of the line name indicates whether to display the line.

Horizontal Lines from the Vertical Axis

Line at Mean of Vertical Variable

Display a line at the mean value of the vertical (horizontal) variable.

Line from Vertical Axis to Data Points

This option lets you display individual lines from the data points to the vertical (or horizontal) axis.

Horizontal Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Vertical Lines from the Horizontal Axis

Line at Mean of Horizontal Variable

Display a line at the mean value of the vertical (horizontal) variable.

Line from Horizontal Axis to Data Points

This option lets you display individual lines from the data points to the vertical (or horizontal) axis.

Vertical Line at Value Below

These options let you display lines at particular values. The value is specified to the right of the line settings.

Axis Tickmarks for Data Points

Vertical Axis Ticks for each Data Point

Display tick marks for each data point along the vertical axis.

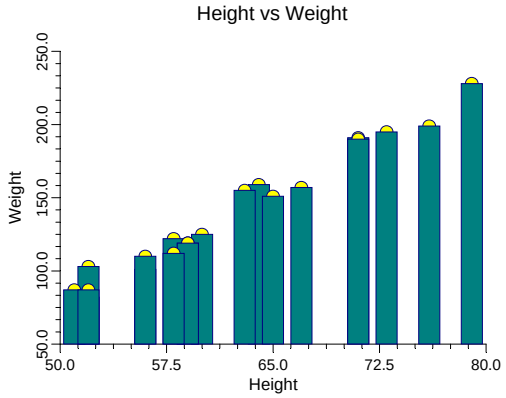
Horizontal Axis Ticks for each Data Point

Display tick marks for each data point along the horizontal axis.

Bars and Sunflower Plot Tab

Bars from Data Points to Horizontal (Vertical) Axis

A line or bar can be displayed that will connect each point to the horizontal or vertical axes. This option is useful if you want to generate a “bar chart” look to your scatter plot. Such a plot emphasizes the horizontal and vertical differences among the data points.



Bar Direction

This option controls where the bar extend to.

- **None**
  No bars are drawn.
- **Down**
  Horizontal bars are drawn from the points down to the bottom of the plot.
- **Up**
  Horizontal bars are drawn up from the data points to the top of the plot.

- **Left**
  Vertical bars are drawn from the points to the left axis.
- **Right**
  Vertical bars are drawn from the data points to the right axis.
- **Both**
  Bars are drawn in both directions.

Bars from Data Points to Horizontal (Vertical) Axis – Fill

Fill Color and Pattern

These options control the fill colors and patterns of the inside of the bars.

Bars from Data Points to Horizontal (Vertical) Axis – Outline

Outline Color and Width

These options control the color and width of the outline of the bar.

Bars from Data Points to Horizontal (Vertical) Axis – Width

Select Bar Width Parameter

Specify the method used to set the width of the bars as either Amount or Percent Space.

Amount

Designates a specific width for each bar.

Percent Empty Space

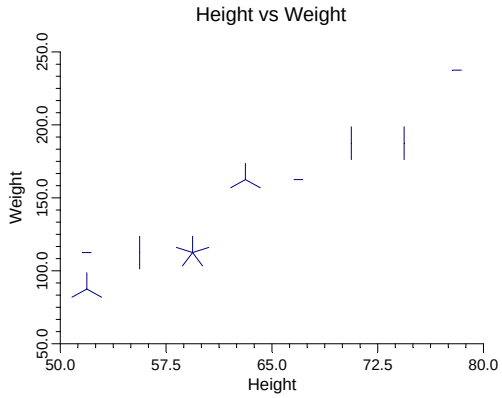
The percent of the horizontal axis length that is space instead of bars. This value determines the width of the bars. The smaller this value, the larger the bar width. Also note that this parameter only works for non-overlapping bars.

Sunflower Plot

The sunflower plot is used when you have so many observations that all you see on your scatter plot is a blob. Because of the large amount of overprinting, you cannot see many of the subtle patterns that occur in your data. Also, the eye tends to concentrate on outliers rather than the body of the data. The sunflower plot summaries the scatter plot by grouping the data so that you only see a few plot points. The algorithm is as follows:

1. Partition the plot with a two-way grid. The number of rows and columns in the grid determines the degree of smoothing.
2. Count the number of points in each cell of the grid. This is the amount that is plotted.

3. Plot a flower according to the following rules: If there is only one point in the cell, plot a point in the center of the cell. If there are more than one observation, make a “flower” with one petal for each point. Arrange the petals evenly about the center of the cell.



Count Pattern

This option specifies whether the sunflowers are displayed and which type of plot to display. Note that the data points will be displayed on top of the sunflowers unless you omit them by changing the Symbol 1 to None.

- **None**

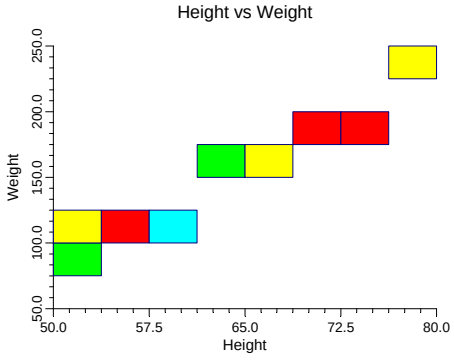
  Selecting this option omits the sunflowers from the scatter plot.

- **Spokes**

  Selecting this option requests the display of the standard sunflower plot.

- **Background Colors**

  Selecting this option requests a variation of the sunflower plot in which the sunflowers are not shown. Instead, the cells of the grid are displayed as rectangles in various colors. The cell’s color is determined by the number of points that fall into it.



Vertical Bins

The number of vertical bins (slices or intervals) that are used. You will have to try a few different values for each plot to find the one that best represents the data.

Horizontal Bins

The number of horizontal bins (slices or intervals) that are used.

Petal Length

This is a percentage adjustment to the length of the petal. A value of 100 here indicates that the petals are to run to the edge of the cell. Percentage values of less than 100 reduce the length of the petals. We have found that a value of 90 works well in most cases.

Maximum Petals

Your data may contain cells that contain hundreds of data points. This option lets you specify a maximum number of petals. Cells with a count greater than this number still display only this many petals.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Scatter Plot

This section presents an example of how to generate a simple scatter plot. The data used are from the SAMPLE database. We will create a scatter plot of variables *Weight* versus *Height*.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Scatter Plots window.

1 Open the SAMPLE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.

2 Open the Scatter Plots window.

- On the menus, select **Graphics**, then **Scatter Plots**. The Scatter Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

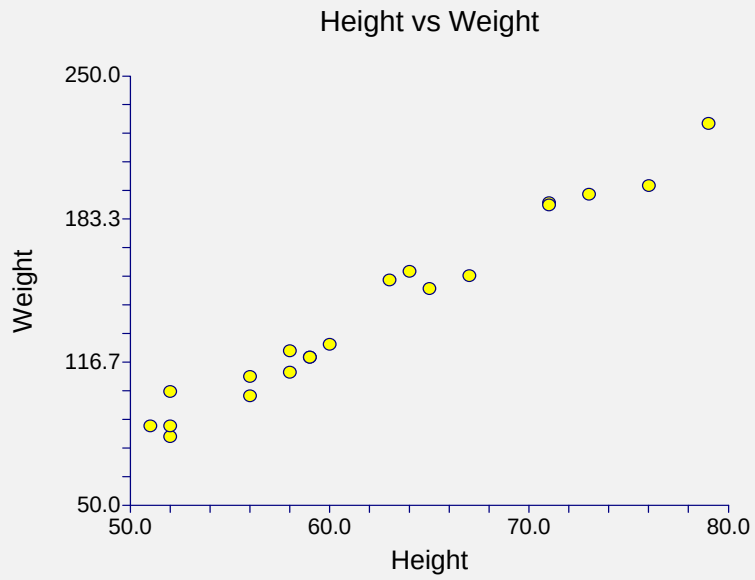
3 Specify the variables.

- On the Scatter Plots window, select the **Variables tab**.
- Double-click in the **Vertical Variable(s)** text box. This will bring up the variable selection window.
- Select **Weight** from the list of variables and then click **Ok**. "Weight" will appear in the Vertical Variable(s) box.
- Double-click in the **Horizontal Variable(s)** text box. This will bring up the variable selection window.
- Select **Height** from the list of variables and then click **Ok**. "Height" will appear in the Horizontal Variable(s) box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Scatter Plot Output



Creating a Scatter Plot Style File

One of the most exciting features of **NCSS** is its ability to mix graphics and text on the output reports. Many of the statistical procedures include scatter plots, box plots, or other graphs as part of their reports. Each of these graphs have over 200 options, so adding the options necessary to specify each graph would greatly increase the number of options that you would have to specify.

To overcome this, we let you create and save scatter plot style files. These style files contain the current settings of all options. When you use the style file in another procedure, such as the Multiple Regression, you only have to set a few of the options. Most of the options come from this style file. A default scatter plot style file was installed with the **NCSS** system. Other style files may be added.

We will now take you through the steps necessary to create a Scatter Plot Style file.

1 Open the SAMPLE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.
- Note: You do not necessarily have to use the SAMPLE database. You can use whatever database is easiest for you. Just open the appropriate database here.

2 Open the Scatter Plots window.

- On the menus, select **Graphics**, then **Scatter Plots**. The Scatter Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the horizontal and vertical variables.

- On the Scatter Plots window, select the **Variables tab**.
- Double-click in the **Vertical Variable(s)** text box. This will bring up the variable selection window.
- Select **Weight** from the list of variables and then click **Ok**. “Weight” will appear in the Vertical Variable(s) box.
- Double-click in the **Horizontal Variable(s)** text box. This will bring up the variable selection window.
- Select **Height** from the list of variables and then click **Ok**. “Height” will appear in the Horizontal Variable(s) box.
- Note: don’t worry about the specification of these variables. The actual variable names are not stored in the style file. They are used here so that you can see what your style will look like.

4 Set your options.

- Set the various options of the scatter plot’s appearance to the way you want them.
- Run the procedure to generate the scatter plot. This gives you a final check on whether the scatter plot appears just how you want it. If it does not appear quite right, go back to the panel and modify the settings until it does.

5 Save the template (optional).

- Although this step is optional, it will usually save a lot of time and effort later if you store the current template. Remember, the template is not the style file.
- To store the template, select the **Template tab** on the Scatter Plots window.
- Enter an appropriate name in the **File Name** box.
- Enter an appropriate phrase at the bottom of the window in the **Template Id** (the long box across the bottom of the Scatter Plot's window). This phrase will be displayed in the Template Id's box to help you identify the template files.
- Select **Save Template** from the File menu. This will save the template.

6 Create and Save the Style File.

- Select **Save Style File** from the File menu. The Save Style File Window will appear.
- Enter an appropriate name in the **Selected File** box. You can either reuse one of the style files that already exist or create a new name. You don't have to worry about drives, directory names, or file extensions. These are all added by the program. Just enter an appropriate file name.
- Press the **Ok** button. This will create and save the style file.

7 Using a Style File.

- Using the style file is easy. For example, suppose you want to use this style file to plot residuals in the Multiple Regression procedure. You do the following:
- Select the **Plot Options tab** in the Multiple Regression procedure.
- Click the button to the right of the **Probability Plot - Plot Style File** box (the initial file name is Default). This will bring up the Scatter Plot Style File Selection window.
- Click on the appropriate file so that it is listed in the Selected File box. Click the **Ok** button.
- The new style file name will appear in the Plot Style File box of the Multiple Regression window. That's it. Your new style has been activated.

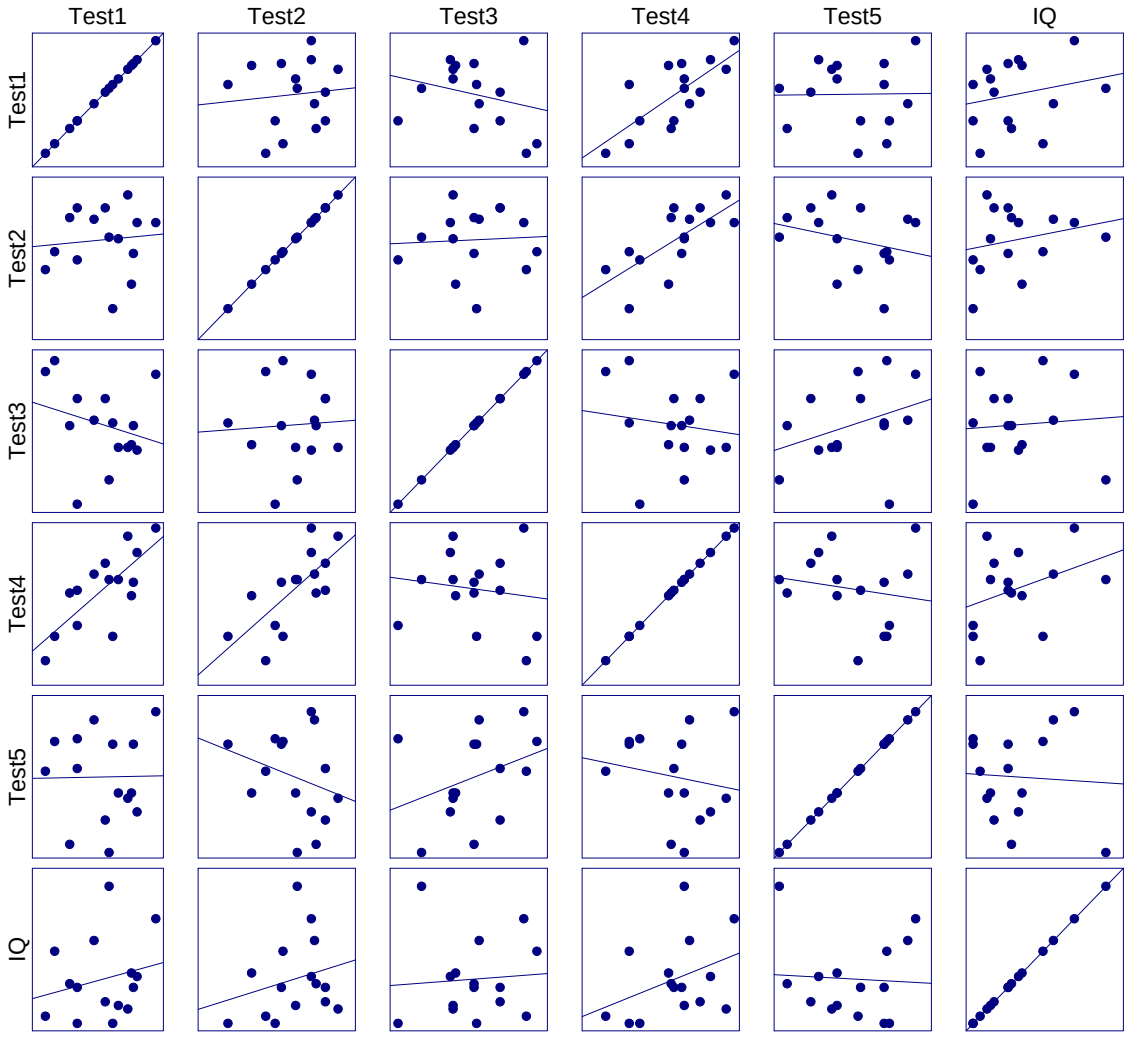
Chapter 162

Scatter Plot Matrix

Introduction

A *scatter plot matrix* is table of scatter plots. Each plot is small so that many plots can be fit on a page. When you need to look at a lot of plots, such as at the beginning of a multiple regression analysis, a scatter plot matrix is a very useful tool.

Following is an example of a scatter plot matrix created during the initial phase of a multiple regression study. Five test scores form the pool of independent variables and the subjects IQ value is the dependent variable. Notice how quickly you can scan the plots for highly correlated variables and for outliers.



Data Structure

Each scatter plot is constructed from two numeric variables. A third, alphanumeric, variable may be used to control the plot symbol.

Procedure Options

Most of the options are identical to those of the Scatter Plot procedure and so they will not be repeated here. We will concentrate on those options that are unique to this procedure.

Variables Tab

This panel specifies which variables are in the scatter plot.

Variables

Vertical Variable(s)

Enter two or more variables to be displayed vertically down the page. If you leave this field blank, the horizontal variables will be used as the vertical variables as well.

Horizontal Variable(s)

Enter two or more variables to be displayed horizontally across the page. If you leave the Vertical Variable(s) field blank, these variables will also be used as the vertical variables.

Grouping (Symbol) Variable

This variable may be used to designate different plotting symbols for different rows of data. For example, you might want to use different plotting symbols to distinguish observations from different groups. The appearance of each symbol is designated on the Symbols Tab.

The symbol variable may also be used to “watch” certain rows throughout the dataset. For example, suppose that two or three rows appear as outliers in one of the plots. It might be useful to observe these rows in other plots. You would do this as follows. Add a new variable to your dataset that consists of zeros for all rows except the two or three of interest. Set the values of those rows to one and specify this new variable as the Symbol Variable. You will be able determine where those rows occur in all the plots in the scatter plot matrix.

Data Label Variable

A data label is text that is displayed beside each point. This field contains the name (or number) of the variable containing the labels. The values may be text or numeric. The size, style, and color of the text may be modified by pressing the second button to the right of the text box. This button brings up the text settings window.

Symbols

These options set the type, color, size, and style of the plotting symbols. Symbols for up to fifteen groups may be used. When no Group Variable is specified, the options of Symbol 1 are used to define the plot symbol.

Symbols – Symbol Size Options

Symbol Size Variable

The variable controls the size of the plot symbol. The size of the smallest and largest points are controlled by the Minimum and Maximum options explained next.

Minimum Symbol Size

This is the size of the smallest plotting symbol—the one that represents the minimum data value. This number represents a percentage adjustment to the normal size of the plot symbol. Hence, the default value of 50 means that the diameter of the plotting symbol is one half of normal. Note that normal is represented by 100.

Maximum Symbol Size

This is the size of the largest plotting symbol—the one that represents the maximum data value. This number represents a percentage adjustment to the normal size of the plot symbol. Hence, the default value of 200 means that the diameter of the plotting symbol is twice that of normal. Note that normal is represented by 100.

Layout Tab

These options control the general appearance of the scatter plot matrix.

Axes Labels

These options specify which individual plots will include axes labels (variable names). You can display them on the plots on the left, the right, the top, or the bottom of the matrix. You can display the labels on all plots by checking the On All Plots options.

Box Plots

These options specify which individual plots will include box plots. You can display them on the plots on the left, the right, the top, or the bottom of the matrix. You can display the box plots on all plots by checking the On All Plots options. Note that you cannot display both right and left box plots or top and bottom box plots on the same graph.

Tick Labels

These options specify which individual plots will include tick labels. You can display them on the plots on the left, the right, the top, or the bottom of the matrix. You can display the tick labels on all plots by checking the On All Plots options.

These options are used in conjunction with Major Ticks, Minor Ticks, and Tickmarks Position on the Axes tab.

Note that displaying tick labels may drastically alter the size of the individual plot.

Dot Plots

These options specify which individual plots will include dot plots. You can display them on the plots on the left, the right, the top, or the bottom of the matrix. You can display the dot plots on all plots by checking the On All Plots options. Note that you cannot display both right and left dot plots or top and bottom dot plots on the same graph.

Individual Plot Dimensions

Plot Height and Plot Width

These two options control the actual size of the plot. They allow you to set the size of the plots using a scale in which 1000 equals one inch. Set this option to 'Automatic' to set the program control the size of the plots.

Margins and Plots Per Row

Left and Right Margin

These options control the amount of space that is used to display the plots. Values are entered in hundreds of an inch. Thus a value of 100 means one inch.

Number of Plots Per Row

This option controls the number of plots per row (line) of output. If more horizontal variables are specified than can fit on a single row, the plot matrix is output in sections.

Axes Tab – Bars & Sunflower Tab

These options function as discussed in the Scatter Plot procedure, so they are not repeated here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Scatter Plot Matrix

This section presents a tutorial on generating a scatter plot matrix. To run this example, take the following steps. Note that the Example1 template will also run this example once the SAMPLE database has been loaded.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Error-Bar Charts window.

1 Open the SAMPLE dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **SAMPLE.S0**.
- Click **Open**.

2 Open the Scatter Plot Matrix window.

- On the menus, select **Graphics**, then **Scatter Plot Matrix**, then **Regular**. The Scatter Plot Matrix procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Scatter Plot Matrix window, select the **Variables tab**.
- Double-click in the **Horizontal Variables** text box. This will bring up the variable selection window.
- Select variables **Test1,Test2,Test3,Test4,Test5,IQ** from the list of variables and then click **Ok**. "Test1-IQ" will appear in this box.

4 Specify the label size.

- On the Scatter Plot Matrix window, select the **Axes tab**.
- Click the down arrow associated with **Label Text under Vertical Axis**.
- Change the **Font Size** to **36** and click **Ok**.
- Click the down arrow associated with **Label Text under Horizontal Axis**.
- Change the **Font Size** to **36** and click **Ok**.

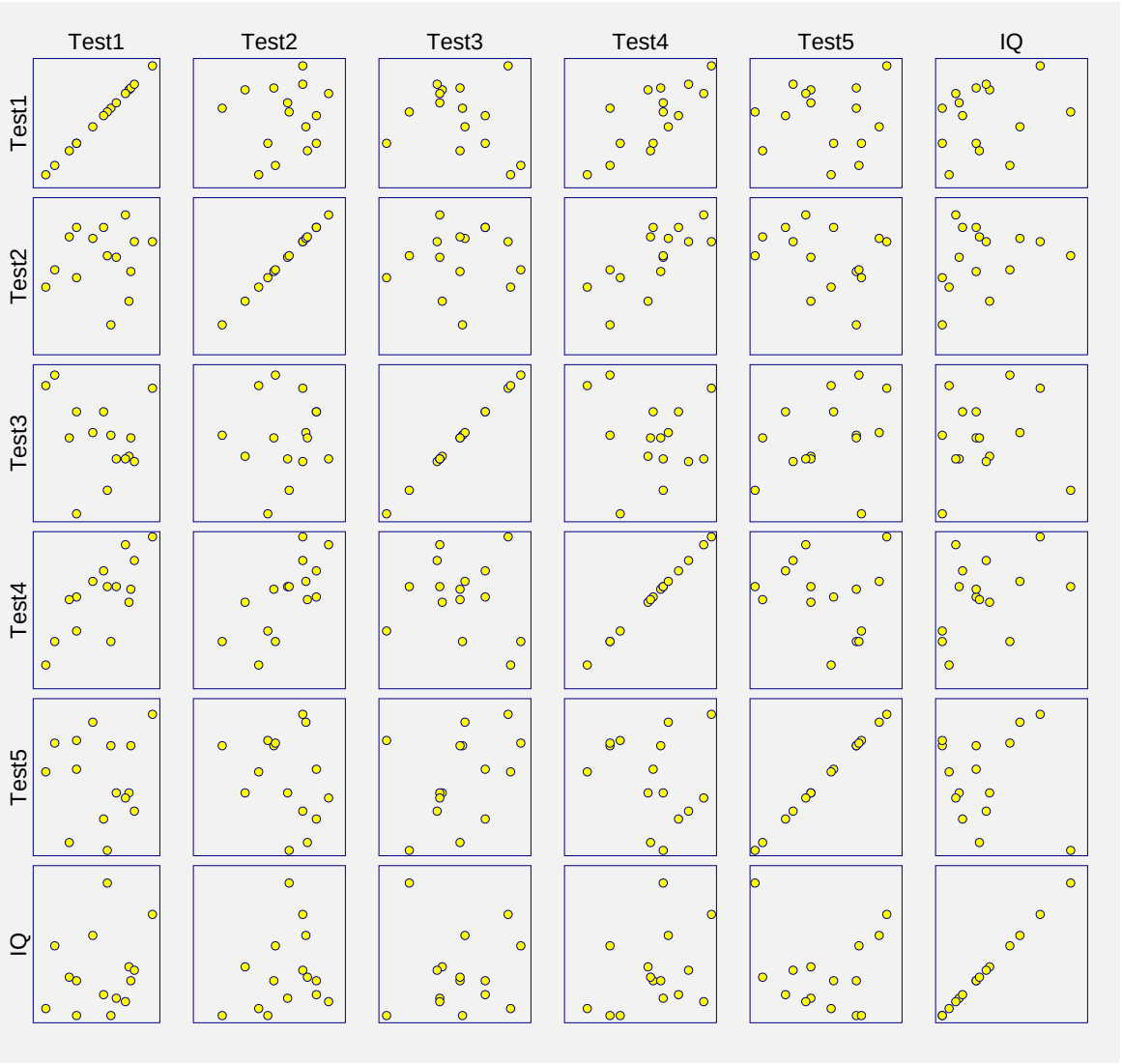
5 Specify the symbols.

- On the Scatter Plot Matrix window, select the **Variables tab**.
- Click the down arrow associated with **Group 1**.
- Change the Symbol Radius to **250** and click **Ok**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Scatter Plot Matrix Output



This report displays the scatter plot matrix.

Chapter 163

Scatter Plot Matrix for Curve Fitting

Introduction

One of the first tasks in curve fitting is to graphically inspect your data. This program lets you view scatter plots of various transformations of both X and Y. These plots are shown in matrix format.

You can look for transformations of both X and Y that give a simple relationship. Usually, your first choice would be to look for transformations of X and Y that yield a straight line. If these cannot be found, the next choice is to find functions that yield a recognizable curve.

Data Structure

The data are entered in two variables: one dependent (vertical) variable and one independent (horizontal) variable.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Y (Vertical) Variable

Variable

Specify the variable to be displayed on the vertical axis.

Y (Vertical) Variable – Select Y Transformations

1/(Y^2), 1/Y, 1/SQRT(Y), LN(Y), SQRT(Y), Y, and Y^2

Specifies whether this transformation of the Y should be plotted.

X (Horizontal) Variable

X Variable

Specifies a single independent (X) variable from the current database. This is the that will appear on the horizontal axis.

X (Horizontal) Variable – Select X Transformations

1/(X^2), 1/X, 1/SQRT(X), LN(X), SQRT(X), X, and X^2

Specifies whether this transformation of the X should be plotted.

Format Options

Plots Per Row

This option controls the size of the plots by specifying how many plots are to be shown on a row.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Vertical and Horizontal Axes

Label Text

This box supplies the vertical axis label. The characters {Y} and {G} are replaced by the corresponding variable names. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Minor Ticks (number)

Tick labels are displayed for the minor tickmarks. This option specifies the number of minor tickmarks displayed along the axis.

Show Grid Lines

Checking this option causes the major grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**
  Specifies the color of the tick labels.
- **Font Size**
  Specifies the size of the tick labels.
- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.
- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.
- **Max Characters**
  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).
- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Scatter Plot Matrix

This section presents an example of how to generate a scatter plot matrix. In this example, we will plot the variables Y1 and X1 of the FNREG3 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Scatter Plot Matrix for Curve Fitting window.

1 Open the FNREG3 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FNREG3.s0**.
- Click **Open**.

2 Open the Scatter Plot Matrix for Curve Fitting window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **Plots**, then **Scatter Plot Matrix**. The Scatter Plot Matrix for Curve Fitting procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

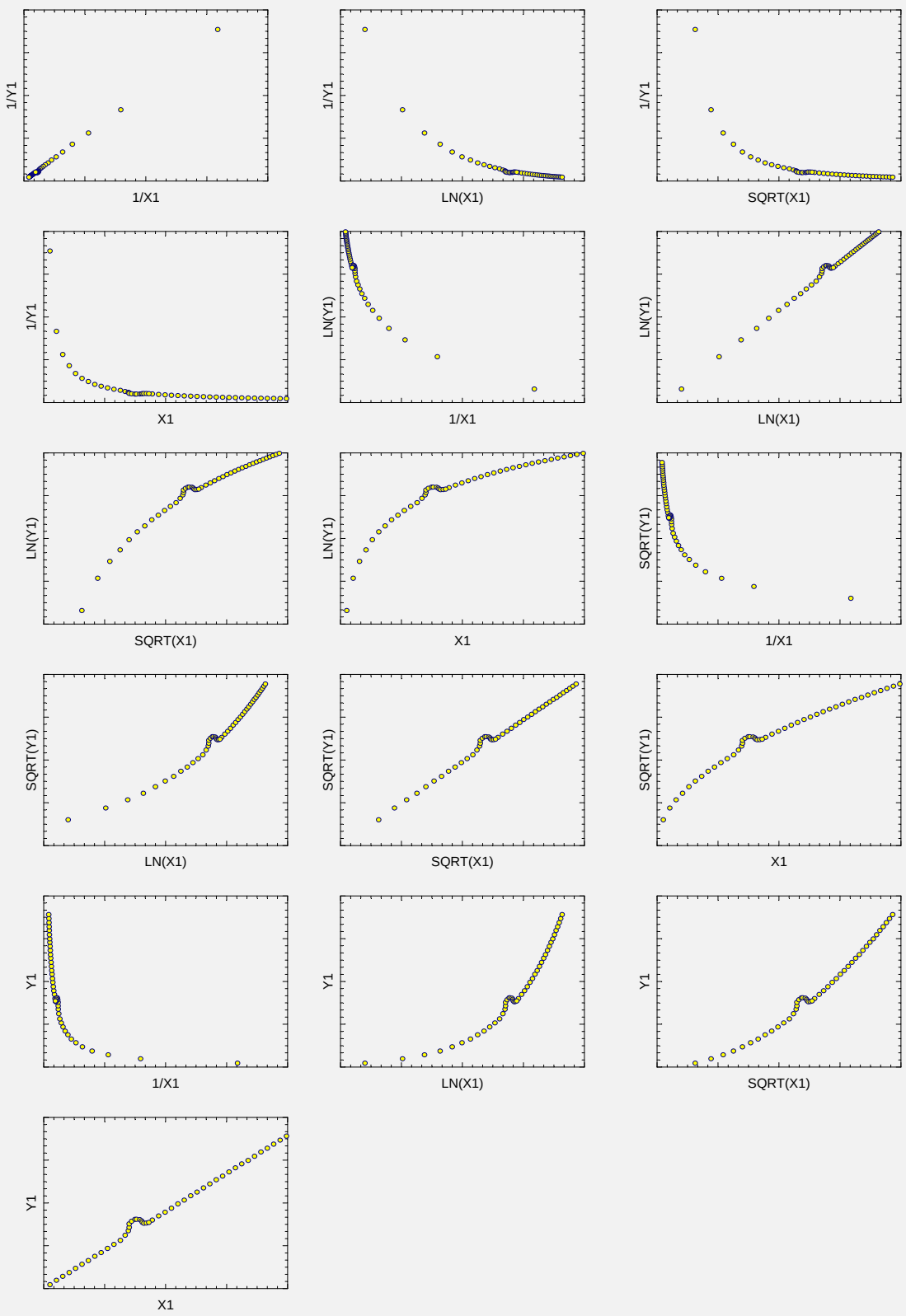
3 Specify the variables.

- On the Scatter Plot Matrix for Curve Fitting window, select the **Variables tab**.
- Double-click in the **Variable** box under **Y (Vertical) Variables**. This will bring up the variable selection window.
- Select **Y1** from the list of variables and then click **Ok**.
- Double-click in the **Variable** box under **X (Horizontal) Variables**. This will bring up the variable selection window.
- Select **X1** from the list of variables and then click **Ok**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Plots Section



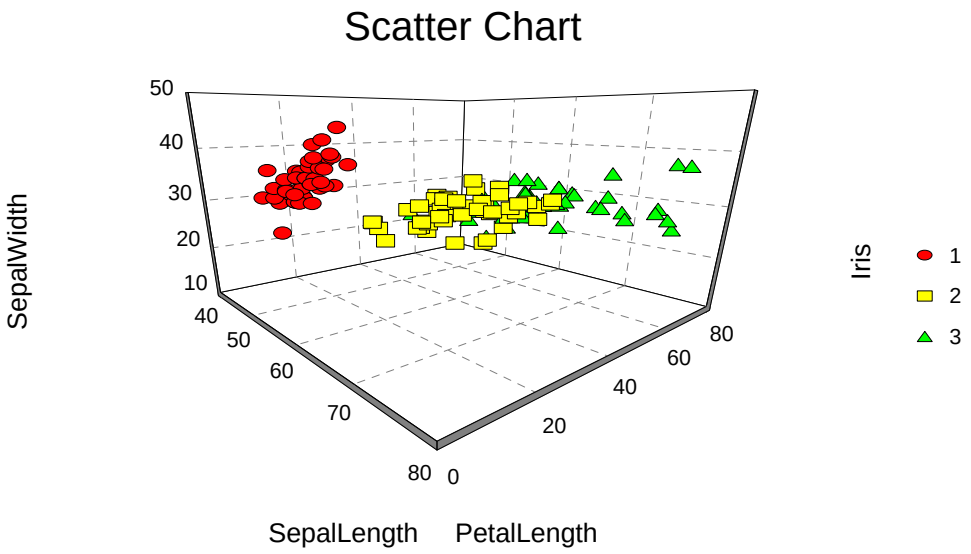
Chapter 170

3D Scatter Plots

Introduction

The 3D scatter plot displays trivariate points plotted in an X-Y-Z grid. It is particularly useful for investigating the relationships among these variables. The influence of a discrete variable may be investigated by using a different plotting symbol for each value of this variable. Hence, up to four variables (three numeric and one discrete) may be displayed on a single graph.

This procedure has the ability to rotate the data, giving the allusion of motion. This allows the human brain to better interpret the data.



Data Structure

The data are entered in three numeric variables. A fourth, discrete variable may be used to define the plotting symbol and color. Below are shown the first few rows of the 150-observation Fisher Iris dataset. These data are contained in the FISHER database.

FISHER dataset (subset)

Iris	SepalLen	SepalWid	PetalLen
1	50	33	14
3	64	28	56
2	65	28	46
3	67	31	56
3	63	28	51
1	46	34	14
3	69	31	51
2	62	22	45

Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Variables Tab

Specify the variables used to make the scatter plot.

Variables

Y (Vertical) Variable

Specify the variable displayed along the Y (vertical) axis. Only numeric values are used.

X1 Variable

Specify the variable displayed across the X1 (horizontal - left) axis. Only numeric values are used.

X2 Variable

Specify the variable displayed across the X2 (depth or horizontal - right) axis. Only numeric values are used.

Symbol Variable

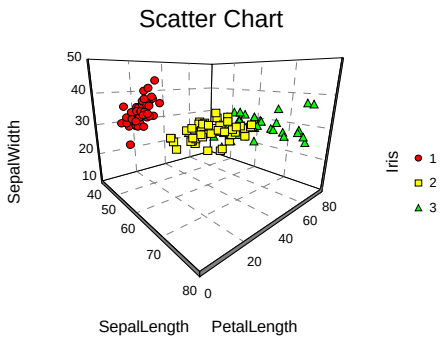
Specify an optional discrete variable whose values are used to defined the plotting color and symbol. If this variable is omitted, the first symbol and color are used.

Options

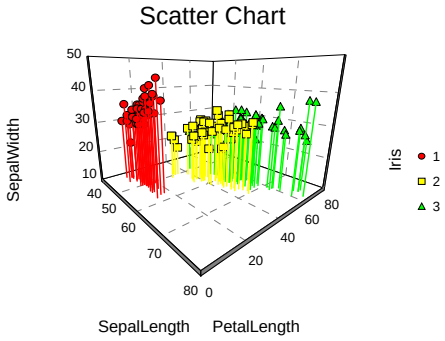
Show Sticks

This option controls the display of the vertical lines between the points and the plot floor.

Unchecked



Checked



Edit Chart Interactively

Checking this option will cause the interactive graphics editor to be displayed. This allows you to modify the graph interactively at run time. This editor is documented in its help file. Once you are through editing the plot, it will be displayed permanently in the output document.

Once the graphics editor comes up, you can use the scroll-bars on the four sides of the graph to interactively rotate the plot to the viewing position and angle that you like the best.

Axes Tab

These options control the appearance and position of the axes.

Y, X1, and X2 Axes

Scaling

This option specifies whether the axis is scaled automatically, with a zero axis origin, or from user specified maximum, minimum, and number of tick marks.

Minimum

Sets the value of the axis minimum. This value must be smaller than the smallest data value along this axis. This value is only used when Scaling is set to User Defined.

Maximum

Sets the value of the axis maximum. This value must be greater than the largest data value. This value is only used when Scaling is set to User Defined.

Number of Ticks

Sets the number of tickmarks. This value is only used when Scaling is set to User Defined.

Grid

Show Vertical Grid

Specify whether to display the vertical grid lines.

Show Horizontal Grid

Specify whether to display the horizontal grid lines.

Line Style

Specify the style (line, dots, dashes, etc.) of the grid lines.

Color

Set the color of the grid lines.

Cage

Thin Walls

This option specifies whether the walls of the axis grid that forms the background of the chart are thick or thin.

Edge Color (Thin Walls Unchecked)

Set the color of the cage edge in 3D charts.

Wall Color

Set the color of the cage wall in 3D charts.

Cage Flip

This option controls whether the back and side walls of the graph cage are allowed to switch to the opposite edge for better viewing.

Titles & Background Tab

These options control the titles that may be placed on all four sides of the chart.

Titles

Chart (Top) Title

This option gives the text that will appear in the title at the top of the chart. The color, font size, and style of the text is controlled by the options to the right.

Bottom Title

This option gives the text that will appear in the title at the bottom of the chart. The color, font size, and style of the text is controlled by the options to the right.

Left Title

This option gives the text that will appear in the title at the left of the chart. This may also serve as a label. The color and orientation of the text is controlled by the options to the right.

Right Title

This option gives the text that will appear in the title at the right of the chart. This may also serve as a label. The color and orientation of the text is controlled by the options to the right.

Variable Names (May be used in Titles of 3D Charts)

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Background Colors and Styles

Entire Graph

Specify the background color and style of the entire chart.

Inside Graph

Specify the background color and style of the chart itself (within the axes).

Graph Title

Specify the background color and style of the chart title.

Left Title

Specify the background color and style of the left title.

Right Title

Specify the background color and style of the right title.

Bottom Title

Specify the background color and style of the bottom title.

Tick Labels & Legend Tab

This panel controls the color and size of the tick labels.

Tick Labels

Text Color

This option sets the color of the reference items along the two axes.

Font Size

This option sets the font size of the reference items along the two axes.

Text Rotation

This option sets the display angle of the reference items along the horizontal axis.

Bold and Italics

This option sets the style of the reference items along the two axes.

Legend

Position

This option sets the position of the legend around the chart. Note that if you choose a position in which the full text of the legend cannot be fit, the legend will not be displayed.

