

StatView® II

StatView® SE+*Graphics*

The Solutions for Data Analysis and Presentation Graphics

Developed by:

Daniel S. Feldman, Jr.

Jim Gagnon

Rich Hofmann

Joe Simpson

The proper citation for this software and manual is:

Abacus Concepts, StatView II. (Abacus Concepts, Inc., Berkeley, CA, 1987).

This manual is printed on recycled paper.



LICENSE. ABACUS CONCEPTS hereby grants to you a nonexclusive license to use the enclosed computer Program subject to the terms of this Agreement. THE PROGRAM AND ANY BACKUP COPIES MAY BE USED ONLY ON THE SINGLE COMPUTING MACHINE OWNED BY YOU AND FOR YOUR OWN PURPOSES.

COPYRIGHT. This Program and accompanying manual are copyrighted and contain proprietary information. All rights reserved. This Program and manual may not, in whole or in part, be copied, photocopied, reproduced, translated or reduced to any electronic medium or machine readable form without the express consent, in writing, of ABACUS CONCEPTS. Duplication of the accompanying diskette is for the sole use of the original purchaser. WILLFUL VIOLATIONS OF THE COPYRIGHT LAW OF THE UNITED STATES CAN RESULT IN CIVIL DAMAGES OF UP TO \$50,000 IN ADDITION TO ACTUAL DAMAGES, PLUS CRIMINAL PENALTIES OF UP TO ONE YEAR IMPRISONMENT AND/OR A \$10,000 FINE.

Program Copyright © 1987 - 1991 Abacus Concepts, Inc.

Manual Copyright © 1990-91 Abacus Concepts, Inc. All rights reserved. Fifth edition.

LIMITED WARRANTY FOR RECORDING MEDIA. ABACUS CONCEPTS or its distributor or agent will repair or replace free of charge any defective recording medium on which any Software product is recorded if the medium is returned to ABACUS CONCEPTS or its distributor or agent by the original customer within ninety (90) days after purchase of this License. This warranty does not cover defects due to accident, abuse, service or modification by any unauthorized person, or any cause occurring after initial delivery of the medium to Licensee. THIS WARRANTY GIVES YOU SPECIFIC LEGAL RIGHTS, AND YOU MAY ALSO HAVE OTHER RIGHTS WHICH VARY FROM STATE TO STATE.

LIMITATION OF IMPLIED WARRANTIES. ALL IMPLIED WARRANTIES WITH RESPECT TO THE RECORDING MEDIUM, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, ARE LIMITED IN DURATION TO NINETY (90) DAYS FROM THE DATE OF RETAIL PURCHASE OF THIS LICENSE. SOME STATES DO NOT ALLOW LIMITATIONS ON HOW LONG AN IMPLIED WARRANTY LASTS, SO THE ABOVE LIMITATION MAY NOT APPLY TO YOU.

DISCLAIMER OF WARRANTY FOR SOFTWARE. Even though ABACUS CONCEPTS has tested and reviewed the software and documentation, ABACUS CONCEPTS' SOFTWARE IS LICENSED ON AN "AS IS" BASIS. THIS MEANS THAT THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE SOFTWARE IS ASSUMED BY LICENSEE. EXCEPT AS MAY BE PROVIDED OTHERWISE IN THE AGREEMENT, SHOULD THE SOFTWARE PROVE DEFECTIVE FOLLOWING ITS PURCHASE, LICENSEE, AND NOT ABACUS CONCEPTS OR ITS AUTHORIZED DISTRIBUTORS OR AGENTS, ASSUMES THE ENTIRE COST OF ALL NECESSARY SERVICE, REPAIR, OR CORRECTION. ABACUS CONCEPTS DISCLAIMS ALL IMPLIED WARRANTIES FOR THE SOFTWARE, INCLUDING WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. ABACUS CONCEPTS MAKES NO REPRESENTATIONS CONCERNING THE QUALITY OF THE SOFTWARE AND DOES NOT PROMISE THAT THE SOFTWARE WILL BE ERROR FREE OR WILL OPERATE WITHOUT INTERRUPTION. SPECIFICATIONS OF THE SOFTWARE INCLUDING THE AMOUNT OF MEMORY, OR TIME REQUIRED FOR EXECUTION OF ANY PROGRAM MAY BE CHANGED IN NEW RELEASES AND VERSIONS.