Percent of Vertical Space

Specify the size of the legend as a percentage of the maximum possible. This option lets you shrink a legend that is too large.

Text Color

Specify the color of the legend text.

Background

Specify the background color of the legend.

Font Size

Specify the size of the legend text.

Bold and Italics

Specify the style of the legend text.

Legend Background Style

Specify the background style of the legend.

Color as Labels

Normally, text in the legend is displayed using the color selected by the Text Color option. This option indicates that each legend entry is to be displayed in the corresponding group color.

3D Options Tab

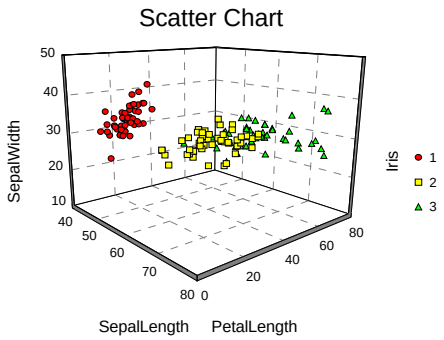
These options control the viewing orientation of the plot. These options may be set interactively by checking the Edit Chart Interactively option and then activating the four scroll bars on the sides of the Graphics Editor window.

Whole Chart Options

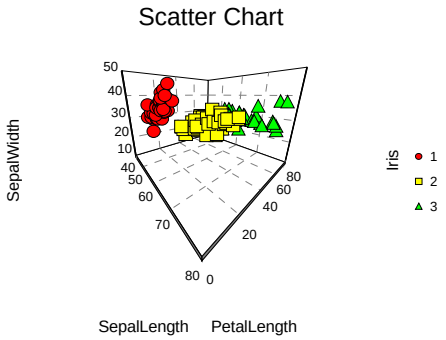
Perspective

This option specifies the perceived distance from which the graph is viewed. The range is from 0 to 100. As the value gets large, the distance gets smaller. A setting of 50 sets the viewing distance at about twice the graph’s width. A setting of 100 sets the viewing distance at about equal to the graph’s width.

Perspective = 10



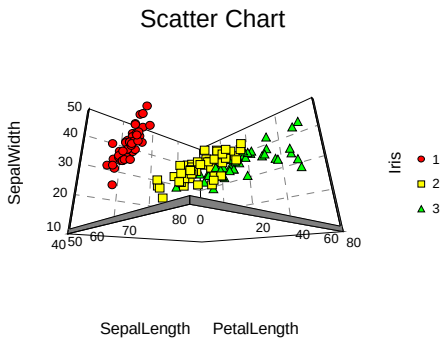
Perspective = 90



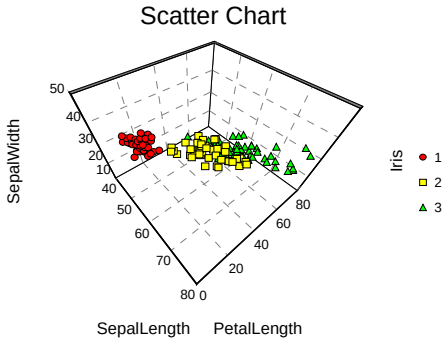
Elevation

This option sets the vertical viewing angle (in degrees). The setting represents an angle above or below a point halfway up the graph. The range is from -60 to 90.

Elevation = -20



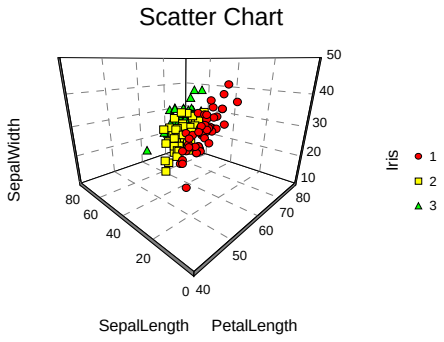
Elevation = 50



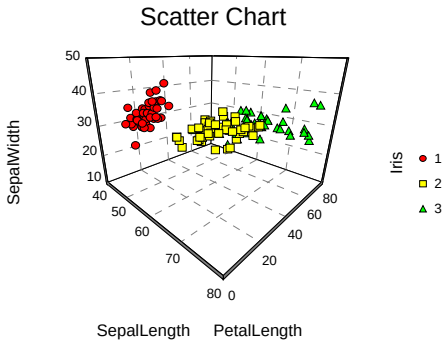
Rotation

This option sets the horizontal viewing angle (in degrees). The setting represents an angle around the base of the graph. The range is from -180 to 180.

Rotation = -45



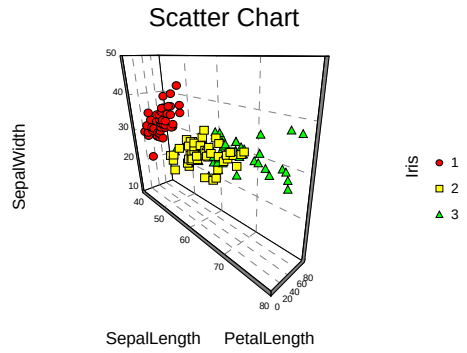
Rotation = 45



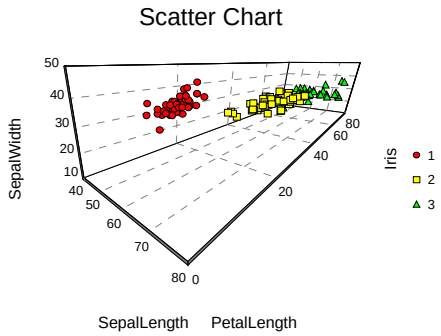
Depth

This option sets the width in the X2 direction. The range is from 1 to 20,000.

Depth = 20



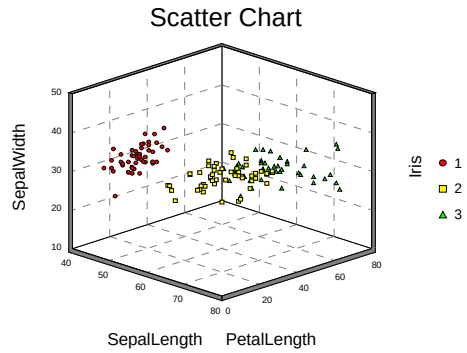
Depth = 400



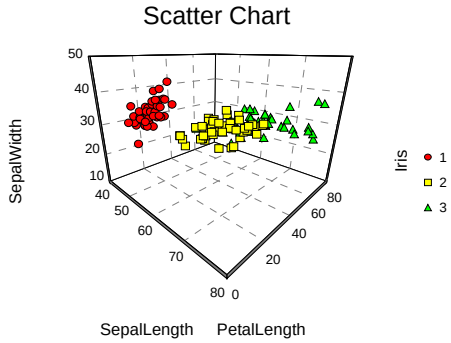
Projection Method

This option specifies the method used to determine viewer position and angle in a 3D graph.

Isometric



Perspective



Symbols Tab

These options specify the colors and patterns used for the plotting symbols. A different plotting symbol is displayed for each unique value of the Symbol Variable. If no Symbol Variable is specified, the settings of Symbol 1 are used for the plot symbol.

Symbol Colors and Styles

Symbol 1 - 15

These options let you specify the color and type of each symbol. The first symbol is associated with the first value, the second symbol with the second value, and so on. If more than fifteen symbols are needed, they are reused so that symbol 16 = symbol 1, symbol 17 = symbol 2, and so on.

Symbol Size

Symbol Size

The radius (size) of the plotting symbol.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a 3D Scatter Plot

This section presents an example of how to create a 3D scatter plot of the data stored on the FISHER database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the 3D Scatter Plots window.

1 Open the Fisher dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the 3D Scatter Plots window.

- On the menus, select **Graphics**, then **3D Scatter Plots**. The 3D Scatter Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

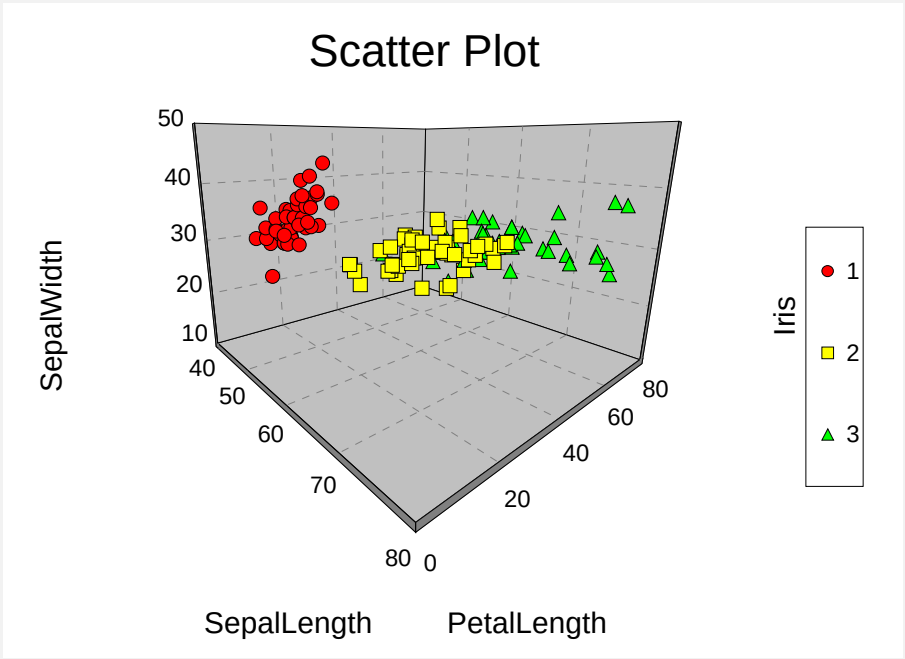
- On the 3D Scatter Plots window, select the **Variables tab**.
- Double-click in the **Y (Vertical) Variable** text box. This will bring up the variable selection window.

- Select **SepalWidth** from the list of variables and then click **Ok**. “SepalWidth” will appear in the Y Variable box.
- Double-click in the **X1 Variable** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. “SepalLength” will appear in the X1 Variable box.
- Double-click in the **X2 Variable** text box. This will bring up the variable selection window.
- Select **PetalLength** from the list of variables and then click **Ok**. “PetalLength” will appear in the X2 Variable box.
- Double-click in the **Symbol Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Symbol Variable box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

3D Scatter Plot Output



Example 2 – Interactive Rotation

This section presents an example of real-time rotation of a 3D scatter plot of the data stored on the FISHER database.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the 3D Scatter Plots window.

1 Open the Fisher dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the 3D Scatter Plots window.

- On the menus, select **Graphics**, then **3D Scatter Plots**. The 3D Scatter Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the 3D Scatter Plots window, select the **Variables tab**.
- Double-click in the **Y (Vertical) Variable** text box. This will bring up the variable selection window.
- Select **SepalWidth** from the list of variables and then click **Ok**. “SepalWidth” will appear in the Y Variable box.
- Double-click in the **X1 Variable** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. “SepalLength” will appear in the X1 Variable box.
- Double-click in the **X2 Variable** text box. This will bring up the variable selection window.
- Select **PetalLength** from the list of variables and then click **Ok**. “PetalLength” will appear in the X2 Variable box.
- Double-click in the **Symbol Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Symbol Variable box.
- Check the **Edit Chart Interactively** box.

4 Specify the axes.

- On the 3D Scatter Plots window, select the **Axes tab**.
- Click the **Show Vertical Grid** option so that it is not checked.
- Click the **Show Horizontal Grid** option so that it is not checked.
- Check the **Thin Walls** option.
- Set the **Cage Wall Color** to white by clicking the button and checking the white color button on the Color of Cage Wall Color window.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top). This will cause the 3D Scatter Plot Editing window to appear.
- The scroll bar on the left controls the vertical view angle. Try moving the thumb of this scroll bar up and down. The scatter plot will have the appearance of motion.
- The scroll bar on the bottom controls the horizontal view angle. Try moving the thumb of this scroll bar right and left. The scatter plot will have the appearance of motion.
- The other two scroll bars control the depth and perspective of the plot.
- The tool bar at the top allows the setting of almost every detail of the plot.

Chapter 171

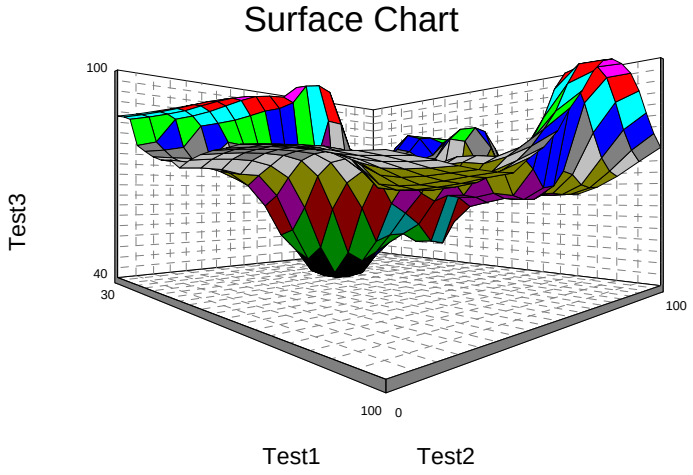
3D Surface Plots

Introduction

Surface plots are diagrams of three-dimensional data. Rather than showing the individual data points, surface plots show a functional relationship between a designated dependent variable (Y), and two independent variables (X1 and X2). The plot is a companion plot to the contour plot.

It is important to understand how these plots are constructed. A two-dimensional grid of X1 and X2 is constructed. The range of this grid is equal to the range of the data. Next, a Y value is calculated for each grid point. This Y value is a weighted average of all data values that are "near" this grid point. (The number of points averaged is user specified.) The three-dimensional surface is constructed using these averaged values. Hence, the surface plot does not show the variation at each grid point.

These plots are useful in regression analysis for viewing the relationship among a dependent and two independent variables. Remember that multiple regression assumes that this surface is a perfectly flat surface. Hence, the surface plot lets you visually determine if multiple regression is appropriate.



Data Structure

A surface plot is constructed from three variables. The X1 and X2 (independent) variables are shown on the horizontal axes. The Y variable is shown along the vertical axis. Note that all three variables must be numeric.

Procedure Options

This section describes the options available in this procedure. To find out more about using a procedure, turn to the Procedures chapter.

Variables Tab

Specify the variables used to make the plot.

Variables

Y (Vertical) Variable

Specify the numeric variable displayed along the Y (vertical) axis.

X1 Variable

Specify the numeric variable displayed across the X1 (horizontal - left) axis.

X2 Variable

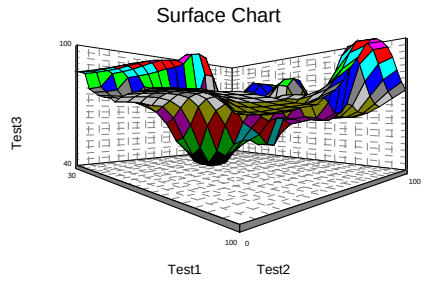
Specify the numeric variable displayed across the X2 (depth or horizontal - right) axis.

Options

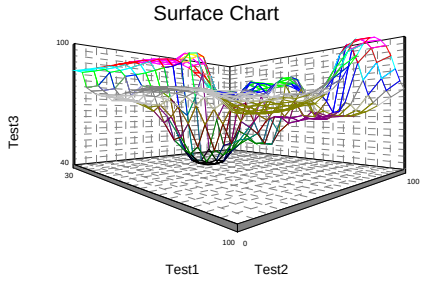
Surface Chart Style

This option selects the style of the surface plot.

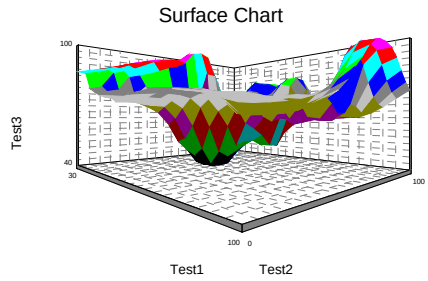
Painted, Lines



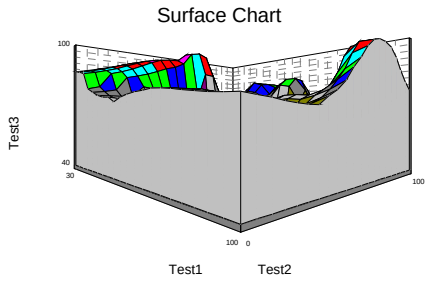
Lines



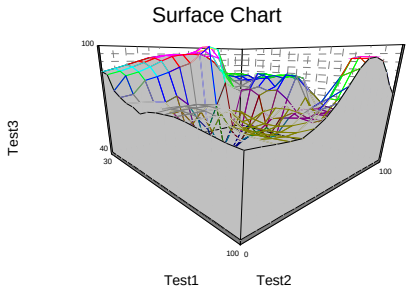
Painted



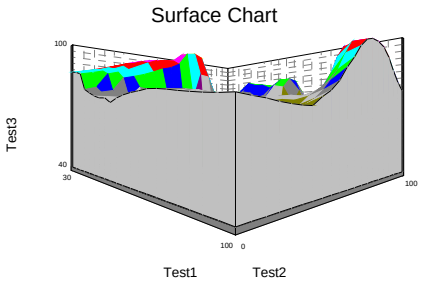
Painted, Lines, Wall



Lines, Wall



Painted, Wall



Rows Per Grid Point

The number of nearest-neighbor data points that are averaged to calculate a Y value at each grid point. This option helps control the smoothness of the plot.

Smoothing Coefficient

This is the exponent that is applied to the relative distance of each data point from the grid point when computing the grid value. The grid value (Y value) is computed as a weighted average of the nearest-neighbor data points. The weights are inversely proportional to this exponentiated distance. Values of 1, 2, or 3 will give you quite a range of alternatives. When the value is 1, relatively distant data points have a larger influence on the average. Hence the contour plot tends to be flatter. Alternative, a value of 3 will give a plot with more hills and valleys.

Edit Chart Interactively

Checking this option will cause the interactive graphics editor to be displayed. This allows you to modify the graph interactively at run time. This editor is documented in its help file. Once you are through editing the plot, it will be displayed permanently in the output document.

Once the graphics editor comes up, you can use the scroll-bars on the four sides of the graph to interactively position the plot to any viewing position and angle.

Axes Tab

These options control the appearance and position of the axes.

Y, X1, and X2 Axes

Decimals

This option specifies the number of decimal places shown in the reference numbers along the corresponding axis.

Minimum

Sets the value of the axis minimum. This value must be smaller than the smallest data value along this axis.

Maximum

Sets the value of the axis maximum. This value must be greater than the largest data value.

Number of Intervals

Sets the number of intervals along this axis. The number of tickmarks is one more than this amount.

Grid

Show Vertical Grid

Specify whether to display the vertical grid lines.

Show Horizontal Grid

Specify whether to display the horizontal grid lines.

Line Style

Specify the style (line, dots, dashes, etc.) of the grid lines.

Color

Set the color of the grid lines.

Cage

Thin Walls

This option specifies whether the walls of the axis grid that forms the background of the chart are thick or thin.

Edge Color (Thin Walls Unchecked)

Set the color of the cage edge in 3D charts.

Wall Color

Set the color of the cage wall in 3D charts.

Axis Color

Set the color of the cage axes in 3D charts.

Cage Flip

This option controls whether the back and side walls of the graph cage are allowed to switch to the opposite edge for better viewing.

Titles & Background Tab

These options control the titles that may be placed on all four sides of the chart.

Titles

Chart (Top) Title

This option gives the text that will appear in the title at the top of the chart. The color, font size, and style of the text is controlled by the options to the right.

Bottom Title

This option gives the text that will appear in the title at the bottom of the chart. The color, font size, and style of the text is controlled by the options to the right.

Left Title

This option gives the text that will appear in the title at the left of the chart. This may also serve as a label. The color and orientation of the text is controlled by the options to the right.

Right Title

This option gives the text that will appear in the title at the right of the chart. This may also serve as a label. The color and orientation of the text is controlled by the options to the right.

Variable Names (May be used in Titles of 3D Charts)

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Background Colors and Styles

Entire Graph

Specify the background color and style of the entire chart.

Inside Graph

Specify the background color and style of the chart itself (within the axes).

Graph Title

Specify the background color and style of the chart title.

Left Title

Specify the background color and style of the left title.

Right Title

Specify the background color and style of the right title.

Bottom Title

Specify the background color and style of the bottom title.

Tick Labels Tab

This panel controls the color and size of the tick labels.

Tick Labels

Text Color

This option sets the color of the reference items along the two axes.

Font Size

This option sets the font size of the reference items along the two axes.

Text Rotation

This option sets the display angle of the reference items along the horizontal axis.

Bold and Italics

This option sets the style of the reference items along the two axes.

3D Options Tab

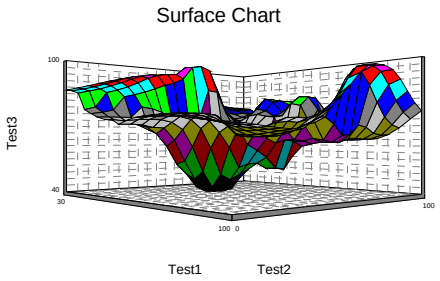
These options control the viewing position of the plot. These options may be set interactively by check the Edit Chart Interactively and then activating the four scroll bars on the sides of the Graphics Editor window.

Whole Chart Options

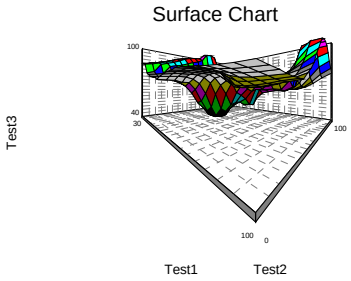
Perspective

This option specifies the perceived distance from which the graph is viewed. The range is from 0 to 100. As the value gets large, the distance gets smaller. A setting of 50 sets the viewing distance at about twice the graph's width. A setting of 100 sets the viewing distance at about equal to the graph's width.

Perspective = 10



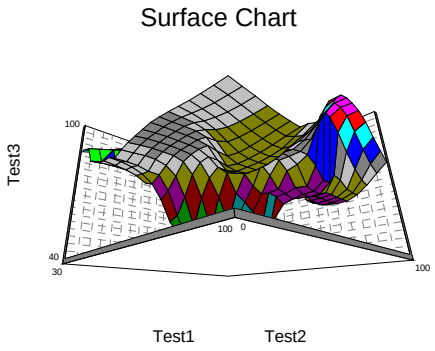
Perspective = 90



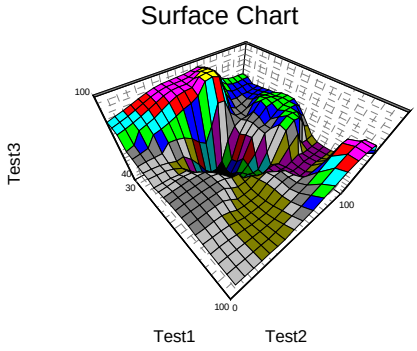
Elevation

This option sets the vertical viewing angle (in degrees). The setting represents an angle above or below a point halfway up the graph. The range is from -60 to 90.

Elevation = -20



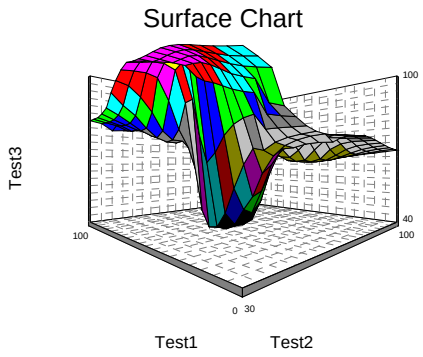
Elevation = 50



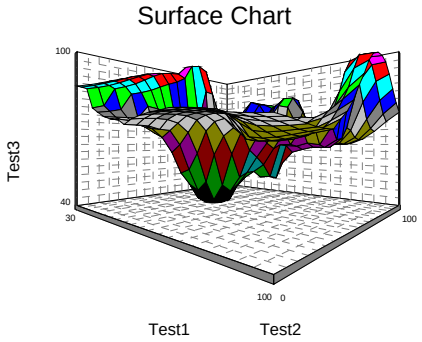
Rotation

This option sets the horizontal viewing angle (in degrees). The setting represents an angle around the base of the graph. The range is from -180 to 180.

Rotation = -45



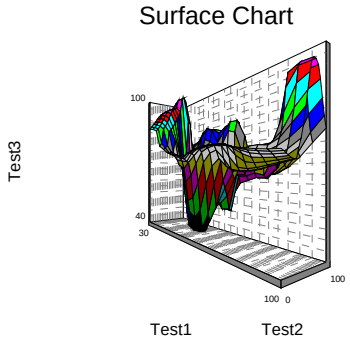
Rotation = 45



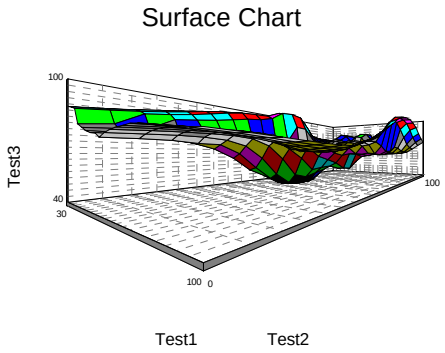
Depth

This option sets the width in the X2 direction. The range is from 1 to 20,000.

Depth = 20



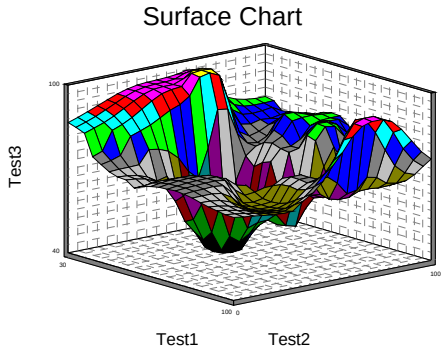
Depth = 400



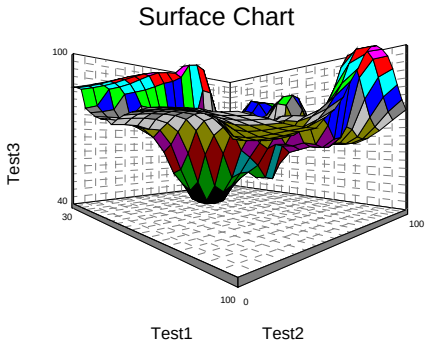
Projection Method

This option specifies the method used to determine viewer position and angle in a 3D graph.

Isometric



Perspective



Surface Chart Type

Surface Color Palette

Specify a color palette for the surface chart. The 128-color palettes may only be used on machines with SuperVGA capabilities. Using a setting here of, for example, Black to Red will allow the surface plot to show a continuous array of red hues from lowest to highest.

Surface Color Min

Specifies the number of the color to be associated with the lowest numerical value. Possible values are 32 to 127. A value near 50 usually works well. Note that this option only works with 128-color palettes.

Surface Color Max

Specifies the number of the color to be associated with the highest numerical value. Possible values are 32 to 127. A value near 120 usually works well. Note that this option only works with 128-color palettes.

Surface Wall Color

The color of the surface wall. The surface wall is a plane extending from the edge of the surface to each axis.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a 3D Surface Plot

This section presents an example of how to create a surface chart using data stored on the SAMPLE database. We will generate the surface using Test3 as the dependent variable and Test1 and Test2 as the independent variables.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the 3D Surface Plots window.

1 Open the Sample dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.

2 Open the 3D Surface Plots window.

- On the menus, select **Graphics**, then **Other Charts and Plots**, then **3D Surface Plots**. The 3D Surface Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the 3D Surface Plots window, select the **Variables tab**.
- Double-click in the **Y (Vertical) Variable** text box. This will bring up the variable selection window.
- Select **Test3** from the list of variables and then click **Ok**. "Test3" will appear in the Y Variable box.
- Double-click in the **X1 Variable** text box. This will bring up the variable selection window.
- Select **Test1** from the list of variables and then click **Ok**. "Test1" will appear in the X1 Variable box.
- Double-click in the **X2 Variable** text box. This will bring up the variable selection window.
- Select **Test2** from the list of variables and then click **Ok**. "Test2" will appear in the X2 Variable box.

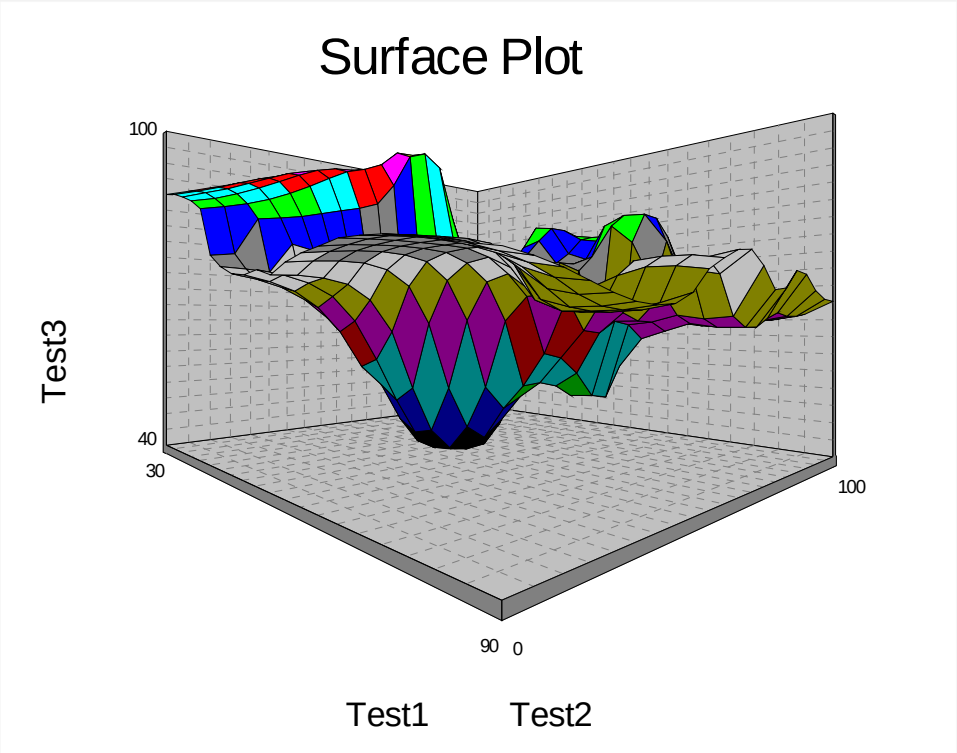
4 Specify the number of decimal places.

- On the 3D Surface Plots window, select the **Axes tab**.
- Set the **Y Decimals** to **0**.
- Set the **X1 Decimals** to **0**.
- Set the **X2 Decimals** to **0**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Surface Plot Output

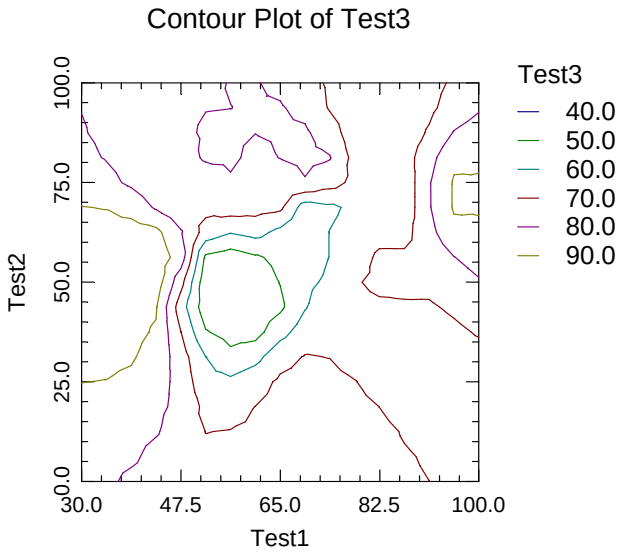


Chapter 172

Contour Plots

Introduction

Contour plots are topographical maps drawn from three-dimensional data. One variable is represented on the horizontal axis and a second variable is represented on the vertical axis. The third variable is represented by isolines (lines of constant value). These plots are often useful in data analysis, especially when you are searching for minimums and maximums in a set of trivariate data. An introduction to contour techniques is contained in Milne (1987).



The program allows input of equally or irregularly spaced data. You can specify the number of grid points, the type of line used to show the isoline (dots, dashes, etc.), and the colors of each of the lines.

Data Structure

A contour plot is constructed from three variables. The X and Y variables are shown on the horizontal and vertical axes, respectively. The Z variable is partitioned and its values are represented by a colored line. Note that all three variables must be numeric.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the contour plot.

Variables

Variable (X, Y, Z)

Each box specifies a numeric variable that will be used to construct the plot. The X variable is displayed along the horizontal axis, the Y variable is displayed along the vertical axis, and the Z variable is displayed using the contour layers.

Minimum and Maximum (X, Y, Z)

The minimum and maximum values along each axis. Care should be used in setting these as they can distort the plot.

Number of Slices (X, Y)

The number of divisions (grid points) along each axis.

Number of Z Slices

The number of contour slices along the Z axis. It is only used if the Contour (Z) Values box is left blank.

Note: In some cases the legend will show one more contour line than this number, corresponding to a minimum value that is not in the data range.

Plot Settings

Obs / Grid Point

The number of nearest-neighbor data points that are averaged to calculate a Z value at each grid point. This option partially controls the smoothness of the plot.

Smooth Coefficient

This is the exponent that is applied to the relative distance of each data point from the grid point when computing the grid value. The grid value (Z value) is computed as a weighted average of the nearest-neighbor data points. The weights are inversely proportional to this exponentiated distance. Values of 1, 2, or 3 will give you quite a range of alternatives. When the value is 1, relatively distant data points have a larger influence on the average. Hence the contour plot tends to be flatter. Alternative, a value of 3 will give a plot with more hills and valleys.

Show Raw X-Y Coordinate Data Using Symbol

The X-Y data may be superimposed over the contour plot. This option controls whether the data are displayed and the appearance (color, size, and symbol) of those data.

Plot Settings – Contour Details

Contour Values (Overrides Number of Slices for Z)

An optional list of values (separated with commas) at which contour lines are drawn. If this option is left blank, the contour values are determined by the number of slices selected (see Slices - Z above).

Contour Value Font Size

When the contour values are displayed, this is the size of the font used to display those values.

Show Contour Values on Plot

This option specifies whether to overlay the values of the contour lines on the plot.

Contour Lines

These options set the color, width, and pattern of the up to fifteen contour lines. Double-clicking the line, or clicking the button to the right of the symbol, brings up a line specification window. This window lets you specify the characteristics of a line in detail. These options are especially useful for making contour maps that may be displayed on black and white printers.

- **Color**
  The color of the line.
- **Width**
  The width of the line.
- **Line Pattern**
  The line pattern (solid, dot, dash, etc.).

Y and X Axes Tab

These options specify the characteristics of the vertical and horizontal axis.

Vertical and Horizontal Axis

Label Text

This box supplies the axis label. The characters {X}, {Y}, and {G} are replaced by the horizontal, vertical, and grouping variable names, respectively. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line and its type (numeric or text).

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**
  Specifies the color of the tick labels.
- **Font Size**
  Specifies the size of the tick labels.
- **Bold, Italic, Underline**
  Specifies the font style of the tick labels.
- **Decimals**
  Specifies the number of decimal places displayed in the tick labels.
- **Max Characters**
  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).
- **Text Rotation**
  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical and Horizontal Axis – Positions

Axis

This option controls the position of the axis: if and where it is displayed.

Label

This option controls the position of the label: if and where it is displayed.

Tick Labels

This option controls the position of the tick labels: if and where they are displayed.

Tickmarks

This option controls the position of the tickmarks: if and where they are displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. The characters {A} and {B} are replaced by the numeric values of the intercept and slope of the regression line, respectively. To display the fitted regression equation, you could use {Y} = {A} + ({B}){X}.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Legend

When data for more than one group are displayed, a legend is desirable. These options specify the legend.

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. The characters *{G}* will be replaced by the name of the group variable. Click the button on the right to specify the font size, color, and style of the legend text.

Legend Decimals

Specify the decimal places displayed in the reference numbers. NOTE: this value may be overridden by the variable's format (set on the Variable Info Sheet).

Storage Tab

These options let you store the X, Y, and Z values of the grid points for further analysis. If a variable is selected, the values are automatically stored. Use care when selecting these variables, since any data currently in the variable will be replaced by the new values.

Storage Variables

Store X, Y, or Z Values in

If a variable is selected the corresponding grid values will be stored in that variable. Any existing data will be overwritten.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Contour Plot

This section presents an example of how to generate a contour plot. The data used are from the SAMPLE database. We will create contour plot of Test1, Test2, and Test3. Test3 will be contoured.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Contour Plots window.

1 Open the Sample dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.

2 Open the Contour Plots window.

- On the menus, select **Graphics**, then **Other Charts and Plots**, then **Contour Plots**. The Contour Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

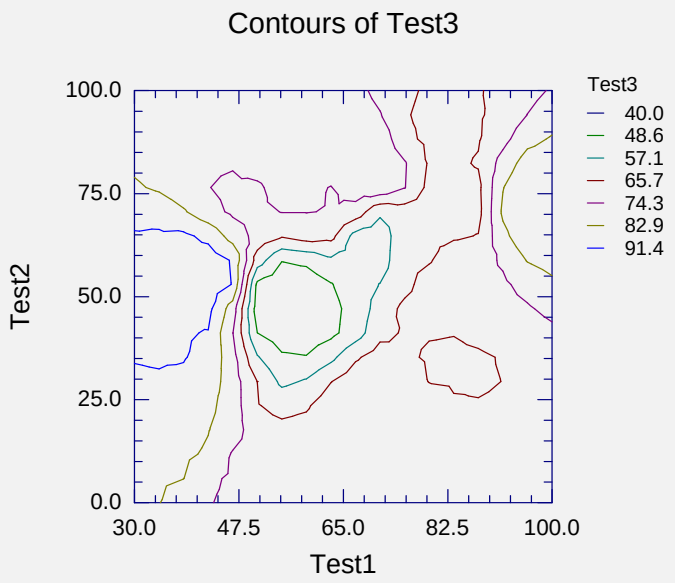
3 Specify the variables.

- On the Contour Plots window, select the **Variables tab**.
- Double-click in the **X Variable** text box. This will bring up the variable selection window.
- Select **Test1** from the list of variables and then click **Ok**. "Test1" will appear in the X Variable box.
- Double-click in the **Y Variable** text box. This will bring up the variable selection window.
- Select **Test2** from the list of variables and then click **Ok**. "Test2" will appear in the Y Variable box.
- Double-click in the **Z Variable** text box. This will bring up the variable selection window.
- Select **Test3** from the list of variables and then click **Ok**. "Test3" will appear in the Z Variable box.
- In the **Z Slices** box, select **6**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Contour Plot Output

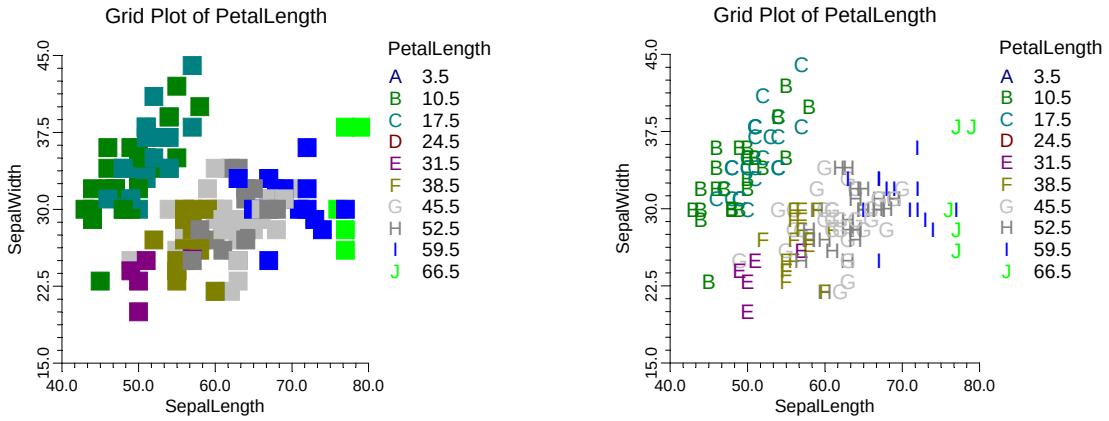


Chapter 173

Grid Plots

Introduction

The grid plot is a type of contour plot developed for displaying three variables. The first two variables are displayed as in the scatter plot on the vertical and horizontal axes. The third variable is displayed either by the color of the block or by a symbol that is coded from low to high.



Data Structure

A grid plot is constructed from three variables. The X and Y variables are shown on the horizontal and vertical axes, respectively. The Z variable is partitioned and its values are represented by either the plot symbol or the block color (see examples above). Note that all three variables must be numeric.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies which variables are used in the grid plot.

Variables

Variable (X, Y, Z)

Each box specifies a numeric variable that will be used to construct the plot. The X variable is displayed along the horizontal axis, the Y variable is displayed along the vertical axis, and the Z variable is displayed using the contour layers.

Minimum and Maximum (X, Y, Z)

The minimum and maximum values along each axis. Care should be used in setting these as they can distort the plot.

Number of Slices (X, Y, Z)

The number of divisions (grid points) along each axis. The number of slices along the z axis is the number of contour lines. It is only used if the Contour (Z) Values box is left blank.

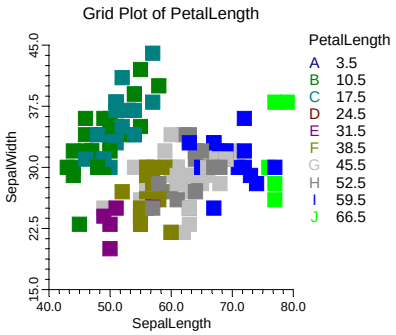
Plot Settings

Plot Style

Specifies the method used to display the value of the Z variable.

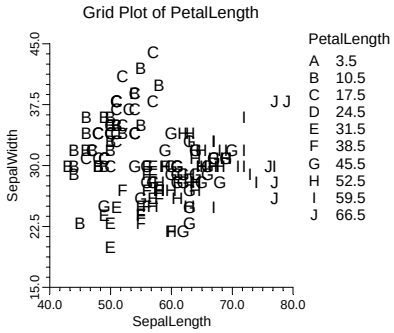
- **Blocks**

  The values of the Z variable are displayed as colored blocks.



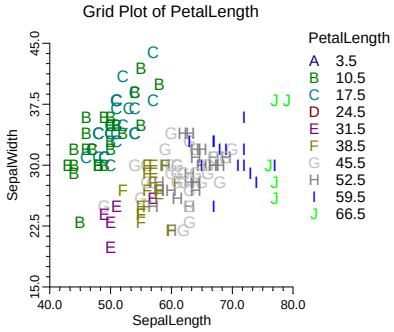
- **Symbols - One Color**

  The values of the Z variable are displayed as letters all with the same color. Each letter represents a range of values.



- **Symbols - Multiple Colors**

  The values of the Z variable are displayed as letters having different colors. Each letter represents a range of values.



Plot Symbols

These options specify the symbols that are used for each layer of the depth (Z) axis. Up to fifteen symbols may be specified.

Symbol 1 - 15

These options specify the symbols that are used for each layer of the depth (Z) axis. The symbols are used if the Plot Style option is set to Symbols. If the Plot Style option is set to Blocks, only the color of the symbol is used.

Click on the symbol or click on the button to the right of the symbol to bring up the symbol window. This will let you change the color, size, and type of symbol.

Axes Tab

These options specify the characteristics of the vertical and horizontal axis.

Vertical and Horizontal Axis

Label Text

This box supplies the axis label. The characters {X}, {Y}, and {G} are replaced by the horizontal, vertical, and grouping variable names, respectively. The font size, color, and style of the label may be modified by pressing the button on the right of the text.

Axis

Clicking this box (or the button to the right) brings up the settings window that controls the size and color of the axis line and its type (numeric or text).

Vertical and Horizontal Axis – Tickmarks and Grid Lines

Major Ticks (number)

Tick labels are displayed for the major tickmarks. This option specifies the number of major tickmarks displayed along the axis.

Major Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the major tickmarks.

Major Grid Lines

Checking this option causes the major grid lines to be displayed.

Minor Ticks (number)

This option specifies the number of minor tickmarks displayed along the axis.

Minor Ticks (settings)

This option sets the color, line width, and line pattern of the grid lines. It also sets the width and length of the minor tickmarks.

Minor Grid Lines

Checking this option causes the minor grid lines to be displayed.

Tick Label Settings...

Clicking this button brings up a window that controls the tick labels that are displayed along this axis. The following options are available in this window:

- **Color**
  Specifies the color of the tick labels.

- **Font Size**
  Specifies the size of the tick labels.

- **Bold, Italic, Underline**

  Specifies the font style of the tick labels.

- **Decimals**

  Specifies the number of decimal places displayed in the tick labels.

- **Max Characters**

  The maximum length (number of characters allowed) of a tick label. This field shifts the axis label away from the axis to make room for the tick labels. Hence, if your tick labels are large, such as 1234.456, you would want a large value here (such as 10 or even 15).

- **Text Rotation**

  Specifies whether the tick labels are displayed vertically or horizontally.

Vertical and Horizontal Axis – Positions

Axis

This option controls the position of the axis: if and where it is displayed.

Label

This option controls the position of the label: if and where it is displayed.

Tick Labels

This option controls the position of the tick labels: if and where they are displayed.

Tickmarks

This option controls the position of the tickmarks: if and where they are displayed.

Titles and Miscellaneous Tab

These options set the titles of the plot. Up to two titles may be specified at the top and at the bottom of the plot.

Titles

Top Title Line 1 and 2

Two title lines may be placed at the top of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. The characters {A} and {B} are replaced by the numeric values of the intercept and slope of the regression line, respectively. To display the fitted regression equation, you could use {Y} = {A} + ({B}){X}.

Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Bottom Title Line 1 and 2

Two title lines may be placed at the bottom of the plot. This option controls value and appearance of these titles. In the text, the characters *{X}*, *{Y}*, *{Z}*, and *{G}* are replaced by the names of the corresponding variables. Clicking the button on the right of the text box brings up a window that sets the color, size, and style of the text.

Background Colors

These options specify plot interior and background colors.

Background

The background color of the plot.

Interior

The color of the area of the plot inside the axes.

Format Options

Variable Names

This option selects whether to display only variable's name, label, or both.

Value Labels

This option selects whether to display only values, value labels, or both. Use this option if you want the group variable to automatically attach labels to the values (like 1=Yes, 2=No, etc.).

Legend

When data for more than one group are displayed, a legend is desirable. These options specify the legend.

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. The characters *{G}* will be replaced by the name of the group variable. Click the button on the right to specify the font size, color, and style of the legend text.

Legend Decimals

Specify the decimal places displayed in the reference numbers. NOTE: this value may be overridden by the variable's format (set on the Variable Info Sheet).

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Creating a Grid Plot

This section presents an example of how to generate a contour plot. The data used are from the FISHER database. We will a create grid plot of *SepalLength*, *SepalWidth*, and *PetalLength*. *PetalLength* will be separated into grid levels.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Grid Plots window.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Fisher.s0**.
- Click **Open**.

2 Open the Grid Plots window.

- On the menus, select **Graphics**, then **Other Charts and Plots**, then **Grid Plots**. The Grid Plots procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

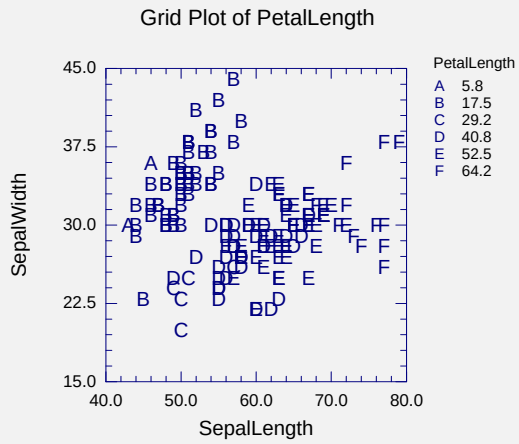
- On the Grid Plots window, select the **Variables tab**.
- Double-click in the **X Variable** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the X Variable box.
- Double-click in the **Y Variable** text box. This will bring up the variable selection window.

- Select **SepalWidth** from the list of variables and then click **Ok**. “SepalWidth” will appear in the Y Variable box.
- Double-click in the **Z Variable** text box. This will bring up the variable selection window.
- Select **PetalLength** from the list of variables and then click **Ok**. “PetalLength” will appear in the Z Variable box.
- In the **Z Number of Slices** box, select **6**.
- In the **Plot Style** list box, select **Symbols - One Color**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Grid Plot Output



Creating a Grid Plot Style File

Some of the statistical procedures include grid plots as part of their reports. Since the grid plot has almost 200 options, adding it to another procedure’s report greatly increases the number of options that you have to specify for that procedure. To overcome this, we let you create and save grid plot style files. These files contain the current settings of all grid plot options. When you use the style file in another procedure you only have to set a few of the options. Most of the options come from this style file. A default grid plot style file was installed with the **NCSS** system. Other style files may be added.

We will now take you through the steps necessary to create a grid plot style file.

1 Open the FISHER dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file Fisher.s0.
- Click **Open**.
- Note: You do not necessarily have to use the FISHER database. You can use whatever database is easiest for you. Just open a database with a column of numeric data.

2 Open the Grid Plots window.

- On the menus, select **Graphics**, then **Other Charts and Plots**, then **Grid Plots**. The Grid Plots procedure will be displayed.

3 Specify the variables.

- On the Grid Plots window, select the **Variables tab**.
- Double-click in the **X Variable** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the X Variable box.
- Double-click in the **Y Variable** text box. This will bring up the variable selection window.
- Select **SepalWidth** from the list of variables and then click **Ok**. "SepalWidth" will appear in the Y Variable box.
- Double-click in the **Z Variable** text box. This will bring up the variable selection window.
- Select **PetalLength** from the list of variables and then click **Ok**. "PetalLength" will appear in the Z Variable box.

4 Set your options.

- Set the various options of the grid plot's appearance to the way you want them.
- Run the procedure to generate the grid plot. This gives you a final check on whether it appears just how you want it. If it does not appear quite right, go back to the panel and modify the settings until it does.

5 Save the template (optional).

- Although this step is optional, it will usually save a lot of time and effort later if you store the current template. Remember, the template file is not the style file.
- To store the template, select the **Template tab** on the Grid Plot window.
- Enter an appropriate name in the File Name box.
- Enter an appropriate phrase at the bottom of the window in the Template Id (the long box across the bottom of the Grid Plot's window). This phrase will be displayed in the Template Id's box to help you identify the template files.
- Select Save Template from the File menu. This will save the template.

6 Create and Save the Style File

- Select Save Style File from the File menu. The Save Style File Window will appear.
- Enter an appropriate name in the Selected File box. You can either reuse one of the style files that already exist or create a new name. You don't have to worry about drives, directory names, or file extensions. These are all added by the program. Just enter an appropriate file name.
- Press the **Ok** button. This will create and save the style file.

7 Using a Style File

- Using the style file is easy. For example, suppose you want to use this Grid Plot Style file in the Response Surface Analysis procedure. You do the following:
- Select the **Grid Plot** tab in the Response Surface Analysis procedure.
- Click the button to the right of the Plot Style File box (the initial file name is Default). This will bring up the Grid Plot Style File Selection window.

- Click on the appropriate file so that it is listed in the Selected File box. Click the **Ok** button.
- The new style file name will appear in the Plot Style File box of the Response Surface Analysis window. Your new style has been activated.

Chapter 180

Color Selection Window

Introduction

The Color Selection window lets you choose an appropriate color from the 16 million colors that are available on today's monitors using the RGB color space. Although choosing a color sounds like a trivial task, it can become time-consuming and frustrating. When you have invested a lot of time and money in a project and now have important results to communicate, you probably want to take the time to make outstanding graphics. A few, well-chosen charts can communicate results quickly and effectively. An important feature of a chart is the color scheme that you use. The goal of the color selection window is to provide a tool that will allow you to pick a set of colors that are pleasing to the eye when viewed together, and let the viewer interpret the results quickly and effectively.

Color Theory

Picking colors is much easier once you discover the simple mathematical rules that should be followed. These rules are based on the theory of color. We suggest you browse the online article on color theory provided at **www.worqx.com**. This site provides a great introduction to color theory and color selection.

Color Models

In the online article, various color models are discussed. Computer users are familiar with the RGB (red, green, blue) model, which gives all the colors that are available on a computer monitor. Printers are more familiar with the CYM or CYMK model. Many other models have been developed. Our favorite model for choosing colors is the HSB (hue, saturation, brightness) model. When using this model, first pick the basic color or hue. Next, select the saturation and brightness of the color. Once a color has been selected, matching colors can be found by selecting other hues while keeping the saturation and brightness the same.

Color Wheel

The task of choosing several matching colors is aided by the use of a color wheel. A color wheel shows several hues all with the same saturation and brightness. Hence, all of the colors on the color wheel 'match'. The colors are arranged on the wheel so that colors on opposite sides of the wheel have high contrast, while adjacent colors have low contrast. Usually, high-contrasting colors are desirable when representing different treatment levels.

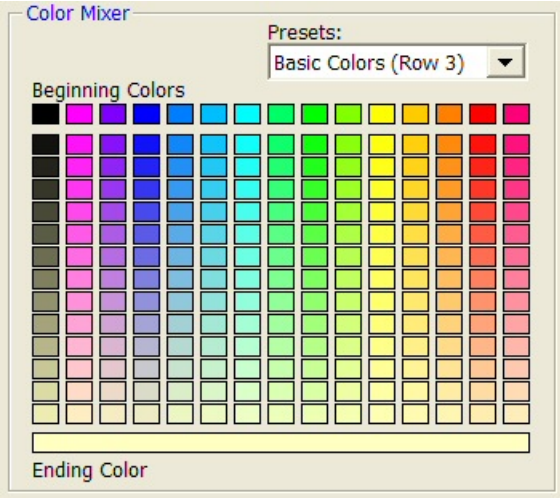
The Color Selection Window

The Color Selection Window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette

These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



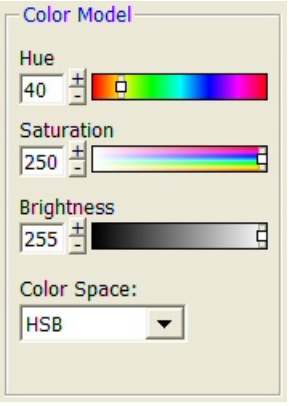
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model

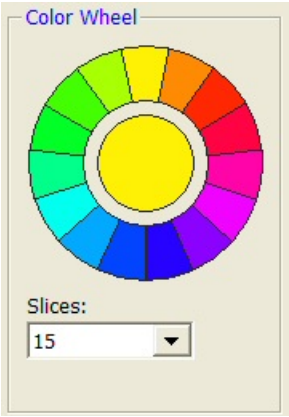


This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



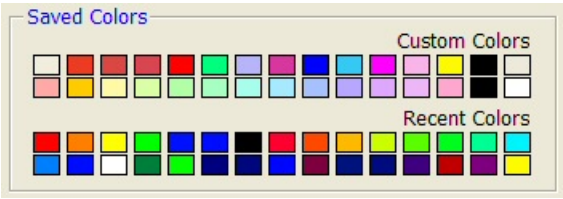
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Active Colors

This box displays the original (Old) color and the selected (New) color.

The resulting color can be set to the original color by dragging and dropping the Old color on the New color.

Either of these colors can be dragged/dropped to any other color box.

180-4 Color Selection Window

Chapter 181

Symbol Settings Window

Introduction

The Symbol Settings window specifies the characteristics of a plotting symbol. The options are placed under two tabs. The first (Color) tab contains the options for specifying the colors of the symbol. The second (Symbol Settings) tab contains the options for specifying the type and size of the symbol.

Window Options

Color Tab

The Color tab lets you pick the fill and border colors of the plot symbol. The window is made up of several components that allow you to pick colors. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

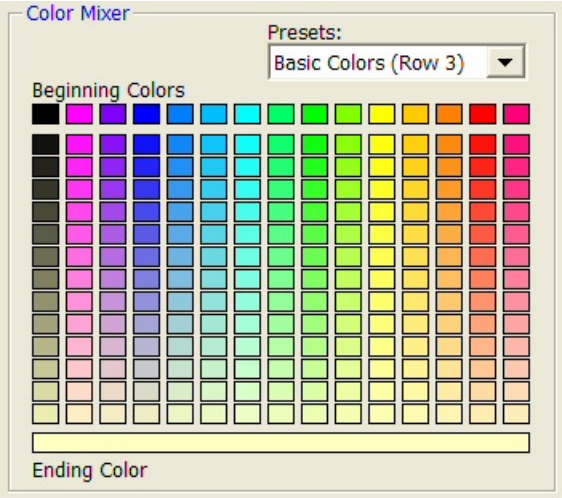
Clicks Modify Fill / Border

This option indicates whether color selections apply to the interior (fill) or border of the symbol.

Basic Palette

These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



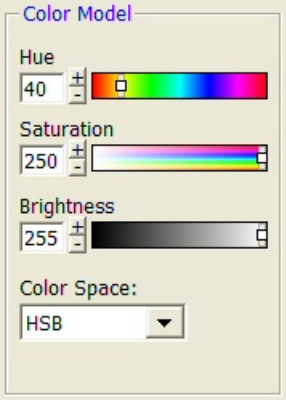
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model

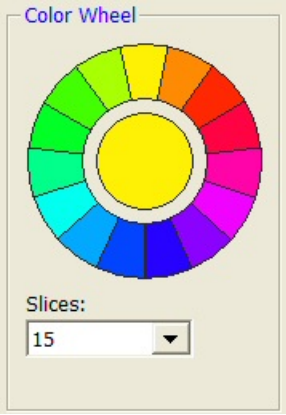


This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors

The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Symbol Settings Tab

The Symbol Settings tab selects various attributes of the symbol.

Symbol Appearance

Type

This option specifies the type of symbol. If the symbol type only uses one color (such as a letter), the border color is used.

Radius

The radius specifies the size of the symbol. The default value of 100 works well in most cases.

Fill

Pattern

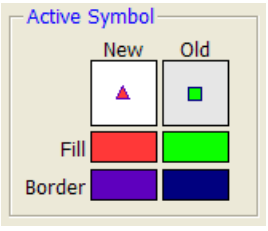
This is the pattern used to fill the interior of a solid symbol. We recommend setting this option to 'solid'.

Border

Width

This option specifies the width of the border.

Active Symbol



This box displays the symbol using the original (Old) settings and the selected (New) settings. Any of the colors can be dragged/dropped to any other color box.

181-4 Symbol Settings Window

Chapter 182

Text Settings Window

Introduction

The Text Settings window specifies the characteristics (color and format) of a line of text, such as a label or title. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Text Settings) tab contains the options for specifying the format of the text.

Window Options

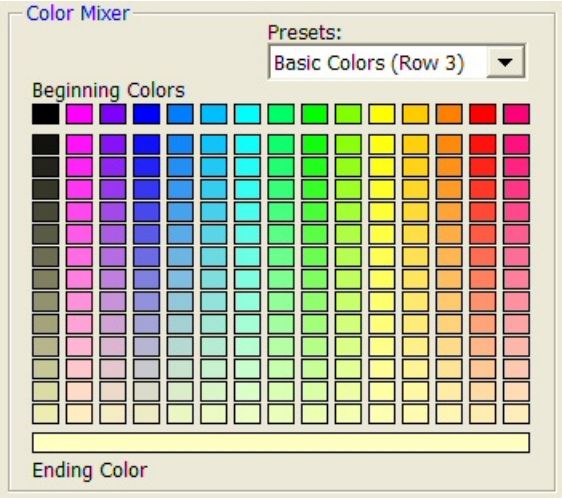
Color Tab

The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette

These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



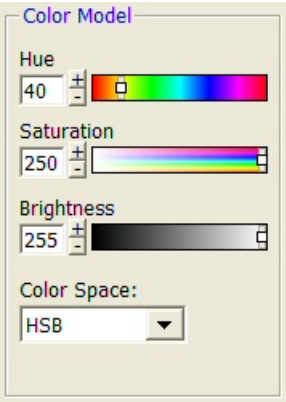
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model

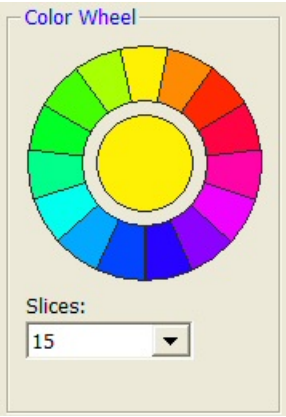


This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors

The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Text Settings Tab

The Text Settings tab selects various attributes of the title or label.

Text Attributes

Font Size

This option sets the font size of the text. The minimum font size is '1'.

Bold, Italic, and Underline

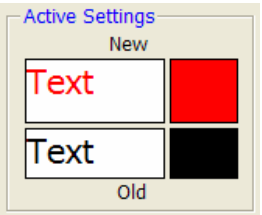
Checking any of these items causes the text to have the corresponding attribute.

Text Value

Actual Text

This option sets the actual text that is to be displayed.

Active Settings



This box displays the original (Old) text format and the selected (New) text format.

The resulting format can be set to the original format by dragging and dropping the Old format on the New format.

Either of these colors can be dragged/dropped to any other color box

182-4 Text Settings Window

Chapter 183

Line Settings Window

Introduction

The Line Settings window specifies the characteristics (color, width, and pattern) of a line. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Settings) tab contains the options for specifying the line width and pattern.

Window Options

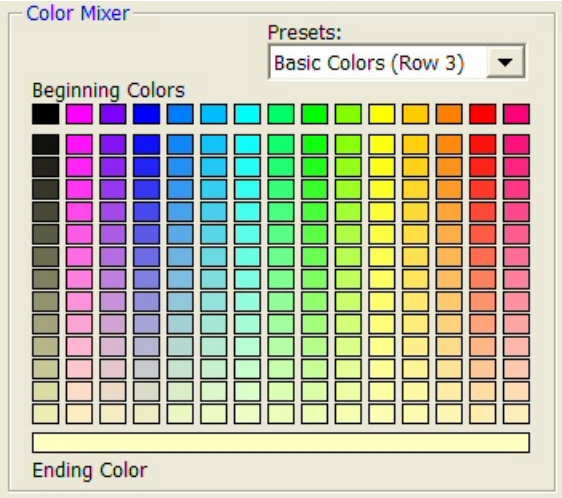
Color Tab

The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette

These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



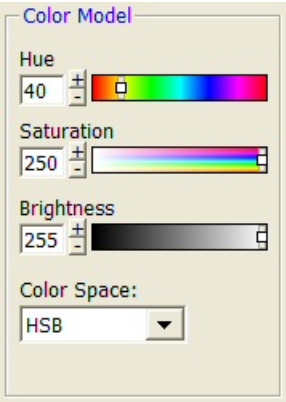
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model

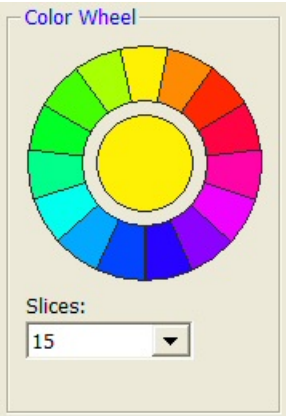


This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors

The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Settings Tab

The Settings tab selects the width and pattern of the line.

Specify Line Settings

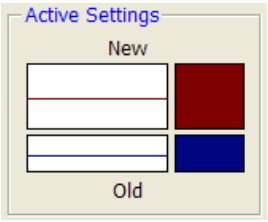
Width

This option sets the width of the line. The minimum line width is '1'.

Line Pattern

This option sets the pattern of the line. Patterns other than 'solid' require the width to be less than 30. Note that the line pattern was more useful in the days before color printers became common. Nowadays, non-solid lines are used less frequently.

Active Settings



This box displays the original (Old) line color and the selected (New) line color.

The resulting line can be set to the original line by dragging and dropping the Old line on the New line.

Either of these colors can be dragged/dropped to any other color box.

183-4 Line Settings Window

Chapter 184

Axis-Line Settings Window

Introduction

The Axis-Line Settings window specifies the characteristics (color and width) of the line used as the axis. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Width & Type) tab contains the options for specifying the line width and type.

Window Options

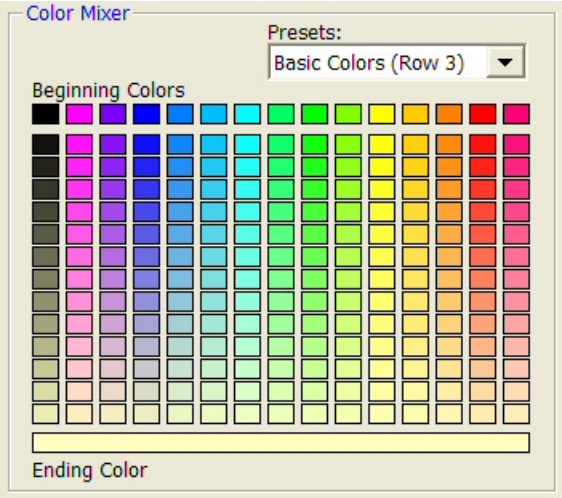
Color Tab

The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette

These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



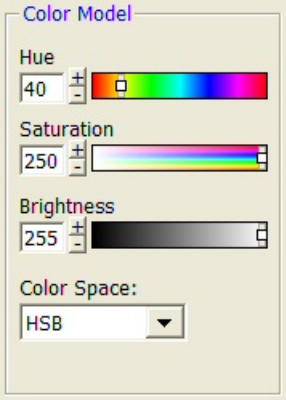
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model

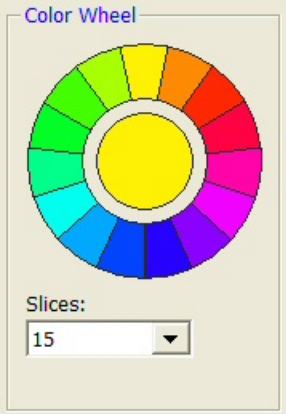


This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors

The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Width & Type Tab

The Width & Type tab selects the width and pattern of an axis line.

Specify Width

Width

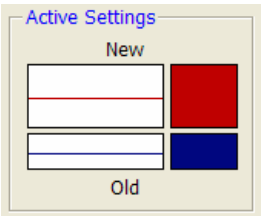
This option sets the width of the line. The minimum line width is '1'.

Specify Type

Type

This option whether the axis is text or numeric. When Text is selected, each unique value is equally-spaced along the axis.

Active Settings



This box displays the original (Old) line color and the selected (New) line color.

The resulting line can be set to the original line by dragging and dropping the Old line on the New line.

Either of these colors can be dragged/dropped to any other color box.

184-4 Axis-Line Settings Window

Chapter 185

Grid / Tick Settings Window

Introduction

The Grid / Tick Settings window specifies the characteristics (color and format) of the tickmarks displayed along the axis as well as the grid lines in the body of the plot. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Grid & Tick Settings) tab contains the options for specifying the format of the grid lines and tickmarks.

Window Options

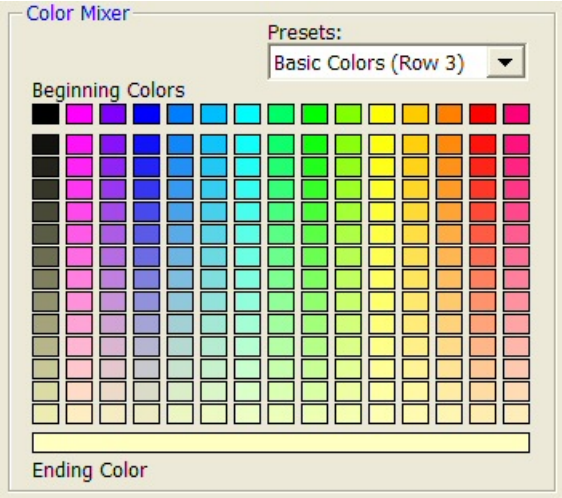
Color Tab

The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette

These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



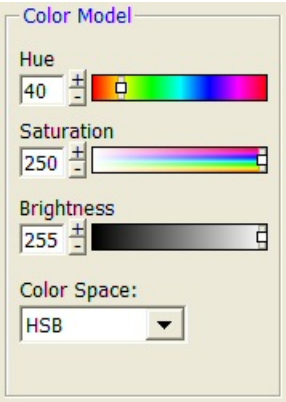
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model

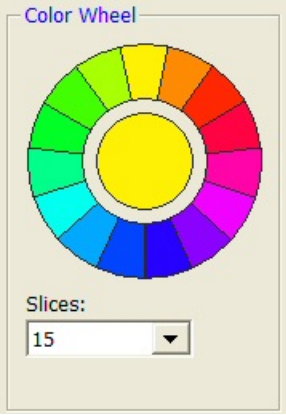


This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors

The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Grid & Tick Settings Tab

The Grid & Tick Settings tab sets various attributes of the grid lines and tickmarks.

Grid Settings

Grid Width

This option sets the width of the grid lines. The minimum width is '1'.

Grid Pattern

This option sets the pattern of the grid line. Patterns other than 'solid' require the width to be less than 30. Note that the grid line pattern was more useful in the days before color printers became common. Nowadays, non-solid lines are used less frequently.

Tick Settings

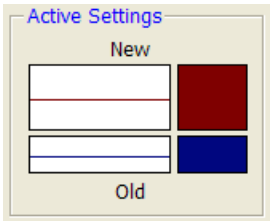
Tick Width

This option sets the width (size) of the tickmarks. The minimum width is '1'.

Tick Length

This option determines the length of the tickmarks.

Active Settings



This box displays the original (Old) grid line and tick color and the selected (New) grid line and tick color.

The resulting line and color can be set to the original line and color by dragging and dropping the Old line onto the New line.

Either of these colors can be dragged/dropped to any other color box.

185-4 Grid / Tick Settings Window

Chapter 186

Tick Label Settings Window

Introduction

The tick label settings window specifies the characteristics (color and format) of the reference numbers displayed at the tickmarks along the axis. The options are placed under two tabs. The first (Color) tab contains the options for specifying the color. The second (Text Settings) tab contains the options for specifying the format of the reference numbers.

Window Options

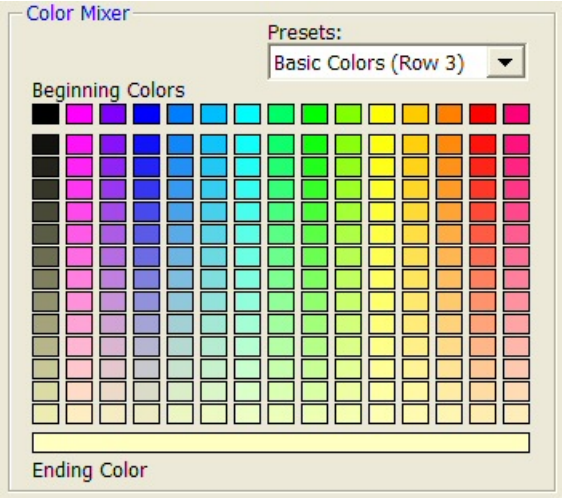
Color Tab

The Color tab window is made up of five basic components that allow you to pick a set of colors with various characteristics. These components are the Basic Palette, the Color Mixer, the Color Model, the Color Wheel, and the Saved Colors. We will now describe each of these components in turn.

Basic Palette

These boxes hold a set of common colors. To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Mixer



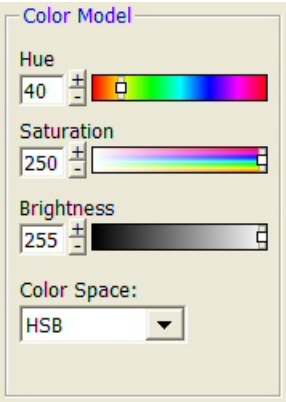
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model

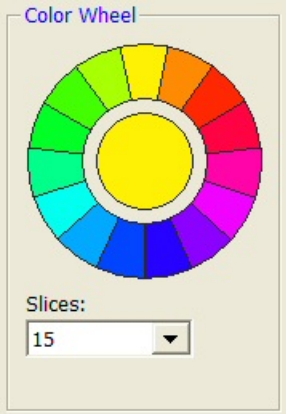


This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



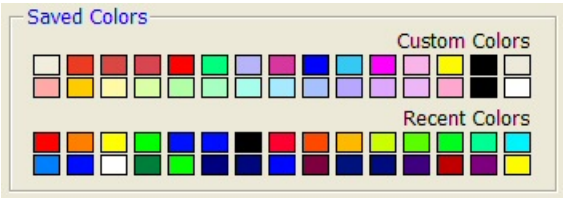
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Text Settings Tab

The Text Settings tab selects various attributes of the reference numbers.

Text Attributes

Font Size

This option sets the font size of the number. The minimum font size is '1'.

Bold, Italic, and Underline

Checking any of these items causes the text to have the corresponding attribute.

Text Format

Decimals

This option determines the number of decimal places that are displayed.

Max Characters

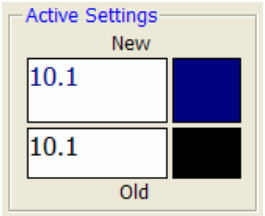
This option determines how much space is provided for displaying the reference numbers.

Text Rotation

Horizontal and Vertical

This option determines whether the text is displayed horizontally or vertically.

Active Settings



This box displays the original (Old) reference number format and the selected (New) reference number format.

The resulting format can be set to the original format by dragging and dropping the Old format on the New format.

Either of these colors can be dragged/dropped to any other color box.

186-4 Tick Label Settings Window

Chapter 187

Heat Map Settings Window

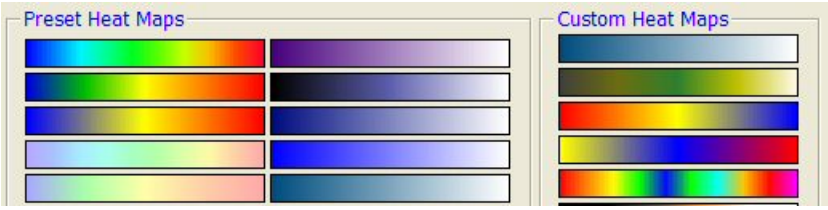
Introduction

The Heat Map Settings window specifies the characteristics of a heat map. The options are placed under three tabs. The first (Heat Maps) tab contains several preconfigured heat maps. The second (Color Selection) tab contains the options for specifying the colors of the heat map. The third (Intervals & Scaling) tab contains the options for specifying the number of color-intervals and the scaling of the heat map.

Window Options

Heat Maps Tab

The Heat Maps tab displays several preset and heat maps heat map patterns that can be selected.



Preset Heat Maps

We have provided several heat map patterns, which show the rich heat map variety that is possible. Simply click on a heat map to activate it. You will see it appear as the selected heat map in the upper right of the window.

Custom Heat Maps

The custom heat maps are heat maps that you create that you want to save for later use. They are stored and will be available until you change them.

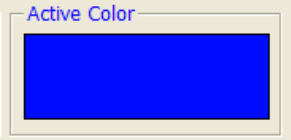
Color Selection Tab

The Color Selection tab lets you pick the colors to be blended to form a heat map. The window is made up of several components that allow you to pick colors. We will now describe each of these components in turn.

Basic Palette

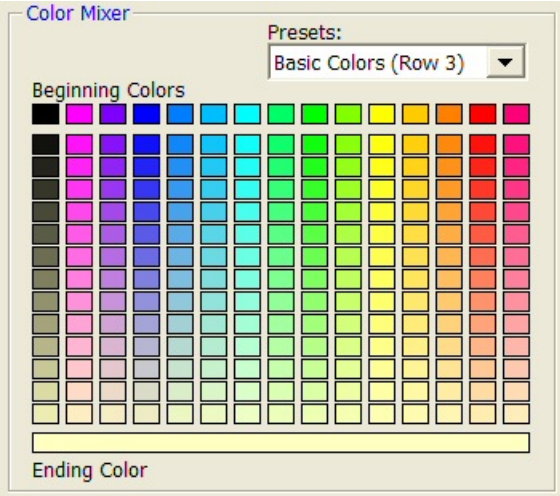
These boxes hold a set of common colors. To use a color, click it and change the currently active color in the Heat Map colors frame or drag/drop it to one of the Heat Map colors.

Active Color



This box provides a large view of the currently-selected color. As you select different colors, this window will change.

Color Mixer



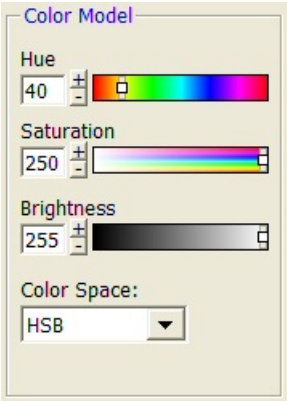
The colors in the body of this section are formed by mixing the colors in the Beginning Colors (the first row) with the Ending Color at the bottom.

The Beginning Colors boxes can be changed by dragging and dropping other colors, or a different pattern can be selected from the Presets box.

Setting the Presets to Mix First and Last allows you to mix colors both horizontally and vertically.

To use a color, click it, double-click it, or drag/drop it somewhere else on the window.

Color Model

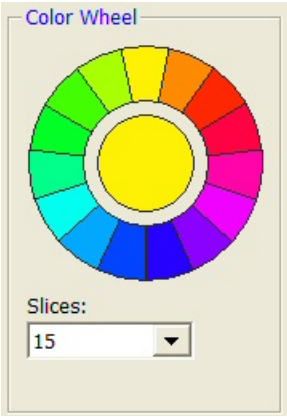


This component specifies individual colors using a specific color model. The default color model is HSB (hue, saturation, brightness). The other models are RGB, CYM, HSI, and HSL. In each case, the individual values range from 0 to 255.

To select several matching colors, set the Saturation and Brightness to the desired value. Vary the Hue. Each time a desirable color is found, drag/drop it to one of the Custom Colors boxes.

Or, colors can be selected from the Color Wheel since it is synchronized with this option.

Color Wheel



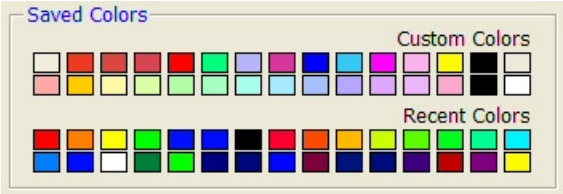
The color wheel displays a harmonious set of colors around the color spectrum. Colors on the wheel have the same saturation and brightness. Colors on opposite sides of the circle are the most contrasting.

Click on a slice to use that color. Drag 'slices' to Custom Colors boxes to save those colors for later.

Set the Color Mixer Presets option to 'Color Wheel' to synchronize Beginning Colors of the Color Mixer with the Color Wheel.

The number of segments on the wheel is controlled by the 'Slices' parameter.

Saved Colors



The Saved Colors are saved after this window is closed. Colors in the Custom Colors are saved from one session to the next.

Colors in the Recent Colors are replaced as new colors are used.

All of these colors can be changed by dragging and dropping another color here.

Intervals & Scaling Tab

The Intervals & Scaling tab has options that determine the intervals and scale of the heatmap.

Number of Intervals

Intervals

This option specifies the number of intervals along the heat map. The default value of 100 seems to work well in most cases.

Scale Along the Heat Map Axis

Scale

This option specifies the type of scaling that is used.

- **Regular**

  The range from the minimum to the maximum is divided up into equal-width intervals.

- **Percentile**

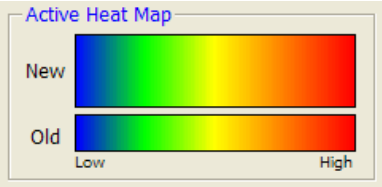
  The range from the minimum to the maximum is divided into percentiles. If the number of intervals is set to 100, then the values in the first percentile receive the first color, the values in the second percentile receive the second color, and so on.

  This scale method forces all colors of the heat map to be displayed. Care must be used when interpreting the data since the intervals are not equally spaced.

- **Log**

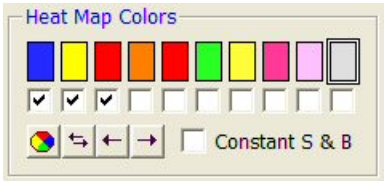
  The values are displayed on according to a logarithmic scale.

Active Heat Map



This box displays the heat map using the original (Old) settings and the selected (New) settings. This heat map can be dragged/dropped to any of the Custom Heat Maps under the Heat Maps tab.

Heat Map Colors



This frame displays the colors that are used to form the heat map. The gray box on the right has a border around it, which indicates that if a color box is clicked, it will change the color of this box.

Even though a color appears in one of these boxes, it will not be used in the heat map unless the check box immediate beneath it is checked.

A row of special buttons appears below the check boxes. We will now describe each of these buttons.

Color Wheel Button

Clicking the first button causes the colors displayed on the Color Wheel to be transferred to these boxes.

Reverse Order Button

Clicking the second button causes the order of the buttons to be reversed.

Shift Left Button

Clicking the third button causes the colors to be shifted one box to the left.

Shift Right Button

Clicking the fourth button causes the colors to be shifted one box to the right.

Constant S & B

Checking this check box causes all of the boxes to by reset so that they have the same saturation and brightness as the highlighted box, while maintaining the same hue.

187-6 Heat Map Settings Window

References

A

Agresti, A. and Coull, B. 1998. “Approximate is Better than ‘Exact’ for Interval Estimation of Binomial Proportions,” *American Statistician*, Volume 52 Number 2, pages 119-126.

A’Hern, R. P. A. 2001. "Sample size tables for exact single-stage phase II designs." *Statistics in Medicine*, Volume 20, pages 859-866.

AIAG (Automotive Industry Action Group). 1995. *Measurement Systems Analysis*. This booklet was developed by Chrysler/Ford/GM Supplier Quality Requirements Task Force. It gives a detailed discussion of how to design and analyze an R&R study. The book may be obtained from ASQC or directly from AIAG by calling 801-358-3570.

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle," In B. N. Petrov & F. Csaki (Eds.), *The second international symposium on information theory*. Budapest, Hungary: Akademiai Kiado.

Akaike, H. 1974. "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, (6): pages 716-723.

Albert, A. and Harris, E. 1987. *Multivariate Interpretation of Clinical Laboratory Data*. Marcel Dekker, New York, New York. This book is devoted to a discussion of how to apply multinomial logistic regression to medical diagnosis. It contains the algorithm that is the basis of our multinomial logistic regression routine.

Allen, D. and Cady, F.. 1982. *Analyzing Experimental Data by Regression*. Wadsworth. Belmont, Calif. This book works completely through several examples. It is very useful to those who want to see complete analyses of complex data.

Al-Sunduqchi, Mahdi S. 1990. *Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics*. Ph.D. dissertation under the direction of William C. Guenther, Dept. of Statistics, University of Wyoming, Laramie, Wyoming.

Altman, Douglas. 1991. *Practical Statistics for Medical Research*. Chapman & Hall. New York, NY. This book provides an introductory discussion of many statistical techniques that are used in medical research. It is the only book we found that discussed ROC curves.

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. 1997. *Statistical Models Based on Counting Processess*. Springer-Verlag, New York. This is an advanced book giving many of the theoretically developments of survival analysis.

Anderson, R.L. and Hauck, W.W. 1983. "A new Procedure for testing equivalence in comparative bioavailability and other clinical trials." *Commun. Stat. Theory Methods.*, Volume 12, pages 2663-2692.

Anderson, T.W. and Darling, D.A. 1954. "A test of goodness-of-fit." *J. Amer. Statist. Assoc*, Volume 49, pages 765-769.

Andrews, D.F., and Herzberg, A.M. 1985. *Data*. Springer-Verlag, New York. This book is a collection of many different data sets. It gives a complete description of each.

Armitage. 1955. “Tests for linear trends in proportions and frequencies.” *Biometrics*, Volume 11, pages 375-386.

Armitage, P., and Colton, T. 1998. *Encyclopedia of Biostatistics*. John Wiley, New York.

Armitage,P., McPherson, C.K., and Rowe, B.C. 1969. “Repeated significance tests on accumulating data.” *Journal of the Royal Statistical Society, Series A,* 132, pages 235-244.

Atkinson, A.C. 1985. *Plots, Transformations, and Regression*. Oxford University Press, Oxford (also in New York). This book goes into the details of regression diagnostics and plotting. It puts together much of the recent work in this area.

Atkinson, A.C., and Donev, A.N. 1992. *Optimum Experimental Designs*. Oxford University Press, Oxford. This book discusses D-Optimal designs.

Austin, P.C., Grootendorst, P., and Anderson, G.M. 2007. “A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study,” *Statistics in Medicine*, Volume 26, pages 734-753.

B

Bain, L.J. and Engelhardt, M. 1991. *Statistical Analysis of Reliability and Life-Testing Models*. Marcel Dekker. New York. This book contains details for testing data that follow the exponential and Weibull distributions.

Baker, Frank. 1992. *Item Response Theory*. Marcel Dekker. New York. This book contains a current overview of IRT. It goes through the details, providing both formulas and computer code. It is not light reading, but it will provide you with much of what you need if you are attempting to use this technique.

Barnard, G.A. 1947. "Significance tests for 2 x 2 tables." *Biometrika* 34:123-138.

Barrentine, Larry B. 1991. *Concepts for R&R Studies*. ASQC Press. Milwaukee, Wisconsin. This is a very good applied work book on the subject of repeatability and reproducibility studies. The ISBN is 0-87389-108-2. ASQC Press may be contacted at 800-248-1946.

Bartholomew, D.J. 1963. "The Sampling Distribution of an Estimate Arising in Life Testing." *Technometrics*, Volume 5 No. 3, 361-374.

Bartlett, M.S. 1950. "Tests of significance in factor analysis." *British Journal of Psychology (Statistical Section)*, 3, 77-85.

Bates, D. M. and Watts, D. G. 1981. "A relative offset orthogonality convergence criterion for nonlinear least squares," *Technometrics*, Volume 23, 179-183.

Beal, S. L. 1987. "Asymptotic Confidence Intervals for the Difference between Two Binomial Parameters for Use with Small Samples." *Biometrics*, Volume 43, Issue 4, 941-950.

Belsley, Kuh, and Welsch. 1980. *Regression Diagnostics*. John Wiley & Sons. New York. This is the book that brought regression diagnostics into the main-stream of statistics. It is a graduate level treatise on the subject.

Benjamini, Y. and Hochberg, Y. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 57, No. 1, 289-300.

Bertsekas, D.P. 1991. *Linear Network Optimization: Algorithms and Codes*. MIT Press. Cambridge, MA.

Blackwelder, W.C. 1993. "Sample size and power in prospective analysis of relative risk." *Statistics in Medicine*, Volume 12, 691-698.

Blackwelder, W.C. 1998. "Equivalence Trials." In *Encyclopedia of Biostatistics*, John Wiley and Sons. New York. Volume 2, 1367-1372.

Bloomfield, P. 1976. *Fourier Analysis of Time Series*. John Wiley and Sons. New York. This provides a technical introduction to fourier analysis techniques.

Bock, R.D., Aiken, M. 1981. "Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika*, 46, 443-459.

Bolstad, B.M., et al. 2003. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, 19, 185-193.

Bonett, Douglas. 2002. "Sample Size Requirements for Testing and Estimating Coefficient Alpha." *Journal of Educational and Behavioral Statistics*, Vol. 27, pages 335-340.

Box, G.E.P. and Jenkins, G.M. 1976. *Time Series Analysis - Forecasting and Control*. Holden-Day.: San Francisco, California. This is the landmark book on ARIMA time series analysis. Most of the material in chapters 6 - 9 of this manual comes from this work.

Box, G.E.P. 1949. "A general distribution theory for a class of likelihood criteria." *Biometrika*, 1949, **36**, 317-346.

Box, G.E.P. 1954a. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variable Problems: I." *Annals of Mathematical Statistics*, **25**, 290-302.

Box, G.E.P. 1954b. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variable Problems: II." *Annals of Mathematical Statistics*, **25**, 484-498.

Box, G.E.P., Hunter, S. and Hunter. 1978. *Statistics for Experimenters*. John Wiley & Sons, New York. This is probably the leading book in the area experimental design in industrial experiments. You definitely should acquire and study this book if you plan anything but a casual acquaintance with experimental design. The book is loaded with examples and explanations.

Breslow, N. E. and **Day, N. E.** 1980. *Statistical Methods in Cancer Research: Volume 1. The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.

Brown, H., and Prescott, R. 2006. *Applied Mixed Models in Medicine.* 2nd ed. John Wiley & Sons Ltd. Chichester, West Sussex, England.

Brush, Gary G. 1988. *Volume 12: How to Choose the Proper Sample Size*, American Society for Quality Control, 310 West Wisconsin Ave, Milwaukee, Wisconsin, 53203. This is a small workbook for quality control workers.

Burdick, R.K. and Larsen, G.A. 1997. "Confidence Intervals on Measures of Variability in R&R Studies." *Journal of Quality Technology, Vol. 29, No. 3, Pages 261-273*. This article presents the formulas used to construct confidence intervals in an R&R study.

Bury, Karl. 1999. *Statistical Distributions in Engineering.*. Cambridge University Press. New York, NY. (www.cup.org).

C

Cameron, A.C. and Trivedi, P.K. 1998. *Regression Analysis of Count Data*. Cambridge University Press. New York, NY. (www.cup.org).

Carmines, E.G. and Zeller, R.A. 1990. *Reliability and Validity Assessment*. Sage University Paper. 07-017. Newbury Park, CA.

Casagrande, J. T., Pike, M.C., and Smith, P. G. 1978. "The Power Function of the "Exact" Test for Comparing Two Binomial Distributions," *Applied Statistics*, Volume 27, No. 2, pages 176-180. This article presents the algorithm upon which our Fisher's exact test is based.

Cattell, R.B. 1966. "The scree test for the number of factors." *Mult. Behav. Res.* 1, 245-276.

Cattell, R.B. and Jaspers, J. 1967. "A general plasmode (No. 30-10-5-2) for factor analytic exercises and research." *Mult. Behav. Res. Monographs*. 67-3, 1-212.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. 1983. *Graphicals Methods for Data Analysis.* Duxbury Press, Boston, Mass. This wonderful little book is full of examples of ways

to analyze data graphically. It gives complete (and readable) coverage to such topics as scatter plots, probability plots, and box plots. It is strongly recommended.

Chatfield, C. 1984. *The Analysis of Time Series*. Chapman and Hall. New York. This book gives a very readable account of both ARMA modeling and spectral analysis. We recommend it to those who wish to get to the bottom of these methods.

Chatterjee and Price. 1979. *Regression Analysis by Example*. John Wiley & Sons. New York. A great hands-on book for those who learn best from examples. A newer edition is now available.

Chen, K.W.; Chow, S.C.; and Li, G. 1997. "A Note on Sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs" *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 25, No. 6, pages 753-765.

Chen, T. T. 1997. "Optimal Three-Stage Designs for Phase II Cancer Clinical Trials." *Statistics in Medicine*, Volume 16, pages 2701-2711.

Chen, Xun. 2002. "A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases." *Statistics in Medicine*, Volume 21, pages 943-956.

Chow, S.C. and Liu, J.P. 1999. *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker. New York.

Chow, S.C.; Shao, J.; Wang, H. 2003. *Sample Size Calculations in Clinical Research*. Marcel Dekker. New York.

Chow, S.-C.; Shao, J.; Wang, H. 2008. *Sample Size Calculations in Clinical Research, Second Edition*. Chapman & Hall/CRC. Boca Raton, Florida.

Cochran and Cox. 1992. *Experimental Designs. Second Edition*. John Wiley & Sons. New York. This is one of the classic books on experimental design, first published in 1957.

Cochran, W.G. and Rubin, D.B. 1973. “Controlling bias in observational studies,” *Sankhya, Ser. A*, Volume 35, Pages 417-446.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, New Jersey. This is a very nice, clearly written book. There are MANY examples. It is the largest of the sample size books. It does not deal with clinical trials.

Cohen, Jacob. 1990. “Things I Have Learned So Far.” *American Psychologist*, December, 1990, pages 1304-1312. This is must reading for anyone still skeptical about the need for power analysis.

Collett, D. 1991. *Modelling Binary Data.* Chapman & Hall, New York, New York. This book covers such topics as logistic regression, tests of proportions, matched case-control studies, and so on.

Collett, D. 1994. *Modelling Survival Data in Medical Research.* Chapman & Hall, New York, New York. This book covers such survival analysis topics as Cox regression and log rank tests.

Conlon, M. and Thomas, R. 1993. “The Power Function for Fisher’s Exact Test.” *Applied Statistics*, Volume 42, No. 1, pages 258-260. This article was used to validate the power calculations of Fisher’s Exact Test in PASS. Unfortunately, we could not use the algorithm to improve the speed because the algorithm requires equal sample sizes.

Conover, W.J. 1971. *Practical Nonparametric Statistics*. John Wiley & Sons, Inc. New York.

Conover, W.J., Johnson, M.E., and **Johnson, M.M.** 1981. *Technometrics***, 23,** 351-361**.**

Cook, D. and Weisberg, S. 1982. *Residuals and Influence in Regression*. Chapman and Hall. New York. This is an advanced text in the subject of regression diagnostics.

Cooley, W.W. and Lohnes, P.R. 1985. *Multivariate Data Analysis*. Robert F. Krieger Publishing Co. Malabar, Florida.

Cox, D. R. 1972. “Regression Models and life tables.” *Journal of the Royal Statistical Society, Series B*, Volume 34, Pages 187-220. This article presents the proportional hazards regression model.

Cox, D. R. 1975. “Contribution to discussion of Mardia (1975a).” *Journal of the Royal Statistical Society, Series B*, Volume 37, Pages 380-381.

Cox, D.R. and Snell, E.J. 1981. *Applied Statistics: Principles and Examples*. Chapman & Hall. London, England.

Cureton, E.E. and D'Agostino, R.B. 1983. *Factor Analysis - An Applied Approach*. Lawrence Erlbaum Associates. Hillsdale, New Jersey. (This is a wonderful book for those who want to learn the details of what factor analysis does. It has both the theoretical formulas and simple worked examples to make following along very easy.)

D

D'Agostino, R.B., Belanger, A., D'Agostino, R.B. Jr. 1990."A Suggestion for Using Powerful and Informative Tests of Normality.", *The American Statistician*, November 1990, Volume 44 Number 4, pages 316-321. This tutorial style article discusses D'Agostino’s tests and tells how to interpret normal probability plots.

D'Agostino, R.B., Chase, W., Belanger, A. 1988."The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations.", *The American Statistician*, August 1988, Volume 42 Number 3, pages 198-202.

D'Agostino, R.B. Jr. 2004. *Tutorials in Biostatistics*. Volume 1. John Wiley & Sons. Chichester, England.

Dallal, G. 1986. "An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality," *The American Statistician*, Volume 40, Number 4, pages 294-296.

Daniel, C. and Wood, F. 1980. *Fitting Equations to Data*. John Wiley & Sons. New York. This book gives several in depth examples of analyzing regression problems by computer.

Daniel, W. 1990. *Applied Nonparametric Statistics.* 2nd ed. PWS-KENT Publishing Company. Boston.

Davies, Owen L. 1971. *The Design and Analysis of Industrial Experiments*. Hafner Publishing Company, New York. This was one of the first books on experimental design and analysis. It has many examples and is highly recommended.

Davis, J. C. 1985. *Statistics and Data Analysis in Geology*. John Wiley. New York. (A great layman's discussion of many statistical procedures, including factor analysis.)

Davison, A.C. and Hinkley, D.V. 1999. *Bootstrap Methods and their Applications*. Cambridge University Press. NY, NY. This book provides and detailed account of bootstrapping.

Davison, Mark. 1983. *Multidimensional Scaling*. John Wiley & Sons. NY, NY. This book provides a very good, although somewhat advanced, introduction to the subject.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. 1988. “Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics*, 44, pages 837-845.

DeMets, D.L. and Lan, K.K.G. 1984. “An overview of sequential methods and their applications in clinical trials.” *Communications in Statistics, Theory and Methods*, 13, pages 2315-2338.

DeMets, D.L. and Lan, K.K.G. 1994. “Interim analysis: The alpha spending function approach.” *Statistics in Medicine*, 13, pages 1341-1352.

Demidenko, E. 2004. *Mixed Models – Theory and Applications*. John Wiley & Sons. Hoboken, New Jersey.

Desu, M. M. and Raghavarao, D. 1990. *Sample Size Methodology*. Academic Press. New York. (Presents many useful results for determining sample sizes.)

DeVor, Chang, and Sutherland. 1992. *Statistical Quality Design and Control*. Macmillan Publishing. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 800 pages.

Devroye, Luc. 1986. *Non-Uniform Random Variate Generation.* Springer-Verlag. New York. This book is currently available online at http://jeff.cs.mcgill.ca/~luc/rnbookindex.html.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. 1994. *Analysis of Longitudinal Data*. Oxford University Press. New York, New York.

Dillon, W. and Goldstein, M. 1984. *Multivariate Analysis - Methods and Applications*. John Wiley. NY, NY. This book devotes a complete chapter to loglinear models. It follows Fienberg's book, providing additional discussion and examples.

Dixon, W. J. and Tukey, J. W. 1968. "Approximate behavior of the distribution of Winsorized t," *Technometrics*, Volume 10, pages 83-98.

Dodson, B. 1994. *Weibull Analysis*. ASQC Quality Press. Milwaukee, Wisconsin. This paperback book provides the basics of Weibull fitting. It contains many of the formulas used in our Weibull procedure.

Donnelly, Thomas G. 1980. "ACM Algorithm 462: Bivariate Normal Distribution," *Collected Algorithms from ACM*, Volume II, New York, New York.

Donner, Allan. 1984. "Approaches to Sample Size Estimation in the Design of Clinical Trials--A Review," *Statistics in Medicine*, Volume 3, pages 199-214. This is a well done review of the clinical trial literature. Although it is becoming out of date, it is still a good place to start.

Donner, A. and Klar, N. 1996. "Statistical Considerations in the Design and Analysis of Community Intervention Trials." *The Journal of Clinical Epidemiology*, Vol. 49, No. 4, 1996, pages 435-439.

Donner, A. and Klar, N. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research.* Arnold. London.

Draghici, S. 2003. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC. London. This is an excellent overview of most areas of Microarray analysis.

Draper, N.R. and Smith, H. 1966. *Applied Regression Analysis*. John Wiley & Sons. New York. This is a classic text in regression analysis. It contains both in depth theory and applications. This text is often used in graduate courses in regression analysis.

Draper, N.R. and Smith, H. 1981. *Applied Regression Analysis - Second Edition*. John Wiley & Sons. New York, NY. This is a classic text in regression analysis. It contains both in-depth theory and applications. It is often used in graduate courses in regression analysis.

Dudoit, S., Shaffer, J.P., and Boldrick, J.C. 2003. "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, Volume 18, No. 1, pages 71-103.

Dudoit, S., Yang, Y.H., Callow, M.J., and **Speed, T.P.** 2002. "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Experiments," *Statistica Sinica*, Volume 12, pages 111-139.

du Toit, S.H.C., Steyn, A.G.W., and Stumpf, R.H. 1986. *Graphical Exploratory Data Analysis.* Springer-Verlag. New York. This book contains examples of graphical analysis for a broad range of topics.

Dunn, O. J. 1964. "Multiple comparisons using rank sums," *Technometrics*, Volume 6, pages 241-252.

Dunnett, C. W. 1955. "A Multiple comparison procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, Volume 50, pages 1096-1121.

Dunteman, G.H. 1989. *Principal Components Analysis*. Sage University Papers, 07-069. Newbury Park, California. Telephone (805) 499-0721. This monograph costs only $7. It gives a very good introduction to PCA.

Dupont, William. 1988. "Power Calculations for Matched Case-Control Studies," *Biometrics*, Volume 44, pages 1157-1168.

Dupont, William and **Plummer, Walton D.** 1990. "Power and Sample Size Calculations--A Review and Computer Program," *Controlled Clinical Trials*, Volume 11, pages 116-128. Documents a nice public-domain program on sample size and power analysis.

Durbin, J. and Watson, G. S. 1950. "Testing for Serial Correlation in Least Squares Regression - I," *Biometrika*, Volume 37, pages 409-428.

Durbin, J. and Watson, G. S. 1951. "Testing for Serial Correlation in Least Squares Regression - II," *Biometrika*, Volume 38, pages 159-177.

Dyke, G.V. and Patterson, H.D. 1952. "Analysis of factorial arrangements when the data are proportions." *Biometrics*. Volume 8, pages 1-12. This is the source of the data used in the LLM tutorial.

E

Eckert, Joseph K. 1990. *Property Appraisal and Assessment Administration*. International Association of Assessing Officers. 1313 East 60th Street. Chicago, IL 60637-2892. Phone: (312) 947-2044. This is a how-to manual published by the IAAO that describes how to apply many statistical procedures to real estate appraisal and tax assessment. We strongly recommend it to those using our *Assessment Model* procedure.

Edgington, E. 1987. *Randomization Tests*. Marcel Dekker. New York. A comprehensive discussion of randomization tests with many examples.

Edwards, L.K. 1993. *Applied Analysis of Variance in the Behavior Sciences*. Marcel Dekker. New York. Chapter 8 of this book is used to validate the repeated measures module of PASS.

Efron, B. and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall. New York.

Elandt-Johnson, R.C. and Johnson, N.L. 1980. *Survival Models and Data Analysis*. John Wiley. NY, NY. This book devotes several chapters to population and clinical life-table analysis.

Epstein, Benjamin. 1960. "Statistical Life Test Acceptance Procedures." *Technometrics*. Volume 2.4, pages 435-446.

Everitt, B.S. and Dunn, G. 1992. *Applied Multivariate Data Analysis*. Oxford University Press. New York. This book provides a very good introduction to several multivariate techniques. It helps you understand how to interpret the results.

F

Farrington, C. P. and Manning, G. 1990. "Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk." *Statistics in Medicine*, Vol. 9, pages 1447-1454. This article contains the formulas used for the Equivalence of Proportions module in PASS.

Feldt, L.S.; Woodruff, D.J.; & Salih, F.A. 1987. "Statistical inference for coefficient alpha." *Applied Psychological Measurement*, Vol. 11, pages 93-103.

Feldt, L.S.; Ankenmann, R.D. 1999. "Determining Sample Size for a Test of the Equality of Alpha Coefficients When the Number of Part-Tests is Small." *Psychological Methods*, Vol. 4(4), pages 366-377.

Fienberg, S. 1985. *The Analysis of Cross-Classified Categorical Data*. MIT Press. Cambridge, Massachusetts. This book provides a very good introduction to the subject. It is a must for any serious student of the subject.

Finney, D. 1971. *Probit Analysis*. Cambridge University Press. New York, N.Y.

Fisher, N.I. 1993. *Statistical Analysis of Circular Data*. Cambridge University Press. New York, New York.

Fisher, R.A. 1936. "The use of multiple measurements in taxonomic problems." *Annuals of Eugenics*, Volume 7, Part II, 179-188. This article is famous because in it Fisher included the 'iris data' that is always presented when discussing discriminant analysis.

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, J. L., Levin, B., Paik, M.C. 2003. *Statistical Methods for Rates and Proportions. Third Edition.* John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, Joseph L. 1986. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons. New York. This book provides a very good introduction to clinical trials. It may be a bit out of date now, but it is still very useful.

Fleming, T. R. 1982. "One-sample multiple testing procedure for Phase II clinical trials." *Biometrics*, Volume 38, pages 143-151.

Flury, B. and Riedwyl, H. 1988. *Multivariate Statistics: A Practical Approach*. Chapman and Hall. New York. This is a short, paperback text that provides lots of examples.

Flury, B. 1988. *Common Principal Components and Related Multivariate Models*. John Wiley & Sons. New York. This reference describes several advanced PCA procedures.

G

Gans. 1984. "The Search for Significance: Different Tests on the Same Data." *The Journal of Statistical Computation and Simulation*, 1984, pages 1-21.

Gart, John J. and Nam, Jun-mo. 1988. "Approximate Interval Estimation of the Ratio in Binomial Parameters: A Review and Corrections for Skewness." *Biometrics*, Volume 44, Issue 2, 323-338.

Gart, John J. and Nam, Jun-mo. 1990. "Approximate Interval Estimation of the Difference in Binomial Parameters: Correction for Skewness and Extension to Multiple Tables." *Biometrics*, Volume 46, Issue 3, 637-643.

Gehlback, Stephen. 1988. *Interpreting the Medical Literature: Practical Epidemiology for Clinicians*. Second Edition. McGraw-Hill. New York. Telephone: (800)722-4726. The preface of this book states that its purpose is to provide the reader with a useful approach to interpreting the quantitative results given in medical literature. We reference it specifically because of its discussion of ROC curves.

Gentle, James E. 1998. *Random Number Generation and Monte Carlo Methods*. Springer. New York.

Gibbons, J. 1976. *Nonparametric Methods for Quantitative Analysis*. Holt, Rinehart and Winston. New York.

Gleason, T.C. and Staelin, R. 1975. "A proposal for handling missing data." *Psychometrika*, 40, 229-252.

Goldstein, Richard. 1989. "Power and Sample Size via MS/PC-DOS Computers," *The American Statistician*, Volume 43, Number 4, pages 253-260. A comparative review of power analysis software that was available at that time.

Gomez, K.A. and Gomez, A. A. 1984. *Statistical Procedures for Agricultural Research*. John Wiley & Sons. New York. This reference contains worked-out examples of many complex ANOVA designs. It includes split-plot designs. We recommend it.

Graybill, Franklin. 1961. *An Introduction to Linear Statistical Models*. McGraw-Hill. New York, New York. This is an older book on the theory of linear models. It contains a few worked examples of power analysis.

Greenacre, M. 1984. *Theory and Applications of Correspondence Analysis*. Academic Press. Orlando, Florida. This book goes through several examples. It is probably the most complete book in English on the subject.

Greenacre, Michael J. 1993. *Correspondence Analysis in Practice*. Academic Press. San Diego, CA. This book provides a self-teaching course in correspondence analysis. It is the clearest exposition on the subject that I have every seen. If you want to gain an understanding of CA, you must obtain this (paperback) book.

Griffiths, P. and Hill, I.D. 1985. *Applied Statistics Algorithms*, The Royal Statistical Society, London, England. See page 243 for ACM algorithm 291.

Gross and Clark 1975. *Survival Distributions*: Reliability Applications in Biomedical Sciences. John Wiley, New York.

Gu, X.S., and Rosenbaum, P.R. 1993. "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms," *Journal of Computational and Graphical Statistics*, Vol. 2, No. 4, pages 405-420.

Guenther, William C. 1977. "Desk Calculation of Probabilities for the Distribution of the Sample Correlation Coefficient," *The American Statistician*, Volume 31, Number 1, pages 45-48.

Guenther, William C. 1977. *Sampling Inspection in Statistical Quality Control*. Griffin's Statistical Monographs, Number 37. London.

H

Haberman, S.J. 1972. "Loglinear Fit of Contingency Tables." *Applied Statistics*. Volume 21, pages 218-225. This lists the fortran program that is used to create our LLM algorithm.

Hahn, G. J. and Meeker, W.Q. 1991. *Statistical Intervals*. John Wiley & Sons. New York.

Hambleton, R.K; Swaminathan, H; Rogers, H.J. 1991. *Fundamentals of Item Response Theory*. Sage Publications. Newbury Park, California. Phone: (805)499-0721. Provides an inexpensive, readable introduction to IRT. A good place to start.

Hamilton, L. 1991. *Regression with Graphics: A Second Course in Applied Statistics*. Brooks/Cole Publishing Company. Pacific Grove, California. This book gives a great introduction to the use of graphical analysis with regression. It is a must for any serious user of regression. It is written at an introductory level.

Hand, D.J. and Taylor, C.C. 1987. *Multivariate Analysis of Variance and Repeated Measures*. Chapman and Hall. London, England.

Hanley, J. A. and McNeil, B. J. 1982. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve." *Radiology*, 143, 29-36. April, 1982.

Hanley, J. A. and McNeil, B. J. 1983. "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases." *Radiology*, 148, 839-843. September, 1983.

Hartigan, J. 1975. *Clustering Algorithms*. John Wiley. New York. (This is the "bible" of cluster algorithms. Hartigan developed the K-means algorithm used in **NCSS**.)

Haupt, R.L. and Haupt, S.E. 1998. *Practical Genetic Algorithms*. John Wiley. New York.

Hernandez-Bermejo, B. and Sorribas, A. 2001. "Analytical Quantile Solution for the S-distribution, Random Number Generation and Statistical Data Modeling." *Biometrical Journal* 43, 1007-1025.

Hintze, J. L. and Nelson, R.D. 1998. "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician* 52, 181-184.

Hoaglin, Mosteller, and Tukey. 1985. *Exploring Data Tables, Trends, and Shapes*. John Wiley. New York.

Hoaglin, Mosteller, and Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons. New York.

Hochberg, Y. and Tamhane, A. C. 1987. *Multiple Comparison Procedures*. John Wiley & Sons. New York.

Hoerl, A.E. and Kennard, R.W. 1970. "Ridge Regression: Biased estimation for nonorthogonal problems." *Technometrics* 12, 55-82.

Hoerl, A.E. and Kennard R.W. 1976. "Ridge regression: Iterative estimation of the biasing parameter." *Communications in Statistics* A5, 77-88.

Howe, W.G. 1969. "Two-Sided Tolerance Limits for Normal Populations—Some Improvements." *Journal of the American Statistical Association*, 64, 610-620.

Hosmer, D. and Lemeshow, S. 1989. *Applied Logistic Regression*. John Wiley & Sons. New York. This book gives an advanced, in depth look at logistic regression.

Hosmer, D. and Lemeshow, S. 1999. *Applied Survival Analysis*. John Wiley & Sons. New York.

Hotelling, H. 1933. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* 24, 417-441, 498-520.

Hsieh, F.Y. 1989. "Sample Size Tables for Logistic Regression," *Statistics in Medicine*, Volume 8, pages 795-802. This is the article that was the basis for the sample size calculations in logistic regression in PASS 6.0. It has been superceded by the 1998 article.

Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. "A Simple Method of Sample Size Calculation for Linear and Logistic Regression," *Statistics in Medicine*, Volume 17, pages 1623-1634. The sample size calculation for logistic regression in PASS are based on this article.

Hsieh, F.Y. and Lavori, P.W. 2000. "Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates," *Controlled Clinical Trials*, Volume 21, pages 552-560. The sample size calculation for Cox regression in PASS are based on this article.

Hsu, Jason. 1996. *Multiple Comparisons: Theory and Methods*. Chapman & Hall. London. This book gives a beginning to intermediate discussion of multiple comparisons, stressing the interpretation of the various MC tests. It provides details of three popular MC situations: all pairs, versus the best, and versus a control. The power calculations used in the MC module of PASS came from this book.

I

Irizarry, R.A., et al. 2003a. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, 4, 249-264.

Irizarry, R.A., et al. 2003b. Summaries of Affymetrix GeneChip Probe Level Data. *Nucleic Acids Research*, 31, e15.

J

Jackson, J.E. 1991. *A User's Guide To Principal Components.* John Wiley & Sons. New York. This is a great book to learn about PCA from. It provides several examples and treats everything at a level that is easy to understand.

James, Mike. 1985. *Classification Algorithms*. John Wiley & Sons. New York. This is a great text on the application of discriminant analysis. It includes a simple, easy-to-understand, theoretical development as well as discussions of the application of discriminant analysis.

Jammalamadaka, S.R. and SenGupta, A. 2001. *Topics in Circular Statistics*. World Scientific. River Edge, New Jersey.

Jobson, J.D. 1992. *Applied Multivariate Data Analysis - Volume II: Categorical and Multivariate Methods*. Springer-Verlag. New York. This book is a useful reference for loglinear models and other multivariate methods. It is easy to follows and provides lots of examples.

Jolliffe, I.T. 1972. "Discarding variables in a principal component analysis, I: Artifical data." *Applied Statistics*, 21:160-173.

Johnson, N.L., Kotz, S., and Kemp, A.W. 1992. *Univariate Discrete Distributions, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions Volume 1, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1995. *Continuous Univariate Distributions Volume 2, Second Edition*. John Wiley & Sons. New York.

Jolliffe, I.T. 1986. *Principal Component Analysis*. Springer-Verlag. New York. This book provides an easy-reading introduction to PCA. It goes through several examples.

Julious, Steven A. 2004. "Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data." *Statistics in Medicine*, 23:1921-1986.

Jung, S.-H. 2005. "Sample size for FDR-control in microarray data analysis" *Bioinformatics*, 21(14):3097-3104.

Juran, J.M. 1979. *Quality Control Handbook*. McGraw-Hill. New York.

K

Kaiser, H.F. 1960. "The application of electronic computers to factor analysis." *Educational and Psychological Measurement*. 20:141-151.

Kalbfleisch, J.D. and Prentice, R.L. 1980. *The Statistical Analysis of Failure Time Data*. John Wiley, New York.

Karian, Z.A and Dudewicz, E.J. 2000. *Fitting Statistical Distributions.* CRC Press, New York.

Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data*. John Wiley. New York. This book gives an excellent introduction to cluster analysis. It treats the forming of the distance matrix and several different types of cluster methods, including fuzzy. All this is done at an elementary level so that users at all levels can gain from it.

Kay, S.M. 1988. *Modern Spectral Estimation.* Prentice-Hall: Englewood Cliffs, New Jersey. A very technical book on spectral theory.

Kendall,M. and Ord, J.K. 1990. *Time Series.* Oxford University Press. New York. This is theoretical introduction to time series analysis that is very readable.

Kendall,M. and Stuart, A. 1987. *Kendall's Advanced Theory of Statistics. Volume 1: Distribution Theory*. Oxford University Press. New York. This is a fine math-stat book for graduate students in statistics. We reference it because it includes formulas that are used in the program.

Kenward, M. G. and Roger, J. H. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*, 53, pages 983-997.

Keppel, Geoffrey. 1991. *Design and Analysis - A Researcher's Handbook.* Prentice Hall. Englewood Cliffs, New Jersey. This is a very readable primer on the topic of analysis of variance. Recommended for those who want the straight scoop with a few, well-chosen examples.

Kirk, Roger E. 1982. *Experimental Design: Procedures for the Behavioral Sciences.* Brooks/Cole. Pacific Grove, California. This is a respected reference on experimental design and analysis of variance.

Klein, J.P. and Moeschberger, M.L.. 1997. *Survival Analysis.* Springer-Verlag. New York. This book provides a comprehensive look at the subject complete with formulas, examples, and lots of useful comments. It includes all the more recent developments in this field. I recommend it.

Koch, G.G.; Atkinson, S.S.; Stokes, M.E. 1986. *Encyclopedia of Statistical Sciences.* Volume 7. John Wiley. New York. Edited by Samuel Kotz and Norman Johnson. The article on Poisson Regression provides a very good summary of the subject.

Kotz and Johnson. 1993. *Process Capability Indices.* Chapman & Hall. New York. This book gives a detailed account of the capability indices used in SPC work. 207 pages.

Kraemer, H. C. and **Thiemann, S.** 1987. *How Many Subjects*, Sage Publications, 2111 West Hillcrest Drive, Newbury Park, CA. 91320. This is an excellent introduction to power analysis.

Kruskal, J. 1964. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika* 29, pages 1-27, 115-129. This article presents the algorithm on which the non-metric algorithm used in NCSS is based.

Kruskal, J. and Wish, M. 1978. *Multidimensional Scaling*. Sage Publications. Beverly Hills, CA. This is a well-written monograph by two of the early pioneers of MDS. We suggest it to all serious students of MDS.

Kuehl, R.O. 2000. *Design of Experiment: Statistical Principles of Research Design and Analysis, 2nd Edition.* Brooks/Cole. Pacific Grove, California. This is a good graduate level text on experimental design with many examples.

L

Lachenbruch, P.A. 1975. *Discriminant Analysis.* Hafner Press. New York. This is an in-depth treatment of the subject. It covers a lot of territory, but has few examples.

Lachin, John M. 2000. *Biostatistical Methods.* John Wiley & Sons. New York. This is a graduate-level methods book that deals with statistical methods that are of interest to biostatisticians such as odds ratios, relative risks, regression analysis, case-control studies, and so on.

Lachin, John M. and **Foulkes, Mary A. 1986.** "Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification," *Biometrics*, Volume 42, September, pages 507-516.

Lan, K.K.G. and DeMets, D.L. 1983. "Discrete sequential boundaries for clinical trials." *Biometrika*, 70, pages 659-663.

Lan, K.K.G. and Zucker, D.M. 1993. "Sequential monitoring of clinical trials: the role of information and Brownian motion." *Statistics in Medicine*, 12, pages 753-765.

Lance, G.N. and Williams, W.T. 1967. "A general theory of classificatory sorting strategies. I. Hierarchical systems." *Comput. J.* 9, pages 373-380.

Lance, G.N. and Williams, W.T. 1967. "Mixed-data classificatory programs I. Agglomerative systems." *Aust.Comput. J.* 1, pages 15-20.

Lawless, J.F. 1982. *Statistical Models and Methods for Lifetime Data*. John Wiley, New York.

Lawson, John. 1987. *Basic Industrial Experimental Design Strategies*. Center for Statistical Research at Brigham Young University. Provo, Utah. 84602. This is a manuscript used by Dr. Lawson in courses and workshops that he provides to industrial engineers. It is the basis for many of our experimental design procedures.

Lebart, Morineau, and Warwick. 1984. *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons. This book devotes a large percentage of its discussion to correspondence analysis.

Lee, E.T. 1974. "A Computer Program for Linear Logistic Regression Analysis" in *Computer Programs in Biomedicine*, Volume 4, pages 80-92.

Lee, E.T. 1980. *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications. Belmont, California.

Lee, E.T. 1992. *Statistical Methods for Survival Data Analysis*. Second Edition. John Wiley & Sons. New York. This book provides a very readable introduction to survival analysis techniques.

Lee, M.-L. T. 2004. *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers. Norwell, Massachusetts.

Lee, S. K. 1977. "On the Asymptotic Variances of u Terms in Loglinear Models of Multidimensional Contingency Tables." *Journal of the American Statistical Association*. Volume 72 (June, 1977), page 412. This article describes methods for computing standard errors that are used in the LLM section of this program.

Lenth, Russell V. 1987. "Algorithm AS 226: Computing Noncentral Beta Probabilities," *Applied Statistics*, Volume 36, pages 241-244.

Lenth, Russell V. 1989. "Algorithm AS 243: Cumulative Distribution Function of the Non-central t Distribution," *Applied Statistics*, Volume 38, pages 185-189.

Lesaffre, E. and Albert, A. 1989. "Multiple-group Logistic Regression Diagnostics" *Applied Statistics*, Volume 38, pages 425-440. See also Pregibon 1981.

Levene, H. 1960. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al., eds. Stanford University Press, Stanford Calif., pp. 278-292.

Lewis, J.A. 1999. "Statistical principles for clinical trials (ICH E9) an introductory note on an international guideline." *Statistics in Medicine*, 18, pages 1903-1942.

Lipsey, Mark W. 1990. *Design Sensitivity Statistical Power for Experimental Research*, Sage Publications, 2111 West Hillcrest Drive, Newbury Park, CA. 91320. This is an excellent introduction to power analysis.

Little, R. and Rubin, D. 1987. *Statistical Analysis with Missing Data*. John Wiley & Sons. New York. This book is completely devoted to dealing with missing values. It gives a complete treatment of using the EM algorithm to estimate the covariance matrix.

Little, R. C. et al. 2006. *SAS for Mixed Models – Second Edition*. SAS Institute Inc., Cary, North Carolina.

Liu, H. and Wu, T. 2005. "Sample Size Calculation and Power Analysis of Time-Averaged Difference," *Journal of Modern Applied Statistical Methods*, Vol. 4, No. 2, pages 434-445.

Lu, Y. and Bean, J.A. 1995. "On the sample size for one-sided equivalence of sensitivities based upon McNemar's test," *Statistics in Medicine*, Volume 14, pages 1831-1839.

Lui, J., Hsueh, H., Hsieh, E., and Chen, J.J. 2002. "Tests for equivalence or non-inferiority for paired binary data," *Statistics in Medicine*, Volume 21, pages 231-245.

Lloyd, D.K. and Lipow, M. 1991. *Reliability: Management, Methods, and Mathematics*. ASQC Quality Press. Milwaukee, Wisconsin.

Locke, C.S. 1984. "An exact confidence interval for untransformed data for the ratio of two formulation means," *J. Pharmacokinet. Biopharm.*, Volume 12, pages 649-655.

Lockhart, R. A. & Stephens, M. A. 1985. "Tests of fit for the von Mises distribution." *Biometrika* 72, pages 647-652.

M

Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. *Sample Size Tables for Clinical Studies, 2nd Edition*. Blackwell Science. Malden, Mass. A very good & easy to read book on determining appropriate sample sizes in many situations.

Makridakis, S. and Wheelwright, S.C. 1978. *Iterative Forecasting*. Holden-Day.: San Francisco, California. This is a very good book for the layman since it includes several detailed examples. It is written for a person with a minimum amount of mathematical background.

Manly, B.F.J. 1986. *Multivariate Statistical Methods - A Primer*. Chapman and Hall. New York. This nice little paperback provides a simplified introduction to many multivariate techniques, including MDS.

Mardia, K.V. and Jupp, P.E. 2000. *Directional Statistics*. John Wiley & Sons. New York.

Marple, S.L. 1987. *Digital Spectral Analysis with Applications.* Prentice-Hall: Englewood Cliffs, New Jersey. A technical book about spectral analysis.

Martinez and Iglewicz. 1981. "A test for departure from normality based on a biweight estimator of scale." *Biometrika*, 68, 331-333).

Marubini, E. and **Valsecchi, M.G.** 1996. *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley: New York, New York.

Mather, Paul. 1976. *Computational Methods of Multivariate Analysis in Physical Geography.* John Wiley & Sons. This is a great book for getting the details on several multivariate procedures. It was written for non-statisticians. It is especially useful in its presentation of cluster analysis. Unfortunately, it is out-of-print. You will have to look for it in a university library (it is worth the hunt).

Matsumoto, M. and Nishimura,T. 1998. "Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator" *ACM Trans. On Modeling and Computer Simulations*.

Mauchly, J.W. 1940. "Significance test for sphericity of a normal n-variate distribution." *Annals of Mathematical Statistics*, 11: 204-209

McCabe, G.P. 1984. "Principal variables." *Technometrics*, 26, 137-144.

McClish, D.K. 1989. "Analyzing a Portion of the ROC Curve." *Medical Decision Making*, 9: 190-195

McHenry, Claude. 1978. "Multivariate subset selection." *Journal of the Royal Statistical Society, Series C*. Volume 27, No. 23, pages 291-296.

McNeil, D.R. 1977. *Interactive Data Analysis*. John Wiley & Sons. New York.

Mendenhall, W. 1968. *Introduction to Linear Models and the Design and Analysis of Experiments*. Wadsworth. Belmont, Calif.

Metz, C.E. 1978. "Basic principles of ROC analysis." *Seminars in Nuclear Medicine*, Volume 8, No. 4, pages 283-298.

Miettinen, O.S. and Nurminen, M. 1985. "Comparative analysis of two rates." *Statistics in Medicine* 4: 213-226.

Milliken, G.A. and Johnson, D.E. 1984. *Analysis of Messy Data, Volume 1*. Van Nostrand Rienhold. New York, NY.

Milne, P. 1987. *Computer Graphics for Surveying.* E. & F. N. Spon, 29 West 35th St., NY, NY 10001

Montgomery, Douglas. 1984. *Design and Analysis of Experiments*. John Wiley & Sons, New York. A textbook covering a broad range of experimental design methods. The book is not limited to industrial investigations, but gives a much more general overview of experimental design methodology.

Montgomery, Douglas and Peck. 1992. *Introduction to Linear Regression Analysis*. A very good book on this topic.

Montgomery, Douglas C. 1991. *Introduction to Statistical Quality Control.* Second edition. John Wiley & Sons. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 700 pages.

Moore, D. S. and McCabe, G. P. 1999. *Introduction to the Practice of Statistics*. W. H. Freeman and Company. New York.

Mosteller, F. and Tukey, J.W. 1977. *Data Analysis and Regression*. Addison-Wesley. Menlo Park, California. This book should be read by all serious users of regression analysis. Although the terminology is a little different, this book will give you a fresh look at the whole subject.

Motulsky, Harvey. 1995. *Intuitive Biostatistics.* Oxford University Press. New York, New York. This is a wonderful book for those who want to understand the basic concepts of statistical testing. The author presents a very readable coverage of the most popular biostatistics tests. If you have forgotten how to interpret the various statistical tests, get this book!

Moura, Eduardo C. 1991. *How To Determine Sample Size And Estimate Failure Rate in Life Testing*. ASQC Quality Press. Milwaukee, Wisconsin.

Mueller, K. E., and Barton, C. N. 1989. "Approximate Power for Repeated-Measures ANOVA Lacking Sphericity." *Journal of the American Statistical Association,* Volume 84, No. 406, pages 549-555.

Mueller, K. E., LaVange, L.E., Ramey, S.L., and Ramey, C.T. 1992. "Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications." *Journal of the American Statistical Association,* Volume 87, No. 420, pages 1209-1226.

Mukerjee, H., Robertson, T., and Wright, F.T. 1987. "Comparison of Several Treatments With a Control Using Multiple Contrasts." *Journal of the American Statistical Association,* Volume 82, No. 399, pages 902-910.

Muller, K. E. and Stewart, P.W. 2006. *Linear Model Theory: Univariate, Multivariate, and Mixed Models*. John Wiley & Sons Inc. Hoboken, New Jersey.

Myers, R.H. 1990. *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company. Boston, Massachusetts. This is one of the bibles on the topic of regression analysis.

N

Naef, F. et al. 2002. "Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays," *Genome Biol.*, 3, RESEARCH0018.

Nam, Jun-mo. 1992. "Sample Size Determination for Case-Control Studies and the Comparison of Stratified and Unstratified Analyses," *Biometrics*, Volume 48, pages 389-395.

Nam, Jun-mo. 1997. "Establishing equivalence of two treatments and sample size requirements in matched-pairs design," *Biometrics*, Volume 53, pages 1422-1430.

Nam, J-m. and Blackwelder, W.C. 2002. "Analysis of the ratio of marginal probabilities in a matched-pair setting," *Statistics in Medicine*, Volume 21, pages 689-699.

Nash, J. C. 1987. *Nonlinear Parameter Estimation*. Marcel Dekker, Inc. New York, NY.

Nash, J.C. 1979. *Compact Numerical Methods for Computers*. John Wiley & Sons. New York, NY.

Nel, D.G. and van der Merwe, C.A. 1986. "A solution to the multivariate Behrens-Fisher problem." *Communications in Statistics—Series A, Theory and Methods*, 15, pages 3719-3735.

Nelson, W.B. 1982. *Applied Life Data Analysis*. John Wiley, New York.

Nelson, W.B. 1990. *Accelerated Testing*. John Wiley, New York.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. 1996. *Applied Linear Statistical Models*. Richard D. Irwin, Inc. Chicago, Illinois. This mammoth book covers regression analysis and analysis of variance thoroughly and in great detail. We recommend it.

Neter, J., Wasserman, W., and Kutner, M. 1983. *Applied Linear Regression Models*. Richard D. Irwin, Inc. Chicago, Illinois. This book provides you with a complete introduction to the methods of regression analysis. We suggest it to non-statisticians as a great reference tool.

Newcombe, Robert G. 1998a. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods." *Statistics in Medicine*, Volume 17, 857-872.

Newcombe, Robert G. 1998b. "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods." *Statistics in Medicine*, Volume 17, 873-890.

Newcombe, Robert G. 1998c. "Improved Confidence Intervals for the Difference Between Binomial Proportions Based on Paired Data." *Statistics in Medicine*, Volume 17, 2635-2650.

Newton, H.J. 1988. *TIMESLAB: A Time Series Analysis Laboratory*. Wadsworth & Brooks/Cole: Pacific Grove, California. This book is loaded with theoretical information about time series analysis. It includes software designed by Dr. Newton for performing advanced time series and spectral analysis. The book requires a strong math and statistical background.

O

O'Brien, P.C. and Fleming, T.R. 1979. "A multiple testing procedure for clinical trials." *Biometrics*, 35, pages 549-556.

O'Brien, R.G. and Kaiser, M.K. 1985. "MANOVA Method for Analyzing Repeated Measures Designs: An Extensive Primer." *Psychological Bulletin*, 97, pages 316-333.

Obuchowski, N. 1998. "Sample Size Calculations in Studies of Test Accuracy." *Statistical Methods in Medical Research*, 7, pages 371-392.

Obuchowski, N. and McClish, D. 1997. "Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal ROC Curve Indices." *Statistics in Medicine*, 16, pages 1529-1542.

Odeh, R.E. and Fox, M. 1991. *Sample Size Choice*. Marcel Dekker, Inc. New York, NY.

O'Neill and Wetherill. 1971 "The Present State of Multiple Comparison Methods," *The Journal of the Royal Statistical Society*, Series B, vol.33, 218-250).

Orloci, L. & Kenkel, N. 1985. *Introduction to Data Analysis*. International Co-operative Publishing House. Fairland, Maryland. This book was written for ecologists. It contains samples and BASIC programs of many statistical procedures. It has one brief chapter on MDS, and it includes a non-metric MDS algorithm.

Ostle, B. 1988. *Statistics in Research. Fourth Edition*. Iowa State Press. Ames, Iowa. A comprehension book on statistical methods.

Ott, L. 1977. *An Introduction to Statistical Methods and Data Analysis*. Wadsworth. Belmont, Calif. Use the second edition.

Ott, L. 1984. *An Introduction to Statistical Methods and Data Analysis, Second Edition*. Wadsworth. Belmont, Calif. This is a complete methods text. Regression analysis is the focus of five or six chapters. It stresses the interpretation of the statistics rather than the calculation, hence it provides a good companion to a statistical program like ours.

Owen, Donald B. 1956. "Tables for Computing Bivariate Normal Probabilities," *Annals of Mathematical Statistics*, Volume 27, pages 1075-1090.

Owen, Donald B. 1965. "A Special Case of a Bivariate Non-Central t-Distribution," *Biometrika*, Volume 52, pages 437-446.

P

Pandit, S.M. and Wu, S.M. 1983. *Time Series and System Analysis with Applications*. John Wiley and Sons. New York. This book provides an alternative to the Box-Jenkins approach for dealing with ARMA models. We used this approach in developing our automatic ARMA module.

Parmar, M.K.B. and Machin, D. 1995. *Survival Analysis*. John Wiley and Sons. New York.

Parmar, M.K.B., Torri, V., and Steart, L. 1998. "Extracting Summary Statistics to Perform Meta-Analyses of the Published Literature for Survival Endpoints." *Statistics in Medicine* 17, 2815-2834.

Pearson, K. 1901. "On lines and planes of closest fit to a system of points in space." *Philosophical Magazine* 2, 557-572.

Pan, Z. and Kupper, L. 1999. "Sample Size Determination for Multiple Comparison Studies Treating Confidence Interval Width as Random." *Statistics in Medicine* 18, 1475-1488.

Pedhazur, E.L. and Schmelkin, L.P. 1991. *Measurement, Design, and Analysis: An Integrated Approach.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. This mammoth book (over 800 pages) covers multivariate analysis, regression analysis, experimental design, analysis of variance, and much more. It provides annotated output from SPSS and SAS which is also useful to our users. The text is emphasizes the social sciences. It provides a "how-to," rather than a theoretical, discussion. Its chapters on factor analysis are especially informative.

Phillips, Kem F. 1990. "Power of the Two One-Sided Tests Procedure in Bioequivalence," *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 18, No. 2, pages 137-144.

Pocock, S.J. 1977. "Group sequential methods in the design and analysis of clinical trials." *Biometrika*, 64, pages 191-199.

Press, S. J. and Wilson, S. 1978. "Choosing Between Logistic Regression and Discriminant Analysis." *Journal of the American Statistical Association*, Volume 73, Number 364, Pages 699-705. This article details the reasons why logistic regression should be the preferred technique.

Press, William H. 1986. *Numerical Recipes*, Cambridge University Press, New York, New York.

Pregibon, Daryl. 1981. "Logistic Regression Diagnostics." *Annals of Statistics*, Volume 9, Pages 705-725. This article details the extensions of the usual regression diagnostics to the case of logistic regression. These results were extended to multiple-group logistic regression in Lesaffre and Albert (1989).

Price, K., Storn R., and Lampinen, J. 2005. *Differential Evolution – A Practical Approach to Global Optimization.* Springer. Berlin, Germany.

Prihoda, Tom. 1983. "Convenient Power Analysis For Complex Analysis of Variance Models." *Poster Session of the American Statistical Association Joint Statistical Meetings*, August 15-18, 1983, Toronto, Canada. Tom is currently at the University of Texas Health Science Center. This article includes FORTRAN code for performing power analysis.

R

Ramsey, Philip H. 1978 "Power Differences Between Pairwise Multiple Comparisons," *JASA*, vol. 73, no. 363, pages 479-485.

Rao, C.R. , Mitra, S.K., & Matthai, A. 1966. *Formulae and Tables for Statistical Work.* Statistical Publishing Society, Indian Statistical Institute, Calcutta, India.

Ratkowsky, David A. 1989. *Handbook of Nonlinear Regression Models*. Marcel Dekker. New York. A good, but technical, discussion of various nonlinear regression models.

Rawlings John O. 1988. *Applied Regression Analysis: A Research Tool*. Wadsworth. Belmont, California. This is a readable book on regression analysis. It provides a thorough discourse on the subject.

Reboussin, D.M., DeMets, D.L., Kim, K, and Lan, K.K.G. 1992. "Programs for computing group sequential boundaries using the Lan-DeMets Method." Technical Report 60, Department of Biostatistics, University of Wisconsin-Madison.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. John Wiley. New York, New York. This book provides a comprehensive mixture of theoretical and applied results in multivariate analysis. My evaluation may be biased since Al Rencher took me fishing when I was his student.

Robins, Greenland, and Breslow. 1986. "A General Estimator for the Variance of the Mantel-Haenszel Odds Ratio," *American Journal of Epidemiology*, vol.42, pages 719-723.

Robins, Breslow, and Greenland. 1986. "Estimators of the Mantel-Haenszel variance consisten in both sparse data and large-strata limiting models," *Biometrics*, vol. 42, pages 311-323.

Rosenbaum, P.R. 1989. "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, vol. 84, no. 408, pages 1024-1032.

Rosenbaum, P.R., and Rubin, D.B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, pages 41-55.

Rosenbaum, P.R., and Rubin, D.B. 1984. "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, vol. 79, pages 516-524.

Rosenbaum, P.R., and Rubin, D.B. 1985a. "Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *American Statistician*, vol. 39, pages 33-38.

Rosenbaum, P.R., and Rubin, D.B. 1985b. "The Bias Due to Incomplete Matching," *Biometrics*, vol. 41, pages 106-116.

Ryan, Thomas P. 1989. *Statistical Methods for Quality Improvement*. John Wiley & Sons. New York. This is a comprehensive treatment of SPC including control charts, process capability, and experimental design. It provides many rules-of-thumb and discusses many non-standard situations. This is a very good 'operators manual' type of book. 446 pages.

Ryan, Thomas P. 1997. *Modern Regression Methods*. John Wiley & Sons. New York. This is a comprehensive treatment of regression analysis. The author often deals with practical issues that are left out of other texts.

S

Sahai, Hardeo & Khurshid, Anwer. 1995. *Statistics in Epidemiology*. CRC Press. Boca Raton, Florida.

Schiffman, Reynolds, & Young. 1981. *Introduction to Multidimensional Scaling*. Academic Press. Orlando, Florida. This book goes through several examples.

Schilling, Edward. 1982. *Acceptance Sampling in Quality Control*. Marcel-Dekker. New York.

Schlesselman, Jim. 1981. *Case-Control Studies*. Oxford University Press. New York. This presents a complete overview of case-control studies. It was our primary source for the Mantel-Haenszel test.

Schmee and Hahn. November, 1979. "A Simple Method for Regression Analysis." *Technometrics*, Volume 21, Number 4, pages 417-432.

Schoenfeld, David A. 1983. "Sample-Size Formula for the Proportional-Hazards Regression Model" *Biometrics,* Volume 39, pages 499-503.

Schoenfeld, David A. and **Richter, Jane R.** 1982**.** "Nomograms for Calculating the Number of Patients Needed for a Clinical Trial with Survival as an Endpoint," *Biometrics,* March 1982, Volume 38, pages 163-170.

Schork, M. and Williams, G. 1980. "Number of Observations Required for the Comparison of Two Correlated Proportions." *Communications in Statistics-Simula. Computa.,* B9(4), 349-357.

Schuirmann, Donald. 1981**.** "On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval," *Biometrics,* Volume 37, pages 617.

Schuirmann, Donald. 1987**.** "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics,* Volume 15, Number 6, pages 657-680.

Seber, G.A.F. 1984. *Multivariate Observations*. John Wiley & Sons. New York. (This book is an encyclopedia of multivariate techniques. It emphasizes the mathematical details of each technique and provides a complete set of references. It will only be useful to those comfortable with reading mathematical equations based on matrices.)

Seber, G.A.F. and Wild, C.J. 1989. *Nonlinear Regression*. John Wiley & Sons. New York. This book is an encyclopedia of nonlinear regression.

Senn, Stephen. 1993. *Cross-over Trials in Clinical Research*. John Wiley & Sons. New York.

Senn, Stephen. 2002. *Cross-over Trials in Clinical Research.* Second Edition. John Wiley & Sons. New York.

Shapiro, S.S. and Wilk, M.B. 1965 "An analysis of Variance test for normality." *Biometrika*, Volume 52, pages 591-611.

Shuster, Jonathan J. 1990. *CRC Handbook of Sample Size Guidelines for Clinical Trials.* CRC Press, Boca Raton, Florida. This is an expensive book ($300) of tables for running log-rank tests. It is well documented, but at this price it better be.

Signorini, David. 1991. "Sample size for Poisson regression," *Biometrika,* Volume 78, 2, pages 446-450.

Simon, Richard. "Optimal Two-Stage Designs for Phase II Clinical Trials," *Controlled Clinical Trials,* 1989, Volume 10, pages 1-10.

Snedecor, G. and Cochran, Wm. 1972. *Statistical Methods*. The Iowa State University Press. Ames, Iowa.

Sorribas, A., March, J., and Trujillano, J. 2002. "A new parametric method based on S-distributions for computing receiver operating characteristic curves for continuous diagnostic tests." *Statistics in Medicine* 21, 1213-1235.

Spath, H. 1985. *Cluster Dissection and Analysis.* Halsted Press. New York. (This book contains a detailed discussion of clustering techniques for large data sets. It contains some heavy mathematical notation.)

Speed, T.P. (editor). 2003. *Statistical Analysis of Gene Expression Microarray Data.* Chapman & Hall/CRC. Boca Raton, Florida.

Stekel, D. 2003. *Microarray Bioinformatics.* Cambridge University Press. Cambridge, United Kingdom.

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., and Song, F. 2000. *Methods for Meta-Analysis in Medical Research.* John Wiley & Sons. New York.

Swets, John A. 1996. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics - Collected Papers.* Lawrence Erlbaum Associates. Mahway, New Jersey.

T

Tabachnick, B. and Fidell, L. 1989. *Using Multivariate Statistics.* Harper Collins. 10 East 53d Street, NY, NY 10022. This is an extremely useful text on multivariate techniques. It presents computer printouts and discussion from several popular programs. It provides checklists for each procedure as well as sample written reports. I strongly encourage you to obtain this book!

Tango, Toshiro. 1998. "Equivalence Test and Confidence Interval for the Difference in Proportions for the Paired-Sample Design." *Statistics in Medicine*, Volume 17, 891-908.

Therneau, T.M. and Grambsch, P.M. 2000. *Modeling Survival Data*. Springer: New York, New York. A the time of the writing of the Cox regression procedure, this book provides a thorough, up-to-date discussion of this procedure as well as many extensions to it. Recommended, especially to those with at least a masters in statistics.

Thomopoulos, N.T. 1980. *Applied Forecasting Methods*. Prentice-Hall: Englewood Cliffs, New Jersey. This book contains a very good presentation of the classical forecasting methods discussed in chapter two.

Thompson, Simon G. 1998. *Encyclopedia of Biostatistics, Volume 4*. John Wiley & Sons. New York. Article on Meta-Analysis on pages 2570-2579.

Tiku, M. L. 1965. "Laguerre Series Forms of Non-Central X^2 and F Distributions," *Biometrika*, Volume 42, pages 415-427.

Torgenson, W.S. 1952. "Multidimensional scaling. I. Theory and method." *Psychometrika* 17, 401-419. This is one of the first articles on MDS. There have been many advances, but this article presents many insights into the application of the technique. It describes the algorithm on which the metric solution used in this program is based.

Tubert-Bitter, P., Manfredi,R., Lellouch, J., Begaud, B. 2000. "Sample size calculations for risk equivalence testing in pharmacoepidemiology." *Journal of Clinical Epidemiology* 53, 1268-1274.

Tukey, J.W. and McLaughlin, D.H. 1963. "Less Vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization." *Sankhya, Series A* 25, 331-352.

Tukey, J.W. 1977. *Exploratory Data Analysis.* Addison-Wesley Publishing Company. Reading, Mass.

U

Upton, G.J.G. 1982."A Comparison of Alternative Tests for the 2 x 2 Comparative Trial.", *Journal of the Royal Statistical Society*, Series A,, Volume 145, pages 86-105.

Upton, G.J.G. and Fingleton, B. 1989. *Spatial Data Analysis by Example: Categorical and Directional Data. Volume 2.* John Wiley & Sons. New York.

V

Velicer, W.F. 1976. "Determining the number of components from the matrix of partial correlations." *Psychometrika*, 41, 321-327.

Velleman, Hoaglin. 1981. *ABC's of Exploratory Data Analysis.* Duxbury Press, Boston, Massachusetts.

Voit, E.O. 1992. "The S-distribution. A tool for approximation and classification of univariate, unimodal probability distributions." *Biometrical J.* 34, 855-878.

Voit, E.O. 2000. "A Maximum Likelihood Estimator for Shape Parameters of S-Distributions." *Biometrical J.* 42, 471-479.

Voit, E.O. and Schwacke, L. 1998. "Scalability properties of the S-distribution." *Biometrical J.* 40, 665-684.

Voit, E.O. and Yu, S. 1994. "The S-distribution. Approximation of discrete distributions." *Biometrical J.* 36, 205-219.

W

Walter, S.D., Eliasziw, M., and Donner, A. 1998. "Sample Size and Optimal Designs For Reliability Studies." *Statistics in Medicine*, 17, 101-110.

Welch, B.L. 1938. "The significance of the difference between two means when the population variances are unequal." *Biometrika*, 29, 350-362.

Welch, B.L. 1947. "The Generalization of "Student's" Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

Welch, B.L. 1949. "Further Note on Mrs. Aspin's Tables and on Certain Approximations to the Tabled Function," *Biometrika*, 36, 293-296.

Westfall, P. et al. 1999. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Institute Inc. Cary, North Carolina.

Westgard, J.O. 1981. "A Multi-Rule Shewhart Chart for Quality Control in Clinical Chemistry," *Clinical Chemistry*, Volume 27, No. 3, pages 493-501. (This paper is available online at the www.westgard.com).

Westlake, W.J. 1981. "Bioequivalence testing—a need to rethink," *Biometrics*, Volume 37, pages 591-593.

Whittemore, Alice. 1981. "Sample Size for Logistic Regression with Small Response Probability," *Journal of the American Statistical Association*, Volume 76, pages 27-32.

Wickens, T.D. 1989. *Multiway Contingency Tables Analysis for the Social Sciences.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. A thorough book on the subject. Discusses loglinear models in depth.

Wilson, E.B.. 1927. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, Volume 22, pages 209-212. This article discusses the 'score' method that has become popular when dealing with proportions.

Winer, B.J. 1991**.** *Statistical Principles in Experimental Design (Third Edition)*. McGraw-Hill. New York, NY. A very complete analysis of variance book.

Wit, E., and McClure, J. 2004. *Statistics for Microarrays*. John Wiley & Sons Ltd, Chichester, West Sussex, England.

Wolfinger, R., Tobias, R. and Sall, J. 1994. "Computing Gaussian likelihoods and their derivatives for general linear mixed models," *SIAM Journal of Scientific Computing*, 15, no.6, pages 1294-1310.

Woolson, R.F., Bean, J.A., and Rojas, P.B. 1986. "Sample Size for Case-Control Studies Using Cochran's Statistic," *Biometrics*, Volume 42, pages 927-932.

Y

Yuen, K.K. and Dixon, W. J. 1973. "The approximate behavior and performance of the two-sample trimmed t," *Biometrika*, Volume 60, pages 369-374.

Yuen, K.K. 1974. "The two-sample trimmed t for unequal population variances," *Biometrika*, Volume 61, pages 165-170.

Z

Zar, Jerrold H. 1984**.** *Biostatistical Analysis (Second Edition)*. Prentice-Hall. Englewood Cliffs, New Jersey. This introductory book presents a nice blend of theory, methods, and examples for a long list of topics of interest in biostatistical work.

Zhou, X., Obuchowski, N., McClish, D. 2002**.** *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc. New York, New York. This is a great book on the designing and analyzing diagnostic tests. It is especially useful for its presentation of ROC curves.

Chapter Index

3

3D Scatter Plots, I - 170
3D Surface Plots, I - 171

A

All Possible Regressions, III - 312
Analysis of Two-Level Designs, II - 213
Analysis of Variance
    Analysis of Two-Level Designs, II - 213
    Analysis of Variance for Balanced Data, II - 211
    General Linear Models (GLM), II - 212
    Mixed Models, II - 220
    One-Way Analysis of Variance, II - 210
    Repeated Measures Analysis of Variance, II - 214
Analysis of Variance for Balanced Data, II - 211
Appraisal Ratios, IV - 485
Area Under Curve, III - 390
ARIMA (Box-Jenkins), IV - 471
Attribute Charts, II - 251
Autocorrelations, IV - 472
Automatic ARMA, IV - 474
Axis-Line Selection Window, I - 184

B

Balanced Incomplete Block Designs, II - 262
Bar Charts, I - 141
Beta Distribution Fitting, V - 551
Binary Diagnostic Tests
    Clustered Samples, V - 538
    Paired Samples, V - 536
    Single Sample, V - 535
    Two Independent Samples, V - 537
Box-Jenkins Method, IV - 470
Box Plots, I - 152

C

Canonical Correlation, III - IV - 400
Circular Data Analysis, II - 230
Clustering
    Double Dendrograms, IV - 450
    Fuzzy Clustering, IV - 448
    Hierarchical Clustering / Dendrograms, IV - 445
    K-Means Clustering, IV - 446
    Medoid Partitioning, IV - 447
    Regression Clustering, IV - 449
Color Selection Window, I - 180
Comparables - Sales Price, IV - 486
Contour Plots, I - 172
Correlation Matrix, IV - 401
Correspondence Analysis, IV - 430
Cox Regression, V - 565
Creating / Loading a Database, I - 2
Cross Tabs on Summarized Data, I - 16
Cross Tabulation, V - 501
Cross-Correlations, IV - 473
Cross-Over Analysis Using T-Tests, II - 235
Cumulative Incidence, V - 560
Curve Fitting
    Area Under Curve, III - 390
    Curve Fitting - General, III - 351
    Growth and Other Models, III - 360
    Introduction to Curve Fitting, III - 350
    Nonlinear Regression, III - 315
    Piecewise Polynomial Models, III - 365
    Ratio of Polynomials Fit
        Many Variables, III - 376
        One Variable, III - 375
    Ratio of Polynomials Search
        Many Variables, III - 371
        One Variable, III - 370
    Sum of Functions Models, III - 380
    User-Written Models, III - 385
Curve Fitting - General, III - 351

D

Data Matching – Optimal and Greedy, I - 123
Data Report, I - 117
Data Screening, I - 118
Data Simulation, I - 15
Data Simulator, I - 122
Data Stratification, I - 124
Data Transformation, I - 3
Data Window, I - 7
Database Subsets, I - 14
Databases, I - 102
    Merging Two Databases, I - 104
Decomposition Forecasting, IV - 469
Dendrograms
    Double Dendrograms, IV - 450
    Hierarchical Clustering / Dendrograms, IV - 445
Descriptive Statistics, II - 200
Descriptive Tables, II - 201
Design of Experiments

Analysis of Two-Level Designs, II - 213
Balanced Incomplete Block Designs, II - 262
Design Generator, II - 268
D-Optimal Designs, II - 267
Fractional Factorial Designs, II - 261
Latin Square Designs, II - 263
Response Surface Designs, II - 264
Screening Designs, II - 265
Taguchi Designs, II - 266
Two-Level Designs, II - 260
Design Generator, II - 268
Diagnostic Tests
Binary
Clustered Samples, V - 538
Paired Samples, V - 536
Single Sample, V - 535
Two Independent Samples, V - 537
ROC Curves, V - 545
Discriminant Analysis, IV - 440
Distribution (Weibull) Fitting, V - 550
D-Optimal Designs, II - 267
Dot Plots, I - 150
Double Dendrograms, IV - 450

E

Equality of Covariance, IV - 402
Error-Bar Charts, I - 155
Exponential Smoothing - Horizontal, IV - 465
Exponential Smoothing - Trend, IV - 466
Exponential Smoothing - Trend / Seasonal, IV - 467
Exporting Data, I - 116

F

Factor Analysis, IV - 420
Filter, I - 121
Filters, I - 10
Forecasting / Time Series
ARIMA (Box-Jenkins), IV - 471
Autocorrelations, IV - 472
Automatic ARMA, IV - 474
Cross-Correlations, IV - 473
Decomposition Forecasting, IV - 469
Exponential Smoothing
Horizontal, IV - 465
Trend, IV - 466
Trend / Seasonal, IV - 467
Spectral Analysis, IV - 468
The Box-Jenkins Method, IV - 470
Theoretical ARMA, IV - 475
Fractional Factorial Designs, II - 261
Frequency Tables, V - 500
Function Plots, I - 160
Fuzzy Clustering, IV - 448

G

Gamma Distribution Fitting, V - 552
General Linear Models (GLM), II - 212
Graphics
Introduction to Graphics, I - 140
Settings Windows
Axis-Line, I - 184
Color, I - 180
Grid / Tick, I - 185
Heat Map, I - 187
Line, I - 183
Symbol, I - 181
Text, I - 182
Tick Label, I - 186
Single-Variable Charts
Bar Charts, I - 141
Histograms, I - 143
Pie Charts, I - 142
Probability Plots, I - 144
Three-Variable Charts
3D Scatter Plots, I - 170
3D Surface Plots, I - 171
Contour Plots, I - 172
Grid Plots, I - 173
Two-Variable Charts
Box Plots, I - 152
Dot Plots, I - 150
Error-Bar Charts, I - 155
Function Plots, I - 160
Histograms - Comparative, I - 151
Percentile Plots, I - 153
Scatter Plot Matrix, I - 162
Scatter Plot Matrix for Curve Fitting, I - 163
Scatter Plots, I - 161
Violin Plots, I - 154
Greedy Data Matching, I - 123
Grid Plots, I - 173
Grid / Tick Selection Window, I - 185
Growth and Other Models, III - 360

H

Heat Map Selection Window, I - 187
Hierarchical Clustering / Dendrograms, IV - 445
Histograms, I - 143
Histograms - Comparative, I - 151
Hotelling's One-Sample T2, IV - 405
Hotelling's Two-Sample T2, IV - 410
Hybrid Appraisal Models, IV - 487

I

If-Then Transformations, I - 120
Importing Data, I - 115

Importing Data, I - 12
Installation, I - 100
Installation and Basics, I - 1
Introduction
    Data
        Data Matching – Optimal and Greedy, I - 123
        Data Report, I - 117
        Data Screening, I - 118
        Data Simulator, I - 122
        Data Stratification, I - 124
        Exporting Data, I - 116
        Filter, I - 121
        If-Then Transformations, I - 120
        Importing Data, I - 115
        Merging Two Databases, I - 104
        Transformations, I - 119
    Essentials
        Databases, I - 102
        Installation, I - 100
        Macros, I - 130
        Merging Two Databases, I - 104
        Navigator, I - 107
        Output, I - 106
        Procedures, I - 105
        Spreadsheets, I - 103
        Tutorial, I - 101
Introduction to Curve Fitting, III - 350
Introduction to Graphics, I - 140
Item Analysis, V - 505
Item Response Analysis, V - 506

K

Kaplan-Meier Curves (Logrank Tests), V - 555
K-Means Clustering, IV - 446

L

Latin Square Designs, II - 263
Levey-Jennings Charts, II - 252
Life-Table Analysis, V - 570
Line Selection Window, I - 183
Linear Programming, IV - 480
Linear Regression and Correlation, III - 300
Logistic Regression, III - 320
Loglinear Models, V - 530
Logrank Tests, V - 555

M

Macros, I - 130
Mantel-Haenszel Test, V - 525
Mass Appraisal
    Appraisal Ratios, IV - 485
    Comparables - Sales Price, IV - 486
    Hybrid Appraisal Models, IV - 487
Matching – Optimal and Greedy, I - 123
Medoid Partitioning, IV - 447
Merging Two Databases, I - 104
Meta-Analysis
    Correlated Proportions, IV - 457
    Hazard Ratios, IV - 458
    Means, IV - 455
    Proportions, IV - 456
Mixed Models, II - 220
Multidimensional Scaling, IV - 435
Multiple Regression, III - 305
Multiple Regression with Serial Correlation
    Correction, III - 306
Multivariate Analysis
    Canonical Correlation, III - IV - 400
    Correlation Matrix, IV - 401
    Correspondence Analysis, IV - 430
    Discriminant Analysis, IV - 440
    Equality of Covariance, IV - 402
    Factor Analysis, IV - 420
    Hotelling's One-Sample T2, IV - 405
    Hotelling's Two-Sample T2, IV - 410
    Multidimensional Scaling, IV - 435
    Multivariate Analysis of Variance
        (MANOVA), IV - 415
    Principal Components Analysis, IV - 425
Multivariate Analysis of Variance
    (MANOVA), IV - 415

N

Navigator, I - 107
Nondetects Analysis, II - 240
Nondetects Regression, III - 345
Nonlinear Regression, III - 315

O

One Proportion, V - 510
One-Way Analysis of Variance, II - 210
Operations Research
    Linear Programming, IV - 480
Optimal Data Matching, I - 123
Output, I - 106
Output Window, I - 9

P

Parametric Survival (Weibull) Regression, V - 566
Pareto Charts, II - 253
Percentile Plots, I - 153
Pie Charts, I - 142
Piecewise Polynomial Models, III - 365
Poisson Regression, III - 325

Principal Components Analysis, IV - 425
Principal Components Regression, III - 340
Probability Calculator, I - 135
Probability Plots, I - 144
Probit Analysis, V - 575
Procedure Window, I - 8
Procedures, I - 105
Proportions
    Loglinear Models, V - 530
    Mantel-Haenszel Test, V - 525
    One Proportion, V - 510
    Two Correlated Proportions (McNemar), V - 520
    Two Independent Proportions, V - 515

Q

Quality Control
    Attribute Charts, II - 251
    Levey-Jennings Charts, II - 252
    Pareto Charts, II - 253
    R & R Study, II - 254
    Xbar R (Variables) Charts, II - 250
Quick Start & Self Help
    Creating / Loading a Database, I - 2
    Cross Tabs on Summarized Data, I - 16
    Data Simulation, I - 15
    Data Transformation, I - 3
    Data Window, I - 7
    Database Subsets, I - 14
    Filters, I - 10
    Importing Data, I - 12
    Installation and Basics, I - 1
    Output Window, I - 9
    Procedure Window, I - 8
    Running a Regression Analysis, I - 6
    Running a Two-Sample T-Test, I - 5
    Running Descriptive Statistics, I - 4
    Value Labels, I - 13
    Writing Transformations, I - 11

R

R & R Study, II - 254
Ratio of Polynomials Fit - Many Variables, III - 376
Ratio of Polynomials Fit - One Variable, III - 375
Ratio of Polynomials Search - Many Variables, III - 371
Ratio of Polynomials Search - One Variable, III - 370
Regression
    Cox Regression, V - 565
    Linear Regression and Correlation, III - 300
    Logistic Regression, III - 320
    Multiple Regression, III - 305
    Multiple Regression with Serial Correlation Correction, III - 306
    Nondetects Regression, III - 345
    Nonlinear Regression, III - 315
    Poisson Regression, III - 325
    Principal Components Regression, III - 340
    Response Surface Regression, III - 330
    Ridge Regression, III - 335
    Variable Selection
        Variable Selection for Multivariate Regression, III - 310
        Stepwise Regression, III - 311
        All Possible Regressions, III - 312
Regression Clustering, IV - 449
Reliability See *Survival*
Repeated Measures Analysis of Variance, II - 214
Response Surface Designs, II - 264
Response Surface Regression, III - 330
Ridge Regression, III - 335
ROC Curves, V - 545
Running a Regression Analysis, I - 6
Running a Two-Sample T-Test, I - 5
Running Descriptive Statistics, I - 4

S

Scatter Plot Matrix, I - 162
Scatter Plot Matrix for Curve Fitting, I - 163
Scatter Plots, I - 161
Screening Designs, II - 265
Settings Windows
    Axis-Line, I - 184
    Color, I - 180
    Grid / Tick, I - 185
    Heat Map, I - 187
    Line, I - 183
    Symbol, I - 181
    Text, I - 182
    Tick Label, I - 186
Spectral Analysis, IV - 468
Spreadsheets, I - 103
Stepwise Regression, III - 311
Stratification of Data, I - 124
Sum of Functions Models, III - 380
Survival / Reliability
    Beta Distribution Fitting, V - 551
    Cox Regression, V - 565
    Cumulative Incidence, V - 560
    Distribution (Weibull) Fitting, V - 550
    Gamma Distribution Fitting, V - 552
    Kaplan-Meier Curves (Logrank Tests), V - 555
    Life-Table Analysis, V - 570
    Parametric Survival (Weibull) Regression, V - 566
    Probit Analysis, V - 575
    Time Calculator, V - 580
    Tolerance Intervals, V - 585
Symbol Selection Window, I - 181

T

Tabulation
Cross Tabulation, V - 501
Frequency Tables, V - 500
Taguchi Designs, II - 266
Text Selection Window, I - 182
Theoretical ARMA, IV - 475
Tick Label Selection Window, I - 186
Time Calculator, V - 580
Time Series, See *Forecasting*
Tolerance Intervals, V - 585
Tools
Data Matching – Optimal and Greedy, I - 123
Data Simulator, I - 122
Data Stratification, I - 124
Macros, I - 130
Probability Calculator, I - 135
Transformations, I - 119
T-Tests
One-Sample or Paired, II - 205
Two-Sample, II - 206
Two-Sample (From Means and SD's), II - 207
Tutorial, I - 101
Two Correlated Proportions (McNemar), V - 520
Two Independent Proportions, V - 515
Two-Level Designs, II - 260

U

User-Written Models, III - 385

V

Value Labels, I - 13
Variable Selection for Multivariate
Regression III - 310
Violin Plots, I - 154

W

Writing Transformations, I - 11

X

Xbar R (Variables) Charts, II - 250

Index

2

2BY2 dataset, 320-62

3

3D scatter plot, 140-10, 170-1
  depth, 170-8
  elevation, 170-7
  perspective, 170-6
  projection method, 170-8
  rotation, 170-7
3D surface plot, 140-10, 171-1
  depth, 171-7
  elevation, 171-6
  perspective, 171-6
  projection method, 171-7
  rotation, 171-7

A

Ability data points
  item response analysis, 506-4
Abs transformation, 119-7
Absolute residuals
  multiple regression, 305-78
Accelerated testing
  parametric survival regression, 566-1
Access exporting, 116-1
Access importing, 115-1
Accuracy
  double-precision, 102-4
Accuracy, 101-2
Active colors, 180-3
Add output to log, 106-2
Adding a datasheet, 103-2
Additive constant, 585-4
  descriptive statistics, 200-5
  tolerance intervals, 585-4
Additive seasonality
  exponential smoothing, 467-1
Adjacent values
  box plot, 152-2
Adjusted average distance
  medoid partitioning, 447-13
Adjusted R-squared
  linear regression, 300-46
A-efficiency
  D-optimal designs, 267-13
AIC
  mixed models, 220-7
  Poisson regression, 325-24
Akaike information criterion
  mixed models, 220-7
  Poisson regression, 325-24
Algorithms
  hierarchical cluster analysis, 450-2
Alias
  two level designs, 260-2
  two-level designs, 213-7
All possible regressions, 312-1
Alone lambda
  discriminant analysis, 440-13
Alpha
  Cronbach's, 401-6, 505-2
  hierarchical clustering, 445-8
  multiple regression, 305-32
Alpha Four exporting, 116-1
Alpha level of C.I.'s
  linear regression, 300-26
Alpha of assumptions
  linear regression, 300-26
Alphas
  Cox regression, 565-9, 565-38
Amplitude
  spectral analysis, 468-1
Analysis of covariance
  GLM, 212-25
Analysis of two-level designs, 213-1
Analysis of variance, 211-2
  balanced data, 211-1
  GLM, 212-1
  linear regression, 300-46
  one-way, 210-1
  repeated measures, 214-1
ANCOVA
  GLM, 212-25
  mixed models, 220-1
  multiple regression, 305-86
ANCOVA dataset, 212-25, 305-86
ANCOVA example
  mixed models, 220-85
And
  if-then transformation, 120-2
Anderson and Hauck's test
  cross-over analysis using t-tests, 235-8
Anderson-Darling test
  descriptive statistics, 200-22
  linear regression, 300-49
Andrew's sine
  multiple regression, 305-26
Angular data, 230-1
ANOVA
  balanced data, 211-1
  multiple regression, 305-49
ANOVA balanced
  assumptions, 211-2
ANOVA detail report
  multiple regression, 305-50
Answer variable
  item response analysis, 506-2
Appraisal models
  hybrid, 487-1
Appraisal ratios, 485-1
Appraisal variables, 485-2
Appraisers
  R & R, 254-11
AR order (P)
  automatic ARMA, 474-8
Arc sine transformation, 119-17
Arc tangent transformation, 119-17
ArCosh transformation, 119-17
ArcSine-square root hazard
  Weibull fitting, 550-4
Area charts, 140-1, 141-1
Area under curve, 390-1
  ROC curves, 545-26
ARIMA
  automatic ARMA, 474-1
  Box-Jenkins, 470-1, 471-1
ARMA
  theoretical, 475-1
ARMA model
  Box Jenkins, 470-2
Armitage proportion trend test
  cross tabulation, 501-5
Armitage test
  cross tabulation, 501-16
ARSENIC dataset, 240-16
Arsine transformation, 119-17
ArSinh transformation, 119-17
ArTan transformation, 119-17
ArTanh transformation, 119-17
ASCII dataset, 12-1
ASCII delimited exporting, 116-1
ASCII files

importing fixed format, 115-3
ASCII fixed format exporting, 116-1
Aspin-Welch, 206-2
ASSESS dataset, 487-11
Assessment models
hybrid appraisal, 487-1
Assignable causes
presence of, 250-9
Association
partial and marginal, 530-5
Assumption tests
linear regression, 300-48
Assumptions
analysis of variance, 210-2
Kruskal-Wallis test, 210-2
linear regression, 300-3
multiple regression, 305-6
one-sample t-test, 205-2
one-way ANOVA, 210-28
t-test, 205-22
two-sample t-test, 206-18, 206-27
two-sample t-tests, 206-1
Asymmetric-binary variables
fuzzy clustering, 448-5
hierarchical clustering, 445-7
medoid partitioning, 447-2
Asymmetry
probability plots, 144-3
Attribute chart, 251-1
AUC, 390-1
ROC curves, 545-1, 545-6
AUC dataset, 390-2, 390-6
AUC1 dataset, 390-2
Autocorrelation, 472-1
multiple regression, 305-7
residuals, 305-53
type of, 300-56
Autocorrelation function
Box Jenkins, 470-1
Autocorrelation plot, 472-8
ARIMA, 471-12
automatic ARMA, 474-12
Automatic ARMA, 474-1
Autoregressive parameters
ARIMA, 471-3
theoretical ARMA, 475-1
Average absolute percent error
multiple regression, 305-45
Average difference plot
t-test, 205-20
Average distance
medoid partitioning, 447-13
Average silhouette
fuzzy clustering, 448-9
medoid partitioning, 447-13
Average squared loadings
canonical correlation, 400-4
Average transformation, 119-15
Axis-line settings window, 184-1

B

Backcasting
exponential smoothing, 465-2, 466-3, 467-3
Backward links
double dendrograms, 450-2
hierarchical clustering, 445-3
Balanced incomplete block designs, 262-1
Band
linear regression, 300-6
Bar charts, 140-1, 141-1
depth, 141-13
elevation, 141-13
gap between bars, 141-14
gap between sets of bars, 141-15
perspective, 141-12
projection method, 141-14
rotation, 141-13
Barnard's test of difference
two proportions, 515-14
Bartlett test
factor analysis, 420-14
principal components analysis, 425-17
T2, 410-10
Bartlett's test, 402-1
Baseline
area under curve, 390-3
Baseline cumulative survival
Cox regression, 565-38
Baseline survival
Cox regression, 565-8
Basic palette, 180-2
Basics, 1-1
BBALL dataset, 445-5, 445-12, 446-2, 446-6, 447-6, 447-12
BEAN dataset, 220-79, 220-82
Best model
all possible regressions, 312-4
Beta
hierarchical clustering, 445-8
BETA dataset, 551-2, 551-11
Beta distribution
probablility calculator, 135-1
simulation, 122-3
Beta distribution fitting, 551-1
Beta trace
PC regression, 340-14
BetaProb transformation, 119-8
BetaValue transformation, 119-8
Between subject
repeated measures, 214-2
Bias
R & R, 254-22
BIB designs, 262-1
Bimodal data
simulation, 122-23
Binary diagnostic tests
clustered samples, 538-1
paired samples, 536-1
two independent samples, 537-1
Binary response variables, 320-1
Binary test
1-sample binary diagnostic test, 535-1
BINCLUST dataset, 538-3, 538-7
Binomial distribution
probablility calculator, 135-2
simulation, 122-5
BinomProb transformation, 119-8
BinomValue transformation, 119-8
Binormal
ROC curves, 545-2
Bioequivalence
cross-over analysis using t-tests, 235-5
Bisquare weights
linear regression, 300-14
Bivariate normal distribution
probablility calculator, 135-2
Biweight
Weibull fitting, 550-17
Biweight estimator of scale, 200-22
Biweight kernel
Kaplan-Meier, 555-9
Weibull fitting, 550-35
Blackwelder test
correlated proportions, 520-5
Bleasdale-Nelder model
curve fitting, 351-5
growth curves, 360-3
Block size
balanced incomplete block designs, 262-3
fractional factorial designs, 261-2
Block variable
fractional factorial designs, 261-1
response surface designs, 264-2
Blocking
two level designs, 260-2
BMDP exporting, 116-1
BMT dataset, 555-43
Bonferroni
one-way ANOVA, 210-4
Bonferroni adjustment
mixed models, 220-14
Bonferroni C.I.'s
T2, 405-9, 410-9
Bootstrap
linear regression, 300-42
Bootstrap C.I. method
linear regression, 300-30
two proportions, 515-28
Bootstrap C.I.'s
multiple regression, 305-31
Bootstrap confidence coefficients
linear regression, 300-30
Bootstrap histograms

linear regression, 300-31, 300-44, 305-42
multiple regression, 305-75
Bootstrap percentile type
linear regression, 300-30
two proportions, 515-28
Bootstrap report
multiple regression, 305-74
Bootstrap retries
linear regression, 300-30
two proportions, 515-28
Bootstrap sample size
linear regression, 300-29
two proportions, 515-28
Bootstrap sampling method
linear regression, 300-30
Bootstrapping
curve fitting, 351-14
linear regression, 300-22
multiple regression, 305-21
t-test, 205-3
two-sample t-test, 206-3
Bootstrapping example
multiple regression, 305-72, 305-76
Box plot
adjacent values, 152-2
fences, 152-6
interquartile range, 152-1
whiskers, 152-5
Box plot style file, 152-13
Box plots, 140-5, 152-1
multiple comparisons, 152-2
Box's M, 214-1
Box's M test, 402-1, 402-7
Hotelling's T2, 410-2
repeated measures, 214-22
T2, 410-10
BOX320 dataset, 213-6
BOX402 dataset, 213-12
Box-Behnken designs, 264-1
Box-Jenkins
ARIMA, 471-1
automatic ARMA, 474-1
Box-Jenkins analysis, 470-1
Box-Pierce-Ljung statistic
automatic ARMA, 474-12
Box's M test
MANOVA, 415-5
BRAIN WEIGHT dataset, 2-2
Breslow ties
Cox regression, 565-6

C

C.I.method
multiple regression, 305-41
Calibration
linear regression, 300-6, 300-41
Caliper matching, 123-4
Caliper radius, 123-5
Candidate points
D-optimal designs, 267-14
Canonical correlation, 400-1
Canonical variate
MANOVA, 415-13
Capability analysis
Xbar R, 250-11
Capacities
Xbar R, 250-30
Carryover effect
cross-over analysis using t-tests, 235-3
Cascade, 106-5
Categorical IV's
Cox regression, 565-20
logistic regression, 320-20
multiple regression, 305-29
Poisson regression, 325-9
Categorical variables
multiple regression, 305-3, 305-87
Cauchy distribution
simulation, 122-5
Cbar
logistic regression, 320-15
C-chart, 251-2
Cell edit box, 103-10
Cell reference, 103-10
Censor variable
parametric survival regression, 566-4
Censored
Cox regression, 565-17
Kaplan-Meier, 555-15
Weibull fitting, 550-11
Censored regression, 566-1
Centering
Cox regression, 565-19
Central moments
descriptive statistics, 200-11
Central-composite designs, 264-1
Centroid
double dendrograms, 450-2
hierarchical clustering, 445-3
Charts
pareto, 253-1
variables, 250-1
Checklist
one sample tests, 205-21
one-way ANOVA, 210-26
two-sample tests, 206-25
Chen's method
two proportions, 515-20
Chi
loglinear models, 530-20
Chi-square
cross tabulation, 501-10
frequency tables, 500-11
Poisson regression, 325-26
Chi-square distribution
probablility calculator, 135-2
Chi-square test
cross tabulation, 501-1
two proportions, 515-6
Chi-square test example, 16-1
CHOWLIU73 dataset, 235-9, 235-15
Circular correlation, 230-12
Circular data analysis, 230-1
Circular histogram, 230-17
Circular histograms, 230-1
Circular statistics, 230-1
Circular uniform distribution, 230-3
CIRCULAR1 dataset, 230-22
Circularity
repeated measures, 214-3, 214-23
Clear, 103-5
Cluster analysis
double dendrograms, 450-1
K-means, 446-1
Cluster centers
K-means clustering, 446-1
Cluster cutoff
hierarchical clustering, 445-8
Cluster means
K-means clustering, 446-8
Cluster medoids section
fuzzy clustering, 448-9
medoid partitioning, 447-14
Cluster randomization
clustered binary diagnostic, 538-1
Cluster variables
K-means clustering, 446-3
Clustering
centroid, 445-7
complete linkage, 445-7
flexible strategy, 445-7
fuzzy, 448-1
group average, 445-7
hierarchical, 445-1
median, 445-7
medoid, 447-1
regression, 449-1
simple average, 445-7
single linkage, 445-7
Ward's minimum variance, 445-7
Cochran's Q test
meta analysis of hazard ratios, 458-4
meta-analysis of correlated proportions, 457-4
meta-analysis of means, 455-3
meta-analysis of proportions, 456-4
Cochran's test
two proportions, 515-7
Cochrane-Orcutt procedure, 306-1
COD
appraisal ratios, 485-8
descriptive statistics, 200-20
hybrid appraisal models, 487-17

Code cross-reference, 310-7
Coefficient alpha
  item analysis, 505-2
Coefficient of dispersion
  appraisal ratios, 485-8
  descriptive statistics, 200-18
  hybrid appraisal models, 487-17
Coefficient of variation
  descriptive statistics, 200-18
  linear regression, 300-38
  multiple regression, 305-45
Coefficients
  regression, 305-47
  stepwise regression, 311-8
Collate transformation, 119-12
COLLETT157 dataset, 565-55
COLLETT266 dataset, 320-73
COLLETT5 dataset, 555-42
Collinearity
  MANOVA, 415-5
Color
  mixer, 180-2
  model, 180-2
  wheel, 180-3
Color selection window, 180-1
Column widths, 103-15
Communality
  factor analysis, 420-3, 420-12, 420-16
  principal components analysis, 425-16
Communality iterations
  factor analysis, 420-8
Comparables
  sales price, 486-1
COMPARABLES dataset, 486-10
Competing risks
  cumulative incidence, 560-1
Complete linkage
  double dendrograms, 450-2
  hierarchical clustering, 445-3
Compound symmetry
  repeated measures, 214-3
CONCENTRATION dataset, 240-21
Concordance
  Kendall's coefficient, 211-15
Condition number
  multiple regression, 305-58
  PC regression, 340-13
  ridge regression, 335-17
Conditional tests
  two proportions, 515-5
Confidence band
  linear regression, 300-6, 300-33, 300-60
Confidence coefficient
  multiple regression, 305-32
  T2, 410-5
Confidence interval
  descriptive statistics, 200-13
  multiple regression, 305-14
  Poisson regression, 325-26
Confidence intervals
  Cox regression, 565-11
  curve fitting, 350-4
  linear regression, 300-6
  T2, 405-9, 410-9
  two proportions, 515-18
Confidence intervals of odds ratio
  two proportions, 515-23
Confidence intervals of ratio
  two proportions, 515-21
Confidence limits, 200-2
  linear regression, 300-33
  Nelson-Aalen hazard, 550-4
Confounding
  two level designs, 260-2
Confounding size, 213-3
Constant distribution
  simulation, 122-6
Constraint section
  linear programming, 480-5
Constraints
  linear programming, 480-1
Contains transformation, 119-17
Contaminated normal simulation, 122-21
Continuity correction
  two proportions, 515-7
Contour plots, 140-11, 172-1
  response surface regression, 330-19
Contrast type
  multiple regression, 305-29
  Poisson regression, 325-9
Contrast variables
  multiple regression, 305-4
Control charts
  attribute, 251-1
  formulas, 250-5
  Xbar R, 250-1
Control limits
  Xbar R, 250-2
Cook's D
  linear regression, 300-20, 300-62, 300-63, 300-65, 300-66
  multiple regression, 305-20, 305-64
Cook's distance
  logistic regression, 320-15
Cophenetic correlation
  hierarchical clustering, 445-14
Cophenetic correlation coefficient, 445-4
Copy, 103-4
Copy output, 106-3
Copying data, 7-2
COR
  correspondence analysis, 430-14
Correlation, 300-1
  canonical, 400-1
  confidence limits, 300-12
  cross, 473-1
  linear regression, 300-2, 300-11, 300-45
  Pearson, 300-45
  Spearman, 300-45
  Spearman rank, 401-1
  Spearman's rank, 300-12
Correlation coefficient
  linear regression, 300-9
Correlation coefficient distribution
  probablility calculator, 135-3
Correlation matrices
  factor analysis, 420-5
  principal components analysis, 425-8
Correlation matrix, 401-1
Correlation matrix report
  multiple regression, 305-46
Correlations
  medoid partitioning, 447-10
  partial, 401-3
  principal components analysis, 425-17
Correlogram
  autocorrelation, 472-1
CORRES1 dataset, 430-6, 430-10, 430-16
Correspondence analysis, 430-1
  eigenvalues, 430-12
CorrProb transformation, 119-8
CorrValue transformation, 119-8
Cos transformation, 119-17
Cosh transformation, 119-17
Cosine transformation, 119-17
Cost benefit analysis
  ROC curves, 545-22
Count tables, 500-1
Count transformation, 119-15
Covariance
  analysis of, 212-25
  multiple regression, 305-86
Covariance matrices, 402-1
Covariance matrix
  repeated measures, 214-3
Covariance pattern models
  mixed models, 220-5
Covariates
  GLM, 212-3
  mixed models, 220-9
  response surface regression, 330-5
CovRatio
  linear regression, 300-21, 300-63
  multiple regression, 305-20, 305-64
Cox model
  Cox regression, 565-1
Cox proportional hazards regression model, 565-1
Cox regression, 565-1
Cox test

circular data, 230-9
Cox-Mantel logrank test
Kaplan-Meier, 555-41
COXREG dataset, 565-51
COXSNELL dataset, 123-23
Cox-Snell residual
parametric survival regression, 566-19
Cox-Snell residuals
Cox regression, 565-13, 565-39
nondetects regression, 345-13
Cp
all possible regressions, 312-8
multiple regression, 305-55
Xbar R, 250-12
Cp variable plot
all possible regressions, 312-10
Cpk
Xbar R, 250-12, 250-31
Cramer's V
cross tabulation, 501-14
Creating a database, 2-1
Creating a new database
tutorial, 101-2
Creating data
simulation, 122-1
Cronbach's alpha
item analysis, 505-2, 505-6
Cronbachs alpha
correlation matrix, 401-6
CROSS dataset, 220-101
Cross tabulation, 501-1
summarized data, 16-1
Cross-correlations, 473-1
Crossed factors
design generator, 268-1
Crossover analysis, 220-1
Cross-over analysis using t-tests, 235-1
Crossover data example
mixed models, 220-101
Crosstabs, 501-1
CsProb transformation, 119-9
CsValue transformation, 119-9
CTR
correspondence analysis, 430-14
Cubic fit
curve fitting, 351-2
Cubic terms
response surface regression, 330-7
Cum transformation, 119-7
Cumulative hazard
Cox regression, 565-2
Cumulative hazard function
Kaplan-Meier, 555-2
Weibull fitting, 550-2
Cumulative incidence analysis, 560-1
Cumulative survival
Cox regression, 565-2
Curve equivalence
curve fitting, 351-16
Curve fitting, 351-1
introduction, 350-1
Curve inequality test
curve fitting, 351-32
Custom model
Cox regression, 565-26
multiple regression, 305-34
CUSUM chart, 250-4, 250-8
CUSUM Charts, 250-37
Cut, 103-4
Cut output, 106-3
Cycle-input variable
decomposition forecasting, 469-5

D

D'Agostino kurtosis
descriptive statistics, 200-24
D'Agostino kurtosis test
linear regression, 300-49
D'Agostino omnibus
descriptive statistics, 200-25
D'Agostino omnibus test
linear regression, 300-49
D'Agostino skewness
descriptive statistics, 200-23
D'Agostino skewness test
linear regression, 300-49
DAT exporting, 116-1
Data
entering, 2-1
estimating missing, 118-1
importing, 12-1
numeric, 102-1
printing, 2-7, 103-3, 117-1
saving, 2-6
simulation, 15-1
simulation of, 122-1
text, 102-1
Data features, 200-1
Data imputation, 118-1
Data matching
caliper, 123-4
caliper radius, 123-5
distance calculation method, 123-3
forced match variable, 123-4
full (variable), 123-3
greedy, 123-1, 123-2
optimal, 123-1, 123-2
propensity score, 123-2
standardized difference, 123-15
Data orientation
bar charts, 141-2
Data report, 103-6, 117-1
Data screening
T2 alpha, 118-3
Data screening, 118-1
Data screening, 200-3
Data simulator, 122-1
Data stratification, 124-1
Data transformation, 3-1
Data type, 102-10
Data window, 1-4, 7-1
Database, 102-1
clearing, 2-9
creating, 2-1, 101-2
Excel compatible, 102-1
exporting, 115-1, 116-1
introduction, 101-1
limits, 102-1
loading, 2-1, 2-10, 7-1
opening, 101-3
printing, 2-7
S0, 102-1
s0 and s1 files, 2-6
S0-type, 2-9
S0Z (zipped), 102-1
S0Z-type, 2-9
saving, 101-2
size, 102-1
sorting, 103-6
subsets, 14-1
Database/spreadsheet comparison, 102-4
Databases
merging two, 104-1
Dataset
2BY2, 320-62
ANCOVA, 212-25, 305-86
ARSENIC, 240-16
ASCII, 12-1
ASSESS, 487-11
AUC, 390-2, 390-6
AUC1, 390-2
BBALL, 445-5, 445-12, 446-2, 446-6, 447-6, 447-12
BEAN, 220-79, 220-82
BETA, 551-2, 551-11
BINCLUST, 538-3, 538-7
BMT, 555-43
BOX320, 213-6
BOX402, 213-12
BRAIN WEIGHT, 2-2
CHOWLIU73, 235-9, 235-15
CIRCULAR1, 230-22
COLLETT157, 565-55
COLLETT266, 320-73
COLLETT5, 555-42
COMPARABLES, 486-10
CONCENTRATION, 240-21
CORRES1, 430-6, 430-10, 430-16
COXREG, 565-51
COXSNELL, 123-23
CROSS, 220-101
DCP, 345-2, 345-9
DIOXIN, 240-2, 240-11

DOPT_MIXED, 267-22
DOPT3, 267-20
DRUGSTUDY, 501-19
DS476, 315-2, 315-9, 385-2, 385-9
EXAMS, 450-12
EXERCISE, 214-6, 214-16
FANFAILURE, 550-49
FISH, 220-90
FISHER, 143-14, 144-15, 150-8, 151-13, 152-12, 153-8, 154-8, 170-2, 170-9, 173-7, 402-2, 402-5, 440-4, 440-10, 440-20, 440-22
FNREG1, 360-15, 380-7
FNREG2, 365-11
FNREG3, 163-4, 370-6, 375-8
FNREG4, 371-6, 376-8
FNREG5, 351-30
FRUIT, 141-1, 141-17
FUZZY, 448-3, 448-8
HAIR, 220-103
HEART, 212-23
HOUSING, 306-4, 306-10
INTEL, 465-7, 466-9, 471-7, 473-5
IQ, 305-27, 305-43, 305-72, 305-76, 305-79
ITEM, 505-2, 505-5, 506-2, 506-6
KLEIN6, 555-45
KOCH36, 325-7, 325-21
LACHIN91, 320-71
LATINSQR, 212-22
LEAD, 240-19
LEE91, 570-4, 570-15
LEUKEMIA, 320-18, 320-34, 320-57
LINREG1, 300-24, 300-37
LOGLIN1, 530-7, 530-11
LP, 480-2, 480-4
LUNGCANCER, 565-15, 565-31, 565-48
MAMMALS, 3-1, 4-1, 10-1
MAMMALS1, 5-1, 6-1
MANOVA1, 410-3, 410-6, 415-5, 415-10
MARUBINI, 560-3, 560-9
MDS2, 435-6, 435-10
MDS2, 435-15
METACPROP, 457-6, 457-14
METAHR, 458-6, 458-12
MLCO2, 470-11
MOTORS, 566-3, 566-11
NC CRIMINAL, 320-64, 320-68
NONDETECTS, 240-4
ODOR, 330-3, 330-11
PAIN, 220-51
PCA2, 420-5, 420-11, 425-9, 425-15
PCA2, 118-4
PET, 538-11
PIE, 142-6
PLANT, 212-27
POISREG, 325-37
POLITIC, 13-1, 14-1
PREPOST, 305-87
PROPENSITY, 123-5, 123-12, 124-4
QATEST, 250-14, 250-27, 250-33, 250-35, 250-37, 251-3, 251-11, 253-2, 253-7, 253-9
RCBD, 220-94
REACTION, 214-29
REACTION, 214-6
READOUT105, 550-47
REGCLUS, 449-2, 449-5
RESALE, 117-4, 151-14, 155-1, 155-7, 201-1, 201-11, 201-12, 201-14, 201-15, 201-17, 201-19, 201-21, 305-81, 500-1, 500-9, 500-10, 500-12, 500-14, 501-1, 501-8, 501-11, 501-17
RIDGEREG, 335-7, 335-15, 340-3, 340-11
RMSF, 545-3
RNDBLOCK, 211-4, 211-11, 212-3, 212-12
ROC, 545-19
RRSTUDY, 254-1, 254-10
RRSTUDY1, 254-24
SALES, 467-9, 469-9
SALESRATIO, 485-1, 485-6, 486-4
SAMPLE, 101-3, 161-20, 162-5, 171-9, 172-7, 200-4, 200-10, 205-12, 206-12, 210-16, 310-3, 310-6, 311-3, 311-6, 312-2, 312-6, 400-8, 401-2, 401-5, 585-8
SERIESA, 470-8, 474-7
SMOKING, 525-2, 525-5
SUNSPOT, 468-9, 472-7
SURVIVAL, 555-14, 555-37, 575-1, 575-5
SUTTON 22, 456-6, 456-14
SUTTON30, 455-6, 455-13
T2, 405-3, 405-5, 405-10
TIMECALC, 580-3
TUTOR, 220-98
TWOSAMPLE, 220-69, 220-72
TWOSAMPLE2, 220-70, 220-73
TWOSAMPLECOV, 220-76
WEIBULL, 550-12, 550-27, 550-44, 552-3, 552-12, 555-27
WEIBULL2, 144-17
WEIGHTLOSS, 220-85
WESTGARD, 252-9
ZHOU 175, 545-33
ZINC, 345-15
Datasheet, 101-1
Datasheets, 102-1
Date formats, 102-8
Date function transformations, 119-6
Day format, 102-8
Day transformation, 119-6
DB, 115-1
Dbase importing, 115-1
DBF exporting, 116-1
DBF importing, 115-1
DCP dataset, 345-2, 345-9
Death density
  life-table analysis, 570-3
Decision variables
  linear programming, 480-1
Decomposition forecasting, 469-1
Default template, 105-1
Defects/defectives variable, 251-4
D-efficiency
  D-optimal designs, 267-12
Degrees of freedom
  factor analysis, 420-14
  two-sample t-test, 206-13
Delta
  cluster goodness-of-fit, 445-4
  loglinear models, 530-8
  Mantel-Haenszel test, 525-4
Dendrogram
  hierarchical clustering, 445-15
Dendrograms, 445-1
  double, 450-1, 450-3
Density trace
  histograms, 143-1
  histograms – comparative, 151-2
  violin plot, 154-1
Dependent variable
  linear regression, 300-25
  multiple regression, 305-1
  Poisson regression, 325-8
Depth
  3D scatter plot, 170-8
  3D surface plot, 171-7
  bar charts, 141-13
Derivatives
  Weibull fitting, 550-16
Descriptive statistics, 4-1, 200-1
  additive constant, 200-5
  Anderson-Darling test, 200-22
  central moments, 200-11
  COD, 200-20
  coefficient of dispersion, 200-18
  coefficient of variation, 200-18
  confidence interval, 200-13
  D'Agostino kurtosis, 200-24
  D'Agostino omnibus, 200-25
  D'Agostino skewness, 200-23
  dispersion, 200-16
  EDF, 200-7
  Fisher's g1, 200-18
  Fisher's g2, 200-18
  geometric mean, 200-14
  harmonic mean, 200-14

Histogram, 200-25
interquartile range, 200-17
IQR, 200-17
Kolmogorov-Smirnov, 200-23
kurtosis, 200-18
Lilliefors' critical values, 200-23
MAD, 200-20
Martinez-Iglewicz, 200-22
mean, 200-13
mean absolute deviation, 200-20
mean deviation, 200-20
mean-deviation, 200-20
median, 200-14
mode, 200-15
moment, 200-11
Normal probability plot, 200-26
normality, 200-21
normality tests, 200-21
percentile type, 200-6
Probability plot, 200-26
quartiles, 200-21
range, 200-17
Shapiro-Wilk test, 200-22
skewness, 200-17
Skewness test, 200-24
standard deviation, 200-16
standard error, 200-13
Stem-leaf plot, 200-27
trim-mean, 200-19
trimmed, 200-19
trim-std dev, 200-19
unbiased Std Dev, 200-17
variance, 200-15
Descriptive statistics report
multiple regression, 305-45
Descriptive tables, 201-1
Design generator, 268-1
Designs
analysis of, 213-1
Box-Behnken, 264-1
central-composite, 264-1
design generator, 268-1
factorial, 260-3
fractional factorial, 261-1
Plackett-Burman, 265-1
response surface, 264-1
screening, 265-1
Taguchi, 266-1
two-level factorial, 260-1, 268-1
Determinant
D-optimal designs, 267-13
Determinant analysis
D-optimal designs, 267-11
Deviance
Cox regression, 565-10
logistic regression, 320-8
Poisson regression, 325-4, 325-5
Deviance residuals
Cox regression, 565-14, 565-40
logistic regression, 320-13
Poisson regression, 325-31
Deviance test
Poisson regression, 325-3
DFBETA
logistic regression, 320-14
DFBETAS
linear regression, 300-21, 300-63
multiple regression, 305-20, 305-65
DFCHI2
logistic regression, 320-15
DFDEV
logistic regression, 320-15
Dffits
linear regression, 300-63
DFFITS
linear regression, 300-20
multiple regression, 305-19, 305-64
Diagnostic test
1-sample binary diagnostic test, 535-1
2-sample binary diagnostic, 537-1
paired binary diagnostic, 536-1
DIF exporting, 116-1
Differencing
ARIMA, 471-2
autocorrelation, 472-2
Box Jenkins, 470-7
spectral analysis, 468-4
Differential evolution
hybrid appraisal models, 487-2
Weibull fitting, 550-11
Digamma
beta distribution fitting, 551-12
Dimensions
multidimensional scaling, 435-4
DIOXIN dataset, 240-2, 240-11
Directional test
meta analysis of hazard ratios, 458-3
meta-analysis of correlated proportions, 457-4
meta-analysis of proportions, 456-4
Disabling the filter, 121-4
Discriminant analysis, 440-1
logistic regression, 320-1
Discrimination parameter
item response analysis, 506-8
Dispersion
descriptive statistics, 200-16
Dissimilarities
medoid partitioning, 447-1
multidimensional scaling, 435-4
Distance
multidimensional scaling, 435-2
Distance calculation
medoid partitioning, 447-2
Distance calculation method
data matching, 123-3
Distance method
fuzzy clustering, 448-5
hierarchical clustering, 445-8
Distances
medoid partitioning, 447-10
Distinct categories
R & R, 254-3, 254-19
Distribution
circular uniform, 230-3
Von Mises, 230-5
Distribution fitting
Weibull fitting, 550-1
Distribution statistics, 200-1
Distributions
combining, 122-13
exponential, 550-1
extreme value, 550-1
logistic, 550-1
log-logistic, 550-1
lognormal, 550-1
mixing, 122-13
simulation, 122-1
Weibull, 550-1
Dmn-criterion value, 206-23
DOPT_MIXED dataset, 267-22
DOPT3 dataset, 267-20
D-optimal designs, 267-1
Dose
probit analysis, 575-1
Dose-response plot
probit analysis, 575-9
Dot plots, 140-4, 150-1
jittering, 150-1
Double dendrograms, 450-1
Double exponential smoothing, 466-1
Double-precision accuracy, 101-2, 102-4
DRUGSTUDY dataset, 501-19
DS476 dataset, 315-2, 315-9, 385-2, 385-9
Dummy variables
multiple regression, 305-3
Duncan’s test
one-way ANOVA, 210-5
Dunn’s partition coefficient
fuzzy clustering, 448-2
Dunn’s test
one-way ANOVA, 210-7
Dunnett’s test
one-way ANOVA, 210-6
Duplicates
D-optimal designs, 267-5
Durbin-Watson
linear regression, 300-17
multiple regression, 305-17
Durbin-Watson test
multiple regression, 305-53
multiple regression with serial correlation, 306-3

E

e - using
  Cox regression, 565-4
E notation, 102-4
EDF
  descriptive statistics, 200-7
EDF plot, 240-15
Edit
  clear, 103-5
  copy, 103-4
  cut, 103-4
  delete, 103-5
  fill, 103-6
  find, 103-6
  insert, 103-5
  paste, 103-4
  undo, 103-4
Efron ties
  Cox regression, 565-7
Eigenvalue
  MANOVA, 415-14
  PC regression, 340-13
Eigenvalues, 425-17
  correspondence analysis, 430-12
  factor analysis, 420-14
  multidimensional scaling, 435-11
  multiple regression, 305-58, 305-59
  principal components analysis, 425-12
  ridge regression, 335-17
Eigenvector
  multiple regression, 305-58, 305-60
Eigenvectors
  factor analysis, 420-15
Elapsed time
  time calculator, 580-1
Elevation
  3D scatter plot, 170-7
  3D surface plot, 171-6
  bar charts, 141-13
Ellipse (probability)
  linear regression, 300-8
Else
  if-then transformation, 120-4
EM algorithm
  principal components analysis, 425-5
Empirical
  ROC curves, 545-2
Empty cells, 102-5
Entry date
  time calculator, 580-2
Entry time
  Cox regression, 565-17
  Kaplan-Meier, 555-15
Epanechnikov
  Weibull fitting, 550-17
Epanechnikov kernel
  Kaplan-Meier, 555-8
  Weibull fitting, 550-34
Epsilon
  Geisser-Greenhouse, 214-4
  repeated measures, 214-20
Equal slopes
  multiple regression, 305-86
Equality of covariance matrices, 402-1
Equivalence
  2-sample binary diagnostic, 537-9
  clustered binary diagnostic, 538-8
  cross-over analysis using t-tests, 235-1
  paired binary diagnostic, 536-7
  ROC curves, 545-30
Equivalence test
  correlated proportions, 520-8
  two proportions, 515-17
  two-sample, 207-1
Equivalence tests
  two proportions, 515-38
Error-bar charts, 140-6, 155-1
Euclidean distance
  medoid partitioning, 447-2
Event date
  time calculator, 580-2
EWMA chart, 250-4, 250-35
EWMA chart limits, 250-8
EWMA parameter, 250-19
Exact test
  two proportions, 515-12
Exact tests
  two proportions, 515-4, 515-36
EXAMS dataset, 450-12
Excel exporting, 116-1
EXERCISE dataset, 214-6, 214-16
Exiting NCSS, 101-4
Exp transformation, 119-7
Experiment (Run)
  two level designs, 260-2
Experimental design, 260-1
  two level designs, 260-2
Experimental error
  two level designs, 260-2
Experimentwise error rate, 210-3
Exponential
  curve fitting, 351-10
  using, 565-4
Exponential distribution
  simulation, 122-6
  Weibull fitting, 550-8
Exponential model
  curve fitting, 351-6
  growth curves, 360-4
Exponential regression, 566-1
Exponential smoothing
  double, 466-1
  horizontal, 465-1
  simple, 465-1
  trend, 466-1
  trend and seasonal, 467-1
ExpoProb transformation, 119-9
Export, 103-3
Export limitations, 116-1
Exporting data, 116-1
Exposure
  Poisson regression, 325-1
Exposure variable
  Poisson regression, 325-12
ExpoValue transformation, 119-9
Extract transformation, 119-18
Extreme value distribution
  Weibull fitting, 550-8

F

F distribution
  probablility calculator, 135-3
  simulation, 122-7
Factor analysis, 420-1
Factor loadings
  factor analysis, 420-16
  principal components analysis, 425-2
Factor rotation
  factor analysis, 420-7
Factor scaling
  D-optimal designs, 267-2
Factorial designs
  two level designs, 260-3
  two-level designs, 260-1
Factors
  how many, 420-3, 425-6
Failed
  parametric survival regression, 566-2
  Weibull fitting, 550-11
Failure
  Cox regression, 565-16
  Kaplan-Meier, 555-15
Failure distribution
  Weibull fitting, 550-37
Familywise error rate, 210-3
FANFAILURE dataset, 550-49
Farazdaghi and Harris model
  curve fitting, 351-5
  growth curves, 360-3
Farrington-Manning test
  two proportions, 515-10
Fast Fourier transform
  spectral analysis, 468-3
Fast initial restart, 250-9
Feedback model, 487-1
Fences
  box plot, 152-6
File function transformation, 119-15
Files

Access, 115-1
ASCII, 115-3
BMDP, 115-1
creating text, 115-1
Dbase, 115-1
Excel, 115-1
NCSS 5.0, 115-1
Paradox, 115-1
SAS, 115-1
SPSS, 115-1
text, 115-1
Fill, 103-6
Fill functions transformations, 119-6
Filter, 121-1
disabling, 10-4
specifying, 103-7
Filter statements, 103-7
Filters, 10-1
Final Tableau section
linear programming, 480-6
Find, 103-6
Find a procedure, 107-1
Find in output, 106-4
Find next in output, 106-4
FIR, 250-9
FISH dataset, 220-90
FISHER dataset, 143-14, 144-15, 150-8, 151-13, 152-12, 153-8, 154-8, 170-2, 170-9, 173-7, 402-2, 402-5, 440-4, 440-10, 440-20, 440-22
Fisher information matrix
beta distribution fitting, 551-14
gamma distribution fitting, 552-15
Weibull fitting, 550-32
Fisher's exact test, 501-1, 501-13
cross tabulation, 501-17
Fisher's Z transformation
linear regression, 300-11
Fisher's exact test
cross tabulation, 501-11
Fisher's g1
descriptive statistics, 200-18
Fisher's g2
descriptive statistics, 200-18
Fisher's LSD
one-way ANOVA, 210-6
Fixed effects
mixed models, 220-9
Fixed effects model
meta-analysis of correlated proportions, 457-5
meta-analysis of hazard ratios, 458-4
meta-analysis of means, 455-4
meta-analysis of proportions, 456-5
Fixed effects models
mixed models, 220-4
Fixed factor
ANOVA balanced, 211-5
GLM, 212-4
repeated measures, 214-8
Fixed sigma
Xbar R, 250-19
Fixed Xbar
Xbar R, 250-18
Fleiss Confidence intervals
two proportions, 515-24
Fleming-Harrington tests
Kaplan-Meier, 555-12
Flexible strategy
double dendrograms, 450-3
hierarchical clustering, 445-4
Flipping constant, 240-2
FNREG1 dataset, 360-15, 380-7
FNREG2 dataset, 365-11
FNREG3 dataset, 163-4, 370-6, 375-8
FNREG4 dataset, 371-6, 376-8
FNREG5 dataset, 351-30
Follow-up
life-table analysis, 570-2
Forced match variable, 123-4
Forced points
D-optimal designs, 267-5
Forced X's
variable selection, 310-4
Forecast
ARIMA, 471-11
automatic ARMA, 474-10
decomposition forecasting, 469-10
exponential smoothing, 465-8, 466-12, 467-10
Forecasts
multiple regression with serial correlation, 306-3
Forest plot
meta analysis of hazard ratios, 458-17
meta-analysis of correlated proportions, 457-20
meta-analysis of means, 455-17
meta-analysis of proportions, 456-20
Format, 102-6
Forward selection
Cox regression, 565-23
logistic regression, 320-17
Poisson regression, 325-6
Forward selection with switching
logistic regression, 320-18
multiple regression, 305-24
Poisson regression, 325-7
Forward variable selection
multiple regression, 305-23
Fourier plot
spectral analysis, 468-10
Fourier series
spectral analysis, 468-2
Fprob transformation, 119-9
Fraction transformation, 119-7
Fractional-factorial designs, 261-1
F-ratio
linear regression, 300-47
Freeman-Tukey standardized residual
loglinear models, 530-20
Frequency
spectral analysis, 468-1
Frequency polygon
histograms, 143-13
Frequency tables, 500-1
Frequency variable
linear regression, 300-25
Poisson regression, 325-8
Friedman's Q statistic, 211-15
Friedman's rank test, 211-3
FRUIT dataset, 141-1, 141-17
F-test
multiple regression, 305-50
FT-SR
loglinear models, 530-20
Full matching, 123-3
Function plots, 160-1
Functions
nonlinear regression, 315-4
Fuzz factor
filter, 121-2
in filter comparisons, 103-8
Fuzzifier
fuzzy clustering, 448-5
Fuzzy clustering, 448-1
FUZZY dataset, 448-3, 448-8
Fvalue transformation, 119-9

G

G statistic test
Poisson regression, 325-3
Gamma
hierarchical clustering, 445-8
Gamma distribution
probablility calculator, 135-4
simulation, 122-7
Gamma distribution fitting, 552-1
GammaProb transformation, 119-9
GammaValue transformation, 119-9
Gap between bars
bar charts, 141-14
Gap between sets of bars
bar charts, 141-15
Gart-Nam test
two proportions, 515-11
Gehan test
Kaplan-Meier, 555-12
nondetects analysis, 240-3
Geisser-Greenhouse adjustment, 214-1, 214-5

Geisser-Greenhouse epsilon, 214-4, 214-20
General linear models, 212-1
Generating data, 122-1
Generations
  hybrid appraisal models, 487-8
Geometric mean
  descriptive statistics, 200-14
Gleason-Staelin redundancy measure
  principal components analysis, 425-17
GLM
  checklist, 212-18
Gompertz model
  curve fitting, 351-7
  growth curves, 360-5
Goodness of fit
  loglinear models, 530-4
  Poisson regression, 325-3
Goodness-of-fit
  hierarchical clustering, 445-4
  K-means clustering, 446-2
  multidimensional scaling, 435-3
  ratio of polynomials, 370-2
Goto in output, 106-4
Graeco-Latin square designs, 263-1
Greedy matching, 123-1, 123-2
Greenwood's formula
  Kaplan-Meier, 555-3, 555-29, 555-33
  Weibull fitting, 550-3
Grid / tick settings window, 185-1
Grid lines, 185-1
Grid plot style file, 173-8
Grid plots, 140-11, 173-1
  response surface regression, 330-19
Grid range
  hybrid appraisal models, 487-9
Group average
  double dendrograms, 450-2
  hierarchical clustering, 445-4
Group variables
  logistic regression, 320-19
Growth curves, 360-1

H

HAIR dataset, 220-103
Harmonic mean
  descriptive statistics, 200-14
Hat diagonal
  linear regression, 300-19, 300-62
  multiple regression, 305-18, 305-64
Hat matrix
  linear regression, 300-18
  logistic regression, 320-14
  multiple regression, 305-18
  Poisson regression, 325-34
Hat values
  Poisson regression, 325-5
Hazard
  baseline, 565-8
  cumulative, 565-3
  Nelson-Aalen, 555-4
Hazard function
  beta distribution fitting, 551-2
  Cox regression, 565-2
  gamma distribution fitting, 552-2
Hazard function plot
  Kaplan-Meier, 555-36
Hazard rate
  Kaplan-Meier, 555-2
  life-table analysis, 570-3
  Weibull fitting, 550-2, 550-36
Hazard rate plot
  Kaplan-Meier, 555-36
Hazard ratio
  confidence interval, 555-40
  Kaplan-Meier, 555-40
Hazard ratio test
  Kaplan-Meier, 555-41
Hazard ratios
  meta analysis, 458-1
Hazard-baseline
  Cox regression, 565-38
HEART dataset, 212-23
Heat map colors, 187-5
Heat map settings window, 187-1
Help system, 1-10, 100-1
Heterogeneity test
  meta-analysis of proportions, 456-4
Heteroscedasticity
  linear regression, 300-3
Hierarchical cluster analysis, 450-1
  dendrograms, 450-3
Hierarchical clustering, 445-1
Hierarchical models
  Cox regression, 565-23
  loglinear models, 530-3
  multiple regression, 305-32
  response surface regression, 330-1
Hierarchical-classification designs, 212-27
Histogram
  bootstrap, 300-31, 305-42
  definition, 140-2
  density trace, 143-1
  descriptive statistics, 200-25
  linear regression, 300-34
  multiple regression, 305-67
  t-test, 205-20
  Xbar R, 250-32
Histogram style file, 143-16
Histograms, 140-2, 143-1
Histograms - comparative, 140-4, 151-1
Histograms – comparative
  density trace, 151-2
Holliday model
  curve fitting, 351-5
  growth curves, 360-4
Holt's linear trend, 466-1
Holt-Winters forecasting
  exponential smoothing, 467-1
Hotelling's one sample T2, 405-1
Hotelling's T2, 410-1
  1-Sample, 405-1
Hotelling's T2 distribution
  probablility calculator, 135-4
Hotelling's T2 value, 410-7
Hotelling's two-sample T2, 410-1
Hour format, 102-8
HOUSING dataset, 306-4, 306-10
Hsu's test
  one-way ANOVA, 210-6
Huber's method
  multiple regression, 305-26
Huynh Feldt epsilon, 214-20
Huynh-Feldt adjustment, 214-1
Hybrid appraisal models, 487-1
Hybrid model, 487-1
HYP(z)
  piecewise polynomial models, 365-6
Hypergeometric distribution
  probablility calculator, 135-4
HypergeoProb transformation, 119-9
Hypothesis tests
  linear regression, 300-6
  multiple regression, 305-13

I

Identicalness
  curve fitting, 350-6
IEEE format, 102-4
If-then transformations, 120-1
Import limitations, 115-1
Importing, 103-2
Importing data, 12-1, 115-1
Imputation, 118-1
  principal components analysis, 425-4
Imputing data values, 118-1
Incidence
  Poisson regression, 325-1
Incidence rate
  Poisson regression, 325-34
Inclusion points
  D-optimal designs, 267-6
Incomplete beta function ratio
  beta distribution fitting, 551-2
Independence tests
  cross tabulation, 501-1
Independent variable

linear regression, 300-25
Independent variables
logistic regression, 320-20
multiple regression, 305-1
multiple regression, 305-28
Poisson regression, 325-8
Indicator variables
creating, 119-19
multiple regression, 305-3
Individuals
hybrid appraisal models, 487-8
Individuals chart, 250-4
Xbar R, 250-33
Inertia
correspondence analysis, 430-13
Influence
multiple regression, 305-17
Influence report
linear regression, 300-66
Influence detection
linear regression, 300-65
Information matrix
Cox regression, 565-7
Inheritance
hybrid appraisal models, 487-9
Weibull fitting, 550-15
Initial communality
factor analysis, 420-3
Initial Tableau section
linear programming, 480-4
Initial values
backcasting, 465-2, 466-3, 467-3
Insert, 103-5
Installation, 1-1, 100-1
folders, 1-1
Int transformation, 119-7
INTEL dataset, 465-7, 466-9, 471-7, 473-5
Interaction
two level designs, 260-3
Interactions
multiple regression, 305-4
Intercept
linear regression, 300-25, 300-39
multiple regression, 305-34
Poisson regression, 325-15
Interquartile range
box plot, 152-1
descriptive statistics, 200-17
Interval censored
parametric survival regression, 566-3
Weibull fitting, 550-11
Interval data
Cox regression, 565-17
Interval failure
Kaplan-Meier, 555-15
Interval variables
fuzzy clustering, 448-4
hierarchical clustering, 445-6
medoid partitioning, 447-1
Intervals
tolerance, 585-1
Inverse prediction
linear regression, 300-6, 300-41, 300-67, 300-68
IQ dataset, 305-27, 305-43, 305-72, 305-76, 305-79
IQR
descriptive statistics, 200-17
Isolines, 140-11
contour plot, 172-1
Item analysis, 505-1
ITEM dataset, 505-2, 505-5, 506-2, 506-6
Item response analysis, 506-1

J

Jittering
dot plots, 150-1
Join transformation, 119-18
Julian date transformation, 119-6

K

K analysis
ridge regression, 335-22
K values
ridge regression, 335-8
Kaplan-Meier
Weibull fitting, 550-1
Kaplan-Meier estimates, 555-1
Kaplan-Meier product limit estimator
Weibull fitting, 550-3
Kaplan-Meier product-limit, 555-32
beta distribution fitting, 551-14
gamma distribution fitting, 552-16
nondetects analysis, 240-14
Weibull fitting, 550-33
Kaplan-Meier product-limit estimator
beta distribution fitting, 551-2
Kappa reliability test
cross tabulation, 501-15
Kaufman and Rousseeuw
medoid partitioning, 447-4
Kendall's coefficient
concordance, 211-15
Kendall's tau-B
cross tabulation, 501-15
Kendall's tau-C
cross tabulation, 501-15
Kenward and Roger method
mixed models, 220-28
Kernel-smoothed estimators
Kaplan-Meier, 555-9
Weibull fitting, 550-35
Keyboard
commands, 103-11
KLEIN6 dataset, 555-45
K-means cluster analysis, 446-1
KOCH36 dataset, 325-7, 325-21
Kolmogorov-Smirnov
descriptive statistics, 200-23
Kolmogorov-Smirnov test
two-sample, 206-1, 206-23
Kruskall-Wallis test statistic, 210-21
Kruskal-Wallis test, 210-1
Kruskal-Wallis Z test
one-way ANOVA, 210-7
Kurtosis, 200-2
descriptive statistics, 200-18
t-test, 205-15

L

L'Abbe plot
meta-analysis of correlated proportions, 457-22
meta-analysis of means, 455-18
meta-analysis of proportions, 456-22
Labeling values, 102-10
Labeling variables, 2-4
Labels
values, 13-1
LACHIN91 dataset, 320-71
Lack of fit
linear regression, 300-16
Lack-of-fit test
response surface regression, 330-1
Lagk transformation, 119-16
Lambda
canonical correlation, 400-10
discriminant analysis, 440-12
loglinear models, 530-18
Lambda A
cross tabulation, 501-14
Lambda B
cross tabulation, 501-15
Latin square designs, 263-1
LATINSQR dataset, 212-22
Latin-square
GLM, 212-21
Lawley-Hotelling trace
MANOVA, 415-3
Lcase transformation, 119-18
LEAD dataset, 240-19
Least squares
linear regression, 300-5
multiple regression, 305-13
Least squares trend, 466-1
Ledk transformation, 119-16

LEE91 database, 570-15
LEE91 dataset, 570-4
Left censored
  parametric survival regression, 566-3
  Weibull fitting, 550-11
Left transformation, 119-18
Length transformation, 119-18
LEUKEMIA dataset, 320-18, 320-34, 320-57
Levenberg-Marquardt algorithm, 385-1
Levene test
  linear regression, 300-27
  modified, 206-20
  modified (multiple-groups), 210-18
Levene test (modified)
  linear regression, 300-50
Levey-Jennings control charts, 252-1
Life-table analysis, 570-1
Like. ratio chi-square
  loglinear models, 530-13
Likelihood
  Cox regression, 565-5
Likelihood ratio
  1-sample binary diagnostic test, 535-3
  logistic regression, 320-8
  ROC curves, 545-24
Likelihood ratio test
  Cox regresion, 565-10
Likelihood ratio test of difference
  two proportions, 515-8
Likelihood-ratio statistic
  loglinear models, 530-4
Likert-scale
  simulation, 122-8, 122-22
Lilliefors' critical values
  descriptive statistics, 200-23
Limitations
  exporting, 116-1
Line charts, 140-1, 141-1
Line granularity
  linear regression, 300-33
Line settings window, 183-1
Linear discriminant functions
  discriminant analysis, 440-2
Linear model, 212-1
Linear programming, 480-1
Linear regression, 300-1
  assumptions, 300-3
Linearity
  MANOVA, 415-5
  multiple regression, 305-6
Linear-linear fit
  curve fitting, 351-11
Linear-logistic model, 320-1
Linkage type
  hierarchical clustering, 445-7
LINREG1 dataset, 300-24, 300-37
Ljung statistic
  automatic ARMA, 474-12
LLM, 530-1
Ln(X) transformation, 119-7
Loading a database, 2-1, 2-10, 7-1
Loess
  robust, 300-14
LOESS
  linear regression, 300-13
LOESS %N
  linear regression, 300-33
LOESS curve
  linear regression, 300-33
LOESS order
  linear regression, 300-33
LOESS robust
  linear regression, 300-34
Loess smooth
  scatter plot, 161-14
Log document, 106-1
Log file
  tutorial, 101-4
Log likelihood
  Poisson regression, 325-23
  Weibull fitting, 550-30
Log odds ratio transformation
  logistic regression, 320-2
Log of output, 9-6
Log transformation, 119-7
Logarithmic fit
  curve fitting, 351-8
LogGamma transformation, 119-9
Logistic distribution
  Weibull fitting, 550-10
Logistic item characteristic curve
  item response analysis, 506-1
Logistic model
  curve fitting, 351-6
  growth curves, 360-5
Logistic regression, 320-1
  parametric survival regression, 566-1
Logit transformation, 119-7
  logistic regression, 320-1
LOGLIN1 dataset, 530-7, 530-11
Loglinear models, 530-1
Log-logistic distribution
  Weibull fitting, 550-10
Log-logistic regression, 566-1
Lognormal
  curve fitting, 351-10, 351-11
  growth curves, 360-9
Lognormal distribution
  nondetects regression, 345-2
  Weibull fitting, 550-5
Lognormal regression, 566-1
Logrank test
  Kaplan-Meier, 555-41
Log-rank test
  Kaplan-Meier, 555-12
  nondetects analysis, 240-3
  randomization, 555-1
Log-rank tests
  Kaplan-Meier, 555-38
Longitudinal data example
  mixed models, 220-51
Longitudinal data models
  mixed models, 220-4
Longitudinal models, 220-1
Lookup transformation, 119-14
Lotus 123 exporting, 116-1
Lotus 123 importing, 115-1
Lowess smooth
  scatter plot, 161-14
LP dataset, 480-2, 480-4
LUNGCANCER dataset, 565-15, 565-31, 565-48

M

MA order (Q)
  automatic ARMA, 474-8
Macros, 130-1
  command list, 130-25
  commands, 130-6
  examples, 130-26
  syntax, 130-2
MAD
  descriptive statistics, 200-20
MAD constant
  multiple regression, 305-40
MAE
  exponential smoothing, 466-4, 467-2
Mallow's Cp
  variable selection and, 312-8
Mallow's Cp statistic
  multiple regression, 305-55
MAMMALS dataset, 3-1, 4-1, 10-1
MAMMALS1 dataset, 5-1, 6-1
Manhattan distance
  medoid partitioning, 447-3
Mann-Whitney U test, 206-1, 206-20
MANOVA, 415-1
  multivariate normality and Outliers, 415-4
MANOVA1 dataset, 410-3, 410-6, 415-5, 415-10
Mantel Haenszel test
  two proportions, 515-7
Mantel-Haenszel logrank test
  Kaplan-Meier, 555-41
Mantel-Haenszel test, 525-1
MAPE
  exponential smoothing, 466-4, 467-2
Maps
  contour plots, 172-1
  contour plots, 140-11
Mardia-Watson-Wheeler test

circular data, 230-10
Marginal association
loglinear models, 530-6
Martinez-Iglewicz
descriptive statistics, 200-22
Martingale residuals
Cox regression, 565-13, 565-39
Cox regression, 565-40
MARUBINI dataset, 560-3, 560-9
Mass
correspondence analysis, 430-13
Matched pairs
correlated proportions, 520-1
Matching
caliper, 123-4
caliper radius, 123-5
distance calculation method, 123-3
forced match variable, 123-4
full (variable), 123-3
greedy, 123-1, 123-2
optimal, 123-1, 123-2
propensity score, 123-2
standardized difference, 123-15
Mathematical functions
transformations, 119-7
Matrix determinant
equality of covariance, 402-8
Matrix type
principal components analysis, 425-11
Mauchley’s test of compound symmetry, 214-5
Mavk transformation, 119-16
Max % change in any beta
multiple regression, 305-78
Max terms
multiple regression, 305-33
Max transformation, 119-16
Maximum likelihood
Cox regression, 565-5
mixed models, 220-17
Weibull fitting, 550-10
Maximum likelihood estimates
beta distribution fitting, 551-12
McHenry's select algorithm, 310-1
McNemar test
correlated proportions, 520-1, 520-6
cross tabulation, 501-16
McNemar's tests, 501-1
MDB exporting, 116-1
MDB importing, 115-1
MDS, 435-1
MDS2 dataset, 435-6, 435-10, 435-15
Mean
confidence interval for, 200-13
descriptive statistics, 200-13
deviation, 200-20
geometric, 200-14
harmonic, 200-14
standard error of, 200-13
Mean absolute deviation
descriptive statistics, 200-20
Mean deviation
descriptive statistics, 200-20
estimate of standard error of, 200-20
Mean square
linear regression, 300-47
Mean squared error
linear regression, 300-19
multiple regression, 305-19
Mean squares
multiple regression, 305-50
Mean-deviation
descriptive statistics, 200-20
Means
meta-analysis of means, 455-1
Measurement error
R & R, 254-19
Measurement error ratio
R & R, 254-3
Median
cluster method, 445-4
confidence interval, 200-14
descriptive statistics, 200-14
Median cluster method
double dendrograms, 450-2
Median remaining lifetime
life-table analysis, 570-4, 570-22
Median smooth
scatter plot, 161-15
Median survival time
Kaplan-Meier, 555-30
Medoid clustering, 447-1
Medoid partitioning, 447-1
Membership
fuzzy clustering, 448-1
Merging two databases, 104-1
M-estimators
multiple regression, 305-25
Meta-analysis
correlated proportions, 457-1
Meta-analysis of hazard ratios, 458-1
Meta-analysis of means, 455-1
Meta-analysis of proportions, 456-1
METACPROP dataset, 457-6, 457-14
METAHR dataset, 458-6, 458-12
Method of moments estimates
beta distribution fitting, 551-12
Metric multidimensional scaling, 435-5
Michaelis-Menten
curve fitting, 351-1, 351-4
Miettinen - Nurminen test
two proportions, 515-8
Mill’s ratio
Kaplan-Meier, 555-2
Weibull fitting, 550-2
Min transformation, 119-16
Minimum Percent Beta Change, 305-40
Minute format, 102-8
Missing
if-then transformation, 120-4
Missing value estimation
factor analysis, 420-7
Missing values, 102-5, 320-18, 425-4
cross tabs, 501-4
descriptive tables, 201-7
estimating, 118-1
GLM, 212-19
principal components analysis, 425-3
Missing-value imputation
principal components analysis, 425-4
Mixed model
defined, 220-2
Mixed models, 220-1
AIC, 220-7
Bonferroni adjustment, 220-14
covariates, 220-9
differential evolution, 220-29
F test, 220-28
Fisher scoring, 220-29
fixed effects, 220-9
G matrix, 220-18
Kenward and Roger method, 220-28
L matrix, 220-26
likelihood formulas, 220-17
maximum likelihood, 220-17
MIVQUE, 220-29
model building, 220-13
multiple comparisons, 220-14
Newton-Raphson, 220-29
R matrix, 220-19
random vs repeated error, 220-7
restricted maximum likelihood, 220-18
technical details, 220-16
time, 220-11
types, 220-4
zero variance estimate, 220-8
Mixture design
D-optimal designs, 267-22
MLCO2 dataset, 470-11
Mod transformation, 119-7
Mode
descriptive statistics, 200-15
Model
Bleasdale-Nelder, 351-5, 360-3
exponential, 351-6, 360-4
Farazdaghi and Harris, 351-5, 360-3
four-parameter logistic, 351-7, 360-5
Gompertz, 351-7, 360-5

Holliday, 351-5, 360-4
Kira, 351-4, 360-2
monomolecular, 351-6, 360-4
Morgan-Mercer-Floding, 351-8, 360-6
multiple regression, 305-33
reciprocal, 351-4, 360-2
Richards, 351-8, 360-7
Shinozaki, 351-4, 360-2
three-parameter logistic, 351-6, 360-5
Weibull, 351-7, 360-6
Model size
all possible regressions, 312-8
Models
growth curves, 360-1
hierarchical, 530-3
multiphase, 365-1
multiple regression, 305-35
piecewise polynomial, 365-1
ratio of polynomials, 370-1, 375-1
sum of functions, 380-1
user written, 385-1
Modified Kuiper's test
circular data, 230-4
Moment
descriptive statistics, 200-11
Monomolecular model
curve fitting, 351-6
growth curves, 360-4
Monte Carlo samples
1-Sample T2, 405-4
linear regression, 300-31
Monte Carlo simulation, 122-1
Month format, 102-8
Month transformation, 119-6
Morgan-Mercer-Floding model
curve fitting, 351-8
growth curves, 360-6
MOTORS dataset, 566-3, 566-11
Moving average chart, 250-4
Moving average chart limits, 250-8
Moving average parameters
ARIMA, 471-3
theoretical ARMA, 475-2
Moving data, 103-14
Moving range
Xbar R, 250-33
Moving range chart, 250-4
MSEi
multiple regression, 305-19
Multicollinearity
canonical correlation, 400-2
discriminant analysis, 440-4
MANOVA, 415-5
multiple regression, 305-7
ridge regression, 335-1
stepwise regression, 311-2
Multicollinearity report
multiple regression, 305-57
Multidimensional scaling, 435-1
metric, 435-1
Multinomial chi-square tests
frequency tables, 500-1
Multinomial distribution
simulation, 122-8
Multinomial test
frequency tables, 500-10
Multiple comparisons
Bonferroni, 210-4
box plots, 152-2
Duncan's test, 210-5
Dunn's test, 210-7
Dunnett's test, 210-6
Fisher's LSD, 210-6
Hsu's test, 210-6
Kruskal-Wallis Z test, 210-7
mixed models, 220-14
Newman-Keuls test, 210-8
one-way ANOVA, 210-3
recommendations, 210-8
Scheffe's test, 210-8
Tukey-Kramer test, 210-8
Multiple regression
robust, 305-24
Multiple regression, 305-1
assumptions, 305-6
Multiple regression
all possible, 312-1
Multiple regression
binary response, 320-6
Multiple regression with serial correlation, 306-1
Multiplicative seasonality
exponential smoothing, 467-2
Multiplicity factor
t-test, 205-19
Multivariate analysis of variance, 415-1
Multivariate normal
factor analysis, 420-7
principal components analysis, 425-11
Multivariate polynomial ratio fit, 376-1
Multivariate variable selection, 310-1
Multiway frequency analysis
loglinear models, 530-1
Mutation rate
hybrid appraisal models, 487-9
Weibull fitting, 550-15

N

Nam and Blackwelder test
correlated proportions, 520-5
Nam test
correlated proportions, 520-7
Nam's score
correlated proportions, 520-2
Navigator, 107-1
NC CRIMINAL dataset, 320-64, 320-68
NcBetaProb transformation, 119-9
NcBetaValue transformation, 119-10
NcCsProb transformation, 119-10
NcCsValue transformation, 119-10
NcFprob transformation, 119-10
NcFvalue transformation, 119-10
NCSS
quitting, 101-4
NcTprob transformation, 119-10
NcTvalue transformation, 119-10
Nearest neighbor
double dendrograms, 450-2
hierarchical clustering, 445-3
Negative binomial distribution
probablility calculator, 135-5
Negative binomial transformation, 119-10
NegBinomProb transformation, 119-10
Neighborhood
appraisal ratios, 485-7
Nelson-Aalen estimates
Weibull fitting, 550-1
Nelson-Aalen estimator, 555-7
Weibull fitting, 550-33
Nelson-Aalen hazard
Kaplan-Meier, 555-1
Weibull fitting, 550-4
Nested factor
GLM, 212-4
Nested factors
design generator, 268-1
New database, 103-1
New spreadsheet, 103-1
New template, 105-1
Newman-Keuls test
one-way ANOVA, 210-8
Newton-Raphson
Weibull fitting, 550-11
Nominal variables
fuzzy clustering, 448-4
hierarchical clustering, 445-7
medoid partitioning, 447-2
Non-central Beta transformation, 119-10
Non-central Chi-square transformation, 119-10
noncentral-F distribution
transformation, 119-10
Noncentral-t distribution
transformation, 119-10
Nondetects analysis, 240-1
confidence limits, 240-7
flipping constant, 240-2
Gehan test, 240-3

Kaplan-Meier product-limit, 240-14
log-rank test, 240-3
Peto-Peto test, 240-3
Tarone-Ware test, 240-3
NONDETECTS dataset, 240-4
Nondetects regression, 345-1
confidence limits, 345-11
Cox-Snell residual, 345-13
R-squared, 345-11
standardized residual, 345-13
Noninferiority
2-sample binary diagnostic, 537-10
clustered binary diagnostic, 538-9
paired binary diagnostic, 536-8
ROC curves, 545-31
Noninferiority test
correlated proportions, 520-8
two proportions, 515-17
Noninferiority tests
two proportions, 515-37
Nonlinear regression, 315-1
appraisal, 487-1
functions, 315-4
starting values, 315-1
user written models, 385-1
Nonparametric tests
t-test, 205-17
Nonstationary models
Box Jenkins, 470-3
Normal
curve fitting, 351-10
growth curves, 360-9
Normal distribution
probablility calculator, 135-5
simulation, 122-9, 122-20
Weibull fitting, 550-4
Normal line
histograms, 143-12
Normal probability plot
descriptive statistics, 200-26
Normality, 200-4
descriptive statistics, 200-21
ROC curves, 545-12
t-test, 205-15
Normality test alpha, 118-3
Normality tests
Anderson-Darling test, 200-22
D'Agostino kurtosis, 200-24
D'Agostino omnibus, 200-25
D'Agostino skewness, 200-23
descriptive statistics, 200-21
Kolmogorov-Smirnov, 200-23
Lilliefors' critical values, 200-23
linear regression, 300-48
Martinez-Iglewicz, 200-22
multiple regression, 305-52
Shapiro-Wilk test, 200-22
skewness test, 200-24
tolerance intervals, 585-11
NormalProb transformation, 119-10
NormalValue transformation, 119-10
NormScore transformation, 119-16
Notes
omitting them in linear regression, 300-26
NP-chart, 251-1
Number exposed
life-table analysis, 570-2
Number of correlations
canonical correlation, 400-5
Number of points
linear regression, 300-33
Numeric data, 102-1
Numeric functions, 119-6

O

Objective function
linear programming, 480-1
Observational study matching, 123-1
Observational study stratification, 124-1
Odds ratio
1-sample binary diagnostic test, 535-4
2-sample binary diagnostic, 537-9
confidence interval of, 515-23
correlated proportions, 520-5
meta-analysis of correlated proportions, 457-2
meta-analysis of proportions, 456-2
two proportions, 515-1, 515-3
Odds ratios
Mantel-Haenszel test, 525-1
ODOR dataset, 330-3, 330-11
Omission report
multiple regression, 305-54
One proportion, 510-1
One-sample tests, 205-1
One-sample t-test, 205-1
One-way analysis of variance, 210-1
One-way ANOVA
Bonferroni, 210-4
Duncan's test, 210-5
Dunn's test, 210-7
Dunnett's test, 210-6
Fisher's LSD, 210-6
Hsu's test, 210-6
Kruskal-Wallis Z test, 210-7
multiple comparisons, 210-3
Newman-Keuls test, 210-8
orthogonal contrasts, 210-11
orthogonal polynomials, 210-11
planned comparisons, 210-10
Scheffe's test, 210-8
Tukey-Kramer test, 210-8
Open database, 103-1
Open log file, 106-2
Open output file, 106-2
Open spreadsheet, 103-1
Open template, 105-1
Opening a database
tutorial, 101-3
Optimal matching, 123-1, 123-2
Optimal solution section
linear programming, 480-5
Optimal value
linear programming, 480-5
Or
if-then transformation, 120-2
Ordinal variables
fuzzy clustering, 448-4
hierarchical clustering, 445-6
medoid partitioning, 447-2
Original cost
linear programming, 480-5
Orthogonal arrays, 266-1
Orthogonal contrasts
one-way ANOVA, 210-11
Orthogonal polynomial
ANOVA balanced, 211-6
GLM, 212-5
repeated measures, 214-11
Orthogonal polynomials
one-way ANOVA, 210-11
Orthogonal regression
linear regression, 300-9, 300-41
Orthogonal sets of Latin squares, 263-2
Outlier detection
linear regression, 300-64
multiple regression, 305-83
Outlier report
linear regression, 300-66
Outliers
Cox regression, 565-14
linear regression, 300-15
multiple regression, 305-1, 305-24, 305-78
stepwise regression, 311-3
t-test, 205-22
Outliers, 200-3
Output, 106-1
log of, 9-6
printing, 9-4
ruler, 106-4
saving, 9-5
Output document, 106-1
Output window, 1-6, 9-1
Overdispersion
Poisson regression, 325-3, 325-12
Overlay
scatter plot, 161-3

P

Page setup, 103-2
PAIN dataset, 220-51
Paired data
  clustered binary diagnostic, 538-11
Paired t-test
  1-Sample T2, 405-1
Paired t-tests, 205-1
Pair-wise removal
  correlation matrix, 401-3
Paradox exporting, 116-1
Paradox importing, 115-1
Parallel slopes
  multiple regression, 305-86
Parameterization
  curve fitting, 350-5
Pareto chart, 253-1
Pareto charts, 250-41
Parsimony
  ratio of polynomials, 370-2
Partial association
  loglinear models, 530-5
Partial autocorrelation, 472-1
Partial autocorrelation function
  Box Jenkins, 470-4
Partial correlation
  multiple regression, 305-56
Partial residual plots, 305-71
Partial variables
  canonical correlation, 400-4
  correlation matrix, 401-3
Partial-regression coefficients, 305-47
Partition coefficient
  fuzzy clustering, 448-3
Paste, 103-4
Paste output, 106-3
Pasting data, 7-2
PCA, 425-1
PCA2 dataset, 118-4, 420-5, 420-11, 425-9, 425-15
P-chart, 251-1
Pearson chi-square
  loglinear models, 530-4, 530-13
Pearson correlation
  linear regression, 300-45
Pearson correlations
  matrix of, 401-1
Pearson residuals
  logistic regression, 320-13
  Poisson regression, 325-5, 325-31
Pearson test
  Poisson regression, 325-3
Pearson's contingency coefficient
  cross tabulation, 501-14
Percentile plots, 140-5
Percentile Plots, 153-1
Percentile type
  descriptive statistics, 200-6
Percentiles, 200-2
Percentiles of absolute residuals
  multiple regression, 305-78
Period effect
  cross-over analysis using t-tests, 235-4
Period plot
  cross-over analysis using t-tests, 235-24
Periodogram
  spectral analysis, 468-1
Perspective
  3D scatter plot, 170-6
  3D surface plot, 171-6
  bar charts, 141-12
PET dataset, 538-11
Peto-Peto test
  Kaplan-Meier, 555-12
  nondetects analysis, 240-3
Phase
  spectral analysis, 468-1
Phi
  cross tabulation, 501-14
  factor analysis, 420-13
  Poisson regression, 325-3, 325-12, 325-27
  principal components analysis, 425-17
Phis
  theoretical ARMA, 475-2
Pie charts, 140-2, 142-1
PIE dataset, 142-6
Piecewise polynomial models, 365-1
Pillai's trace
  MANOVA, 415-3
Plackett-Burman design, 265-1
Planned comparisons
  one-way ANOVA, 210-10
PLANT dataset, 212-27
Plot size
  linear regression, 300-29
Plots
  3D scatter plots, 140-10, 170-1
  3D surface plots, 140-10, 171-1
  area charts, 140-1, 141-1
  bar charts, 140-1, 141-1
  box plots, 140-5, 152-1
  contour plots, 140-11, 172-1
  density trace, 143-1
  dot plots, 140-4, 150-1
  error-bar charts, 140-6, 155-1
  function plots, 160-1
  grid plots, 140-11, 173-1
  histograms, 140-2, 143-1
  histograms - comparative, 140-4, 151-1
  line charts, 140-1, 141-1
  percentile plots, 140-5, 153-1
  pie charts, 140-2
  probability plots, 140-3, 144-1
  scatter plot matrix, 140-8, 162-1
  scatter plot matrix (curve fitting), 163-1
  scatter plot matrix for curve fitting, 140-9
  scatter plots, 140-7, 161-1
  single-variable charts, 140-1
  surface charts, 140-1, 141-1
  surface plots, 140-10, 171-1
  three-variable charts, 140-10
  two-variable charts, 140-4, 140-7
  violin plots, 140-6, 154-1
POISREG dataset, 325-37
Poisson distribution
  probablility calculator, 135-5
  simulation, 122-9
Poisson regression, 325-1
PoissonProb transformation, 119-11
POLITIC dataset, 13-1, 14-1
Polynomial
  logistic regression, 320-23
  multiple regression, 305-31
  multivariate ratio fit, 376-1
  Poisson regression, 325-11
Polynomial fit
  scatter plot, 161-13
Polynomial model
  response surface regression, 330-1
Polynomial models, 365-1
Polynomial ratio fit, 375-1
Polynomial ratios
  model search (many X variables), 371-1
Polynomial regression model, 330-1
Polynomials
  ratio of, 370-1, 375-1
Pooled terms, 213-2
POR exporting, 116-1
Portmanteau test
  ARIMA, 471-12
  automatic ARMA, 474-12
  Box Jenkins, 470-10
Power
  multiple regression, 305-47
Power spectral density
  spectral analysis, 468-3
Power spectrum
  theoretical ARMA, 475-8
PRD
  appraisal ratios, 485-8
Precision-to-tolerance
  R & R, 254-20
Precision-to-tolerance ratio
  R & R, 254-3
Predicted value
  Poisson regression, 325-32
Predicted values
  linear regression, 300-27, 300-52
  multiple regression, 305-61
Prediction interval

multiple regression, 305-61
Prediction limits
linear regression, 300-33, 300-53, 300-59
multiple regression, 305-61
Pre-post
multiple regression, 305-87
PREPOST dataset, 305-87
PRESS
linear regression, 300-21, 300-51
multiple regression, 305-21, 305-51
PRESS R2
multiple regression, 305-52
Press R-squared
multiple regression, 305-21
PRESS R-squared
linear regression, 300-22
Prevalence
ROC curves, 545-5
Price related differential
appraisal ratios, 485-8
hybrid appraisal models, 487-17
Principal axis method
factor analysis, 420-1
Principal components
linear regression, 300-9
Principal components analysis, 425-1
Principal components regression, 340-1
Print
output, 106-3
Printer setup, 103-2
Printing
data, 2-7, 103-3
output, 9-4
output reports, 4-5
Printing data, 117-1
Prior probabilities
discriminant analysis, 440-5
Prob level, 415-13
linear regression, 300-47
Prob to enter
stepwise regression, 311-4
Prob to remove
stepwise regression, 311-4
Probability Calculator, 135-1
Beta distribution, 135-1
Binomial distribution, 135-2
Bivariate normal distribution, 135-2
Chi-square distribution, 135-2
Correlation coefficient distribution, 135-3
F distribution, 135-3
Gamma distribution, 135-4
Hotelling's T2 distribution, 135-4
Hypergeometric distribution, 135-4
Negative binomial distribution, 135-5
Normal distribution, 135-5
Poisson distribution, 135-5
Student's t distribution, 135-6
Studentized range distribution, 135-6
Weibull distribution, 135-6
Probability ellipse
linear regression, 300-8, 300-33
Probability functions
transformations, 119-8
Probability plot
descriptive statistics, 200-26
linear regression, 300-57
multiple regression, 305-67
t-test, 205-20
Weibull, 144-17
Probability plot style file, 144-19
Probability plots, 140-3
asymmetry, 144-3
quantile scaling, 144-7
Probability Plots, 144-1
Probit analysis, 575-1
Probit plot
probit analysis, 575-10
Procedure, 105-1
running, 101-3
Procedure window, 1-5, 8-1
Product-limit survival distribution
beta distribution fitting, 551-14
gamma distribution fitting, 552-16
Kaplan-Meier, 555-32
Weibull fitting, 550-33
Product-moment correlation
correlation matrix, 401-3
Profiles
correspondence analysis, 430-1
Projection method
3D scatter plot, 170-8
3D surface plot, 171-7
bar charts, 141-14
PROPENSITY dataset, 123-5, 123-12, 124-4
Propensity score, 123-2
stratification, 124-1
Proportion trend test
Armitage, 501-5
Proportions
2-sample binary diagnostic, 537-1
clustered binary diagnostic, 538-1
confidence interval of ratio, 515-21
correlated, 520-1
Meta-analysis of correlated proportions, 457-1
meta-analysis of proportions, 456-1
one, 510-1
paired binary diagnostic, 536-1
two, 515-1
Proportions test
1-sample binary diagnostic test, 535-1
Proximity matrix
multidimensional scaling, 435-1
Proximity measures
multidimensional scaling, 435-4
Pseudo R-squared
multidimensional scaling, 435-12
Poisson regression, 325-4
Pure error
linear regression, 300-16

Q

QATEST dataset, 250-14, 250-27, 250-33, 250-35, 250-37, 251-3, 251-11, 253-2, 253-7, 253-9
Quadratic fit
curve fitting, 351-2
Qualitative factors
D-optimal designs, 267-6, 267-25
Quality
correspondence analysis, 430-13
Quantile scaling
probability plots, 144-7
Quantile test, 205-17
Quantiles
Kaplan-Meier, 555-30
Quartiles
descriptive statistics, 200-21
Quartimax rotation
factor analysis, 420-4
principal components analysis, 425-8
Quatro exporting, 116-1
Quick launch window, 107-1, 107-2
Quick start, 100-1
Quitting NCSS, 101-4

R

R & R study, 254-1
Radial plot
meta analysis of hazard ratios, 458-18
meta-analysis of correlated proportions, 457-21
meta-analysis of means, 455-18
meta-analysis of proportions, 456-21
Random coefficients example
mixed models, 220-103
Random coefficients models
mixed models, 220-5

Random effects model
meta-analysis of correlated proportions, 457-5
meta-analysis of hazard ratios, 458-5
meta-analysis of means, 455-4
meta-analysis of proportions, 456-5
Random effects models, 220-1
mixed models, 220-4
Random factor
ANOVA balanced, 211-5
GLM, 212-4
repeated measures, 214-8
Random numbers, 122-1
uniform, 15-1
Randomization
Latin square designs, 263-2
Randomization test
curve fitting, 351-16
linear regression, 300-24
log-rank, 555-1
T2, 410-7
Randomization tests
1-Sample T2, 405-1, 405-8
T2, 410-1
Randomized block design
repeated measures, 214-6
RandomNormal transformation, 119-11
Random-number functions
transformations, 119-11
Range
descriptive statistics, 200-17
interquartile, 200-17
Range chart, 250-1
Rank transformation, 119-16
Rate ratio
Poisson regression, 325-30
Ratio of polynomials
model search (many X variables), 371-1
model search (one X variable), 370-1
Ratio of polynomials fit, 375-1
many variables, 376-1
Ratio of two proportions
two proportions, 515-6
Ratio plot
decomposition forecasting, 469-12
Ratio section
appraisal ratios, 485-7
Ratio study
appraisal ratios, 485-1
Ratio variables
fuzzy clustering, 448-4
hierarchical clustering, 445-6
medoid partitioning, 447-2
Rayleigh test
circular data, 230-4
Rbar-squared
linear regression, 300-8
multiple regression, 305-15
RCBD data example
mixed models, 220-94
RCBD dataset, 220-94
REACTION dataset, 214-6, 214-29
Readout
parametric survival regression, 566-3
Weibull fitting, 550-11
READOUT105 dataset, 550-47
Rearrangement functions
transformations, 119-12
Recalc all, 103-9, 119-4
Recalc current, 103-8, 119-4
Reciprocal model
curve fitting, 351-4
growth curves, 360-2
Recode functions transformations, 119-14
Recode transformation, 3-4, 119-15
Recoding, 11-1
Reduced cost
linear programming, 480-5
Redundancy indices
canonical correlation, 400-4
Reference group
logistic regression, 320-19
Reference value
logistic regression, 320-21
multiple regression, 305-3, 305-29
Poisson regression, 325-9
Xbar R, 250-23
Reflection C.I. method
multiple regression, 305-41
Reflection method
linear regression, 300-30
two proportions, 515-28
REGCLUS dataset, 449-2, 449-5
Regression
all possible, 312-1
appraisal model, 487-1
backward selection, 311-2
binary response, 320-1, 320-8
clustering, 449-1
Cox, 565-1
diagnostics, 305-63
exponential, 566-1
extreme value, 566-1
forward selection, 311-1
growth curves, 360-1
hybrid appraisal model, 487-1
linear, 300-1
logistic, 320-1, 566-1
log-logistic, 566-1
lognormal, 566-1
model search (many X variables), 371-1
multiple, 312-8
nondetects, 345-1
nonlinear, 315-1
normal, 566-1
orthogonal regression, 300-9
Poisson, 325-1
polynomial ratio, 375-1
polynomial ratio (search), 370-1
principal components, 340-1
proportional hazards, 565-1
response surface regression, 330-1
ridge, 335-1
stepwise, 311-1
sum of functions models, 380-1
user written, 385-1
variable selection, 311-1
Weibull, 566-1
Regression analysis, 6-1
multiple regression, 305-1
Regression clustering, 449-1
Regression coefficients
Cox regression, 565-32
Regression coefficients report
multiple regression, 305-48
Regression equation report
multiple regression, 305-46
Relative risk
meta-analysis of correlated proportions, 457-2
meta-analysis of proportions, 456-2
two proportions, 515-1
Reliability
beta distribution fitting, 551-1, 551-15
gamma distribution fitting, 552-1
item analysis, 505-1
Kaplan-Meier, 555-1
kappa, 501-15
Weibull fitting, 550-1
Reliability analysis
Weibull fitting, 550-1
Reliability function
beta distribution fitting, 551-2
gamma distribution fitting, 552-2
Weibull fitting, 550-2
Remove last sheet, 103-2
Remove transformation, 119-18
Removed lambda
discriminant analysis, 440-12
Repeat transformation, 119-18
Repeatability
R & R, 254-1, 254-14
Repeated measures, 214-1
1-Sample T2, 405-6
mixed models, 220-1
Repeated measures data example
mixed models, 220-51
Repeated measures design
generating, 268-7
Repeated-measures design

GLM, 212-23
Replace, 103-6
Replace in output, 106-4
Replace transformation, 119-18
Replication
two level designs, 260-4
Reporting data, 117-1
Reports
selecting in linear regression, 300-26
Reproducibility
R & R, 254-1, 254-14
RESALE dataset, 117-4, 151-14, 155-1, 155-7, 201-1, 201-11, 201-12, 201-14, 201-15, 201-17, 201-19, 201-21, 305-81, 500-1, 500-9, 500-10, 500-12, 500-14, 501-1, 501-8, 501-11, 501-17
Resampling tab
linear regression, 300-29
Residual
diagnostics, 305-63
linear regression, 300-2, 300-18
multiple regression, 305-17
Residual diagnostics
linear regression, 300-15
multiple regression, 305-15
Poisson regression, 325-33
Residual life
life-table analysis, 570-22
Weibull fitting, 550-40
Residual plots
linear regression, 300-53
multiple regression, 305-67, 305-70
partial residuals, 305-71
Residual report
linear regression, 300-61
multiple regression, 305-62
Residuals
Cox regression, 565-13
Cox regression, 565-39
logistic regression, 320-11
multiple regression, 305-1
Poisson regression, 325-4, 325-31
Residuals-deviance
Cox regression, 565-14
Residuals-Martingale
Cox regression, 565-13
Residuals-scaled Schoenfeld
Cox regression, 565-15
Residuals-Schoenfeld
Cox regression, 565-14
Response surface regression, 330-1
Response-surface designs, 264-1
Restart method
Xbar R, 250-23
Restricted maximum likelihood
mixed models, 220-18
Richards model
curve fitting, 351-8
growth curves, 360-7
Ridge regression, 335-1
Ridge trace
ridge regression, 335-4, 335-18
RIDGEREG dataset, 335-7, 335-15, 340-3, 340-11
Right censored
parametric survival regression, 566-2
Weibull fitting, 550-11
Right transformation, 119-19
Right-hand sides
linear programming, 480-1
Risk ratio
correlated proportions, 520-4
Cox regression, 565-33, 565-35
meta-analysis of correlated proportions, 457-2
meta-analysis of proportions, 456-2
Risk set
Cox regression, 565-16
Kaplan-Meier, 555-3
RMSF dataset, 545-3
RNDBLOCK dataset, 211-4, 211-11, 212-3, 212-12
Robins odds ratio C. L.
Mantel-Haenszel test, 525-11
Robust estimation
principal components analysis, 425-5
Robust iterations
Xbar R, 250-18
Robust loess
linear regression, 300-14
Robust method
multiple regression, 305-39
Robust regression
multiple regression, 305-24, 305-31
Robust regression reports
multiple regression, 305-77
Robust regression tutorial
multiple regression, 305-76
Robust sigma multiplier
Xbar R, 250-18
Robust tab
multiple regression, 305-39
Robust weight
factor analysis, 420-7
principal components analysis, 425-11
Robust weights
multiple regression, 305-78
ROC curves, 545-1
comparing, 545-9
ROC dataset, 545-19
Root MSE
all possible regressions, 312-8
Rose plot
circular data, 230-16
Rose plots, 230-1
Rotation
3D scatter plot, 170-7
3D surface plot, 171-7
bar charts, 141-13
factor analysis, 420-7
principal components analysis, 425-11
Round transformation, 119-7
Row heights, 103-15
Row profiles
correspondence analysis, 430-1
Rows, 251-4, 251-5
Row-wise removal
correlation matrix, 401-3
Roy’s largest root
MANOVA, 415-4
RRSTUDY dataset, 254-1, 254-10
RRSTUDY1 dataset, 254-24
R-squared
adjusted, 300-46
adjusted, 305-45
all possible regressions, 312-8
Cox regression, 565-11
definition, 305-44
linear regression, 300-7, 300-46
logistic regression, 320-10
multiple regression, 305-14
Poisson regression, 325-4, 325-24
R-squared increment
stepwise regression, 311-8
R-squared report
multiple regression, 305-53
R-squared vs variable count plot, 310-8
RStudent
linear regression, 300-20, 300-62
multiple regression, 305-19, 305-63
RStudent plot
multiple regression, 305-69
Rstudent residuals
scatter plot of, 300-55
RTF, 106-3
tutorial, 101-4
RTF output format, 106-1
Ruler
output, 106-4
Run summary report
multiple regression, 305-44
Running a procedure
tutorial, 101-3
Running a regression analysis, 6-1
Running a two-sample t-test, 5-1
Running descriptive statistics, 4-1
Runs tests
attribute charts, 251-3
Xbar R, 250-9

S

S0 database, 102-1
S0/S0Z comparison, 102-4
S0Z/S0 comparison, 102-4
Sale date variable
  appraisal ratios, 485-4
  comparables, 486-7
Sale price variables
  appraisal ratios, 485-2
SALES dataset, 467-9, 469-9
Sales price
  multiple regression, 305-81
SALESRATIO dataset, 485-1, 485-6, 486-4
SAMPLE dataset, 101-3, 161-20, 162-5, 171-9, 172-7, 200-4, 200-10, 205-12, 206-12, 210-16, 310-3, 310-6, 311-3, 311-6, 312-2, 312-6, 400-8, 401-2, 401-5, 585-8
SAS exporting, 116-1
SAS importing, 115-1
Saturated model
  loglinear models, 530-3
Save, 103-3
Save as, 103-3
Save output, 106-3
Saved colors, 180-3
Saving
  data, 2-6
    tutorial, 101-2
  output, 9-5
  template, 8-5
Saving a template, 105-2
Saving results
  multiple regression, 305-42
SC
  medoid partitioning, 447-5
Scaled Schoenfeld residuals
  Cox regression, 565-15, 565-42
Scaling
  multidimensional, 435-1
Scaling factors
  D-optimal designs, 267-2
Scaling method
  fuzzy clustering, 448-5
  hierarchical clustering, 445-8
Scatter plot
  loess smooth, 161-14
  lowess smooth, 161-14
  median smooth, 161-15
  overlay, 161-3
  polynomial fit, 161-13
  spline, 161-15
  sunflower plot, 161-18
Scatter plot matrix, 140-8, 162-1
Scatter plot matrix (curve fitting), 163-1
Scatter plot matrix for curve fitting, 140-9
Scatter plot style file, 161-22
Scatter plots, 140-7, 161-1
  3D, 140-10, 170-1
Scheffe's test
  one-way ANOVA, 210-8
Schoenfeld residuals
  Cox regression, 565-14, 565-41
Schuirmann's test
  cross-over analysis using t-tests, 235-7
Scientific notation, 102-4
Score, 320-45
Score coefficients
  factor analysis, 420-17
  principal components analysis, 425-2
Scores plots
  canonical correlation, 400-12
Scree graph
  factor analysis, 420-3
Scree plot
  factor analysis, 420-15
  principal components analysis, 425-18
Screening data, 118-1, 200-3
Screening designs, 265-1
Searches
  ratio of polynomials, 370-1, 371-1
Seasonal adjustment
  exponential smoothing, 467-1
Seasonal autoregressive parameters
  ARIMA, 471-3
Seasonal decomposition forecasting, 469-1
Seasonal differencing
  ARIMA, 471-2
Seasonal moving average parameters
  ARIMA, 471-3
Seasonal time series
  Box Jenkins, 470-4
Second format, 102-8
Select all output, 106-4
Selecting procedures, 1-7
Selection method
  stepwise regression, 311-4
Selection procedure
  forward, 311-1
Sensitivity
  1-sample binary diagnostic test, 535-2
  2-sample binary diagnostic, 537-2
  clustered binary diagnostic, 538-8
  paired binary diagnostic, 536-2
  ROC curves, 545-1, 545-24
Sequence plot
  multiple regression, 305-69
Sequence transformation, 119-6
Sequential models report
  multiple regression, 305-56
Ser transformation, 119-6
Serial correlation
  linear regression, 300-4
  residuals, 305-53
Serial correlation plot
  multiple regression, 305-68
Serial numbers, 1-3, 100-1
Serial-correlation
  linear regression, 300-50
SERIESA dataset, 470-8, 474-7
Shapiro-Wilk
  linear regression, 300-18
  multiple regression, 305-17
Shapiro-Wilk test
  descriptive statistics, 200-22
  linear regression, 300-49
Shinozaki and Kari model
  curve fitting, 351-4
  growth curves, 360-2
Short transformation, 119-7
Sigma
  Xbar R, 250-19
Sigma multiplier
  Xbar R, 250-17
Sign test, 205-17
Sign transformation, 119-8
SIGN(z)
  piecewise polynomial models, 365-6
Signal-to-noise ratio
  R & R, 254-3
Silhouette
  fuzzy clustering, 448-9
  medoid partitioning, 447-13
Silhouettes
  medoid partitioning, 447-5
Similarities
  multidimensional scaling, 435-4
Simple average
  double dendrograms, 450-2
  hierarchical clustering, 445-3
Simplex algorithm
  linear programming, 480-1
Simulation, 122-1
  Beta distribution, 122-3
  Binomial distribution, 122-5
  Cauchy distribution, 122-5
  Constant distribution, 122-6
  contaminated normal, 122-21
  data, 15-1
  Exponential distribution, 122-6
  F distribution, 122-7
  Gamma distribution, 122-7
  Likert-scale, 122-8, 122-22
  Multinomial distribution, 122-8
  Normal distribution, 122-9, 122-20
  Poisson distribution, 122-9
  skewed distribution, 122-10

Student's T distribution, 122-10
syntax, 122-13
T distribution, 122-10
Tukey's lambda distribution, 122-10
Uniform distribution, 122-11
Weibull distribution, 122-12
Simultaneous C.I.'s
T2, 405-9, 410-10
Sin transformation, 119-17
Single linkage
double dendrograms, 450-2
hierarchical clustering, 445-3
Single-to-noise ratio
R & R, 254-19
Single-variable charts, 140-1
Sinh transformation, 119-17
Skewed distribution
simulation, 122-10
Skewness, 200-2
descriptive statistics, 200-17
t-test, 205-15
Skewness test
descriptive statistics, 200-24
Slices
pie charts, 142-1
Slope
linear regression, 300-39
Slopes
testing for equal
multiple regression, 305-86
SMOKING dataset, 525-2, 525-5
Smooth transformation, 119-16
Smoothing constant
exponential smoothing, 465-1, 466-2
Smoothing constants
exponential smoothing, 467-2
Smoothing interval
item response analysis, 506-4
Solo exporting, 116-1
Solo exporting, 116-1
Solo importing, 115-1
Sort, 103-6
Sort transformation, 119-12
Spath
medoid partitioning, 447-4
SPC fundamentals
Xbar R, 250-38
Spearman correlation
linear regression, 300-45
Spearman rank
correlation matrix, 401-3
Spearman rank correlation
linear regression, 300-12
Specificity
1-sample binary diagnostic test, 535-2
2-sample binary diagnostic, 537-2
clustered binary diagnostic, 538-8
paired binary diagnostic, 536-2
ROC curves, 545-1, 545-24
Spectral analysis, 468-1
Spectral density
spectral analysis, 468-3
Spectrum
spectral analysis, 468-1
Sphericity test
factor analysis, 420-14
Splice transformation, 119-12
Spline
scatter plot, 161-15
Split plot analysis
mixed models, 220-1
Split plot data example
mixed models, 220-98
Spread, 140-5
Spreadsheet
limits, 102-1
overview, 102-1
Spreadsheet/database comparison, 102-4
SPSS importing, 115-1
Sqrt transformation, 119-8
Standard deviation, 200-16
confidence limits, 207-2
descriptive statistics, 200-16
ratio, 207-2
unbiased, 200-17
Standard error, 200-13
linear regression, 300-40
Poisson regression, 325-26
Standardization
PC regression, 340-1
ridge regression, 335-3
Standardize transformation, 119-16
Standardized coefficients
linear regression, 300-40
multiple regression, 305-49
Standardized difference, 123-15
Standardized residual
linear regression, 300-19, 300-61, 300-64
multiple regression, 305-18, 305-63
nondetects regression, 345-13
Start time variable
Weibull fitting, 550-12
Starting NCSS, 1-2, 2-1, 100-1, 101-2
Starting values
curve fitting, 350-3
nonlinear regression, 315-1
Stata file exporting, 116-1
Statistical functions transformations, 119-15
Std error
of kurtosis, 200-18
of skewness, 200-18
of standard deviation, 200-16
of variance, 200-15
of X-mean, 200-20
Std Error
of Coefficient of Variation, 200-18
Stddev transformation, 119-16
StdRangeProb transformation, 119-11
StdRangeValue transformation, 119-11
Stem-leaf
depth, 200-27
leaf, 200-28
stem, 200-27
unit, 200-28
Stem-leaf plot
descriptive statistics, 200-27
Stephens test
circular data, 230-7
Stepwise regression, 311-1
Cox regression, 565-11
logistic regression, 320-17
multiple regression, 305-23
Poisson regression, 325-6
Storing results
linear regression, 300-35
multiple regression, 305-42
Stratification based on propensity scores, 124-1
Stratification of a database, 124-1
Stress
multidimensional scaling, 435-3
Stress A
parametric survival regression, 566-6
Stress B
parametric survival regression, 566-6
Stress plot
parametric survival regression, 566-19
Stress variable
parametric survival regression, 566-6
Student's t distribution
probablility calculator, 135-6
Studentized deviance residuals
Poisson regression, 325-5
Studentized Pearson residuals
Poisson regression, 325-5
Studentized range
one-way ANOVA, 210-5
Studentized range distribution
probablility calculator, 135-6
Studentized residuals
Poisson regression, 325-34
Studentized-range distribution
transformation, 119-11
Student's T distribution
simulation, 122-10
Style file
grid plot, 173-8

Style file
box plot, 152-13
histogram, 143-16
probability plot, 144-19
scatter plot, 161-22
Style files
multiple regression, 305-38
Subset of a database, 14-1
Subset selection
Cox regression, 565-11, 565-48
logistic regression, 320-17
multiple regression, 305-23, 305-32
Poisson regression, 325-6, 325-37
Subset selection report
multiple regression, 305-80
Subset selection tutorial
multiple regression, 305-79
Sum of exponentials
curve fitting, 351-9
growth curves, 360-8
Sum of functions models, 380-1
Sum of squares
multiple regression, 305-49, 305-55
Sum transformation, 119-16
Sunflower plot
scatter plot, 161-18
SUNSPOT dataset, 468-9, 472-7
Support services, 100-2
Surface charts, 140-1, 141-1
Surface plot
depth, 171-7
elevation, 171-6
perspective, 171-6
projection method, 171-7
rotation, 171-7
Surface plots, 140-10, 171-1
Survival
cumulative, 565-4
Survival analysis
Kaplan-Meier, 555-1
life-table analysis, 570-1
time calculator, 580-1
Weibull fitting, 550-1
Survival curves
Kaplan-Meier, 555-1
SURVIVAL dataset, 555-14, 555-37, 575-1, 575-5
Survival distribution
Cox regression, 565-2
Survival function
Kaplan-Meier, 555-2
Weibull fitting, 550-2
Survival plot
Kaplan-Meier, 555-35
Survival quantiles
Kaplan-Meier, 555-6, 555-30
SUTTON 22 dataset, 456-6, 456-14
SUTTON30 dataset, 455-6, 455-13
Symbol settings window, 181-1
Symmetric-binary variables
fuzzy clustering, 448-4
hierarchical clustering, 445-6
medoid partitioning, 447-2
Symmetry, 200-2, 206-25
Symphony exporting, 116-1
Syntax
macros, 130-2
SYS exporting, 116-1
Systat exporting, 116-1
Systat importing, 115-1
System requirements, 1-1

T

T distribution
simulation, 122-10
T2 alpha
data screening, 118-3
T2 Dataset, 405-3, 405-5, 405-10
T2 value, 410-7
Tables
descriptive, 201-1
Taguchi designs, 266-1
Tan transformation, 119-17
Tanh transformation, 119-17
Target specification, 250-20
Tarone-Ware test
Kaplan-Meier, 555-12
nondetects analysis, 240-3
Template, 105-1
default, 105-1
new, 105-1
open, 105-1
save, 105-2
saving, 8-5
Terms
multiple regression, 305-35
Text data, 102-1
Text functions transformations, 119-17
Text settings window, 182-1
Theoretical ARMA, 475-1
Thetas
theoretical ARMA, 475-2
Three-variable charts, 140-10
Threshold limit
Xbar R, 250-23
Tick label settings window, 186-1
Tick settings window, 185-1
Tickmarks, 185-1
Ties method
Cox regression, 565-17
Tile horizontally, 106-5
Tile vertically, 106-5
Time calculator, 580-1
Time format, 102-8
Time remaining
life-table analysis, 570-4
Time variable
Cox regression, 565-16
life-table analysis, 570-6
parametric survival regression, 566-4
TIMECALC dataset, 580-3
TNH(Z)
piecewise polynomial models, 365-6
Tolerance
multiple regression, 305-57
PC regression, 340-13
ridge regression, 335-17
Tolerance intervals, 585-1
Toolbar
customizing, 107-3
Topic search
goto window, 106-4
TOST
two-sample, 207-1
Tprob transformation, 119-11
TPT exporting, 116-1
Transformation
recoding, 3-4
Transformation operators, 119-4
Transformations, 3-1, 102-6, 119-1
Abs, 119-7
Arc sine, 119-17
Arc tangent, 119-17
ArCosh, 119-17
Arsine, 119-17
ArSinh, 119-17
ArTan, 119-17
ArTanh, 119-17
Average, 119-15
BetaProb, 119-8
BetaValue, 119-8
BinomProb, 119-8
BinomValue, 119-8
BinormProb transformation, 119-8
Collate, 119-12
conditional, 120-1
Contains, 119-17
CorrProb, 119-8
CorrValue, 119-8
Cos, 119-17
Cosh, 119-17
Cosine, 119-17
Count, 119-15
CsProb, 119-9
CsValue, 119-9
Cum, 119-7
date functions, 119-6
Day, 119-6
Exp, 119-7
ExpoProb, 119-9
ExpoValue, 119-9
Extract, 119-18
file function, 119-15
fill functions, 119-6

Fprob, 119-9
Fraction, 119-7
Fvalue, 119-9
GammaProb, 119-9
GammaValue, 119-9
HypergeoProb, 119-9
if-then, 120-1
indicator variables, 119-19
Int, 119-7
Join, 119-18
Julian date, 119-6
Lagk, 119-16
Lcase, 119-18
Ledk, 119-16
Left, 119-18
Length, 119-18
Ln(X), 119-7
Log, 119-7
LogGamma, 119-9
logic operators, 119-5
Logit, 119-7
Lookup, 119-14
mathematical functions, 119-7
Mavk, 119-16
Max, 119-16
Min, 119-16
Mod, 119-7
Month, 119-6
NcBetaProb, 119-9
NcBetaValue, 119-10
NcCsProb, 119-10
NcCsValue, 119-10
NcFprob, 119-10
NcFvalue, 119-10
NcTprob, 119-10
NcTvalue, 119-10
Negative binomial, 119-10
NegBinomProb, 119-10
Non-central Beta, 119-10
Non-central Chi-square, 119-10
noncentral-F distribution, 119-10
noncentral-t distribution
  transformation, 119-10
NormalProb, 119-10
NormalValue, 119-10
NormScore, 119-16
numeric functions, 119-6
PoissonProb, 119-11
probability functions, 119-8
RandomNormal, 119-11
random-number functions, 119-11
Rank, 119-16
rearrangement functions, 119-12
Recode, 119-15
recode functions, 119-14
recoding, 11-1
Remove, 119-18
Repeat, 119-18
Replace, 119-18
Right, 119-19
Round, 119-7
Sequence, 119-6
Ser, 119-6
Short, 119-7
Sign, 119-8
simulation, 15-1
Sin, 119-17
Sinh, 119-17
Smooth, 119-16
Sort, 119-12
Splice, 119-12
Sqrt, 119-8
Standardize, 119-16
statistical functions, 119-15
Stddev, 119-16
StdRangeProb, 119-11
StdRangeValue, 119-11
Studentized-range distribution, 119-11
Sum, 119-16
Tan, 119-17
Tanh, 119-17
text functions, 119-17
Tprob, 119-11
trigonometric functions, 119-17
Tvalue, 119-11
Ucase, 119-19
UnCollate, 119-13
Uniform, 119-11
Uniques, 119-13
UnSplice, 119-14
WeibullProb, 119-11
WeibullValue, 119-11
Year, 119-6
Transition type
  piecewise polynomial models, 365-6
Tricube weights
  linear regression, 300-13
Trigamma
  beta distribution fitting, 551-14
Trigonometric functions
  transformations, 119-17
Trim-mean
  descriptive statistics, 200-19
Trimmed
  descriptive statistics, 200-19
Trim-std dev
  descriptive statistics, 200-19
Tschuprow's T
  cross tabulation, 501-14
T-test
  1-Sample T2, 405-1
  assumptions, 205-22
  average difference plot, 205-20
  bootstrapping, 205-3
  histogram, 205-20
  kurtosis, 205-15
  multiplicity factor, 205-19
  nonparametric tests, 205-17
  normality, 205-15
  outliers, 205-22
  probability plot, 205-20
  skewness, 205-15
T-test of difference
  two proportions, 515-8
T-tests
  meta-analysis of means, 455-1
  one sample, 205-1
  paired, 205-1
  two-sample, 206-1
  two-sample (means/SDs), 207-1
Tukey's biweight
  multiple regression, 305-27
Tukey-Kramer test
  one-way ANOVA, 210-8
Tukey's lambda distribution
  simulation, 122-10
TUTOR dataset, 220-98
Tutorial
  general, 101-1
  linear regression, 300-37
Tvalue transformation, 119-11
Two correlated proportions, 520-1
Two independent proportions, 515-1
Two proportions, 515-1
Two sample t-test (from means/SDs), 207-1
Two-level designs, 260-1
Two-level factorial designs, 260-1
TWOSAMPLE dataset, 220-69, 220-72
Two-sample t-test, 5-1, 206-1
  assumptions, 206-18, 206-27
  bootstrapping, 206-3
  degrees of freedom, 206-13
TWOSAMPLE2 dataset, 220-70, 220-73
TWOSAMPLECOV dataset, 220-76
Two-variable charts, 140-4, 140-7
Two-way tables
  cross tabulation, 501-1
TXT exporting, 116-1
TXT importing, 115-1

U

Ucase transformation, 119-19
U-chart, 251-2
Unbiased std dev
  descriptive statistics, 200-17
UnCollate transformation, 119-13
Unconditional tests
  two proportions, 515-5
Undo, 103-4
Unequal variance t-test, 206-2
Uniform distribution
  simulation, 122-11
Uniform kernel
  Kaplan-Meier, 555-8

Weibull fitting, 550-34
Uniform transformation, 119-11
Uniformity test
circular data, 230-3
Uniques transformation, 119-13
Unknown censor
Cox regression, 565-18
Kaplan-Meier, 555-17
life-table analysis, 570-6
UnSplice transformation, 119-14
Unweighted means F-tests, 211-1
User written models, 385-1
UWM F-tests, 211-1
properties of, 211-1

V

Validation
Cox regression, 565-55
life-table analysis, 570-24
Validity
item analysis, 505-1
Value labels, 13-1, 102-10
Variable
data type, 102-10
format, 102-6
labels, 102-6
names, 101-1, 102-5
numbers, 102-5
transformations, 102-6
Variable format, 102-6
Variable info, 102-5
tutorial, 101-2
Variable info file, 102-1
Variable info sheet, 102-1
Variable info tab, 2-4
Variable labeling, 2-4
Variable labels, 102-6
Variable matching, 123-3
Variable name, 2-4
Variable names, 102-5
rules for, 2-5
Variable numbers, 102-5
Variable selection, 310-1
Cox regression, 565-11
logistic regression, 320-17
multiple regression, 305-23
Poisson regression, 325-6
principal components analysis, 425-8
Variables
naming, 101-2
Variables charts, 250-1
Variance
descriptive statistics, 200-15
linear regression, 300-5
multiple regression, 305-13
Variance components
R & R, 254-3, 254-11
Variance inflation factor
multiple regression, 305-8, 305-57
PC regression, 340-12
ridge regression, 335-16
Variance inflation factor plot
ridge regression, 335-19
Variance inflation factors
ridge regression, 335-2
Variance ratio test, 206-19
Variance test
equal, 206-19
linear regression, 300-50
Variances
equality of, 206-20
testing equality of multiple, 210-18
Variates
canonical correlation, 400-1
Varimax rotation
factor analysis, 420-4
principal components analysis, 425-7
VIF
multiple regression, 305-8
ridge regression, 335-2
Violin plot
density trace, 154-1
Violin plots, 140-6, 154-1
Von Mises distribution
circular data, 230-5

W

W mean
appraisal ratios, 485-8
Wald method
correlated proportions, 520-4
Wald statistic
Poisson regression, 325-26
Wald test
Cox regression, 565-11, 565-33
logistic regression, 320-9
Walter's confidence intervals
two proportions, 515-22
Ward's minimum variance
double dendrograms, 450-3
hierarchical clustering, 445-4
Watson & Williams test
circular data, 230-7
Watson test
circular data, 230-4
Watson-Williams F test
circular data, 230-10
WEIBULL dataset, 550-12, 550-27, 550-44, 552-3, 552-12, 555-27
Weibull distribution
probablility calculator, 135-6
simulation, 122-12
Weibull fitting, 550-6
Weibull fitting, 550-1
Weibull model
curve fitting, 351-7
growth curves, 360-6
Weibull probability plot, 144-17
Weibull regression, 566-1
WEIBULL2 dataset, 144-17
WeibullProb transformation, 119-11
WeibullValue transformation, 119-11
Weight variable
linear regression, 300-25
multiple regression, 305-28
WEIGHTLOSS dataset, 220-85
WESTGARD dataset, 252-9
Westgard rules, 252-1
Westlake's confidence interval, 235-6
Whiskers
box plot, 152-5
Wilcoxon rank-sum test, 206-1, 206-20
Wilcoxon signed-rank test, 205-18
Wilcoxon-Mann-Whitney test
cross-over analysis using t-tests, 235-8
Wilks' lambda
canonical correlation, 400-10
discriminant analysis, 440-2
MANOVA, 415-2
Wilson score limits
one proportion, 510-2
Wilson's score
correlated proportions, 520-3
two proportions, 515-19
Window
data, 7-1
output, 9-1
Windows
navigating, 1-4
Winters forecasting
exponential smoothing, 467-1
Within factor
repeated measures, 214-9
Within subject
repeated measures, 214-2
WK exporting, 116-1
WKQ exporting, 116-1
Woolf's odds ratio analysis
Mantel-Haenszel test, 525-11
Word processor, 9-1
Working-Hotelling C.I. band
linear regression, 300-6
Working-Hotelling limits
linear regression, 300-60
WR1 exporting, 116-1
WRK exporting, 116-1

X

Xbar chart, 250-1
Xbar R chart, 250-1
XLS exporting, 116-1

Y

Year format, 102-8
Year transformation, 119-6
Yule-Walker
  automatic ARMA, 474-1

Z

Zero time replacement
  beta distribution fitting, 551-3
  cumulative incidence, 560-4
  gamma distribution fitting, 552-4
  parametric survival regression, 566-4
  Weibull fitting, 550-13
ZHOU 175 dataset, 545-33
ZINC dataset, 345-15