LIMITATION OF LIABILITY. IN NO EVENT WILL ABACUS CONCEPTS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL OR OTHER DAMAGES ARISING OUT OF THE USE OF THE RECORDING MEDIUM OR THE SOFTWARE BY ANY PERSON, WHETHER OR NOT INFORMED OF THE POSSIBILITY OF DAMAGES IN ADVANCE. ABACUS CONCEPTS' TOTAL LIABILITY WITH RESPECT TO ALL CAUSES OF ACTION TOGETHER WILL NOT EXCEED AMOUNTS PAID BY LICENSEE TO ABACUS CONCEPTS FOR THIS LICENSE. THESE LIMITATIONS APPLY TO ALL CAUSES OF ACTION, INCLUDING BREACH OF CONTRACT, BREACH OF WARRANTY, ABACUS CONCEPTS' NEGLIGENCE, STRICT LIABILITY, MISREPRESENTATION AND OTHER TORTS. SOME STATES DO NOT ALLOW THE EXCLUSION OF LIMITATION OF INCIDENTAL OR CONSEQUENTIAL DAMAGES, SO THE ABOVE LIMITATION OR EXCLUSION MAY NOT APPLY TO YOU.

TERM AND TERMINATION. This agreement shall continue in force until terminated. Failure of the customer to abide by the terms of this license agreement, in particular the prohibition against unauthorized reproduction of the program and/or manual will terminate this agreement and result in withdrawal of technical support, and forfeiture of all rights under this Customer User Agreement.

GOVERNING LAW. This Agreement shall be governed by the laws of the State of California.

Abacus Concepts, Inc.
1984 Bonita Avenue
Berkeley, CA 94704
(415) 540-1949

The following names are trademarked products of the corresponding companies. Apple, ImageWriter, LaserWriter, Macintosh and MultiFinder: Apple Computer, Inc. MacDraw and MacPaint: Claris Corporation. Microsoft and MS-DOS: Microsoft Corporation. 4th Dimension: Axiis, Inc. dBase: Ashton-Tate Corporation. Lotus 1-2-3: Lotus Development Corporation. SAS: SAS Institute, Inc. PixelPaint: SuperMac Software. PageMaker: Aldus Corporation.

ISBN: 0-944800-00-9

Acknowledgements

Several people helped create StatView. Rich Hofmann and Joe Simpson provided expert advice on all aspects of the program's operation and statistics. They also authored the sections in this manual covering the statistics. Emily Roberson offered opinions, insights, and ideas from the program's inception to its completion. Paul Hoffman guided the manual through its production as well as writing many sections. Samantha Sager edited the manuscript. Dr. Elaine B. Feldman helped interpret the Lipid Data. Consuelo Lorenzo designed the package. Kathy Krieger provided patience and moral support throughout the development of the various versions of StatView.

Table of Contents

Introduction

System Requirements	2
What You Need to Use StatView II and StatView SE+Graphics.....	2
Terminology	3
Printing.....	4
Working with Other Applications.....	4
Sample Data.....	5
Technical Support	5

Chapter 1 — Quick Start

Installation.....	7
Creating StatView Datasets.....	8
Defining a New Dataset	8
Entering Data in Your Dataset.....	10
Descriptive Statistics.....	10
Opening a Dataset	11
Selecting Columns	11
Performing an Analysis.....	12
Analyzing Additional Columns.....	12
Graphs	13
Creating a Graph	13
Customizing Parts of Your Graph.....	14
Adding a Regression Line	14
Make Subset by Gender.....	15
Adding Display Features to Your Graph.....	16
Adding a Caption	17
New Columns.....	18
Where To Go Next	19

Chapter 2 — Datasets

The Data Window	21
Datasets	22
Columns and Rows	22
Opening, Creating and Altering Datasets	23
Category Data	26
New Categories.....	27
Editing Categories	28
Entering and Editing Data.....	29
Missing Values	29
Selecting Data	30
Cutting, Clearing, and Deleting Data	31
Copying to the Clipboard.....	32
Pasting into StatView	32
Pasting Records.....	32

Pasting Cells	32
Pasting Different Sizes and Types.....	33
Importing Data	34
Methods of Importing Text Files	34
Text File Format	35
The Import Command	35
How StatView Imports Datasets	36
Example of Importing	37
Importing from Excel and Other Applications	38
Exporting Data and Pictures	39
Exporting Data with Text Files.....	39
Exporting Data and Results Through the Clipboard	39
Sorting Datasets	41
Printing.....	41

Chapter 3 — Introduction to Graphs and Analyses

Assigning Variables	43
Assigning with the Mouse and Menus.....	44
Quick Assignment Command	44
Creating Graphs and Tables.....	45
Automatic Updating	45
Modifying Views	45
Graphing Data	46
Statistical Analysis	47
Variable Assignments	47
Grouped Data.....	48
Views of Analyses	48
Multiple Analyses	49
Missing Values.....	49
Calculation.....	50
Interactive Analysis	50
Excluding Rows	50
Ranges	50

Chapter 4 — Graphs and Drawing with StatView

Graphs You Can Make	56
Univariate Chart.....	56
Percentile Plot (Cumulative Frequency Curve).....	58
Scattergram.....	60
Comparison Percentile Chart.....	63
Line Chart	64
Bar Chart.....	66
Histogram.....	67
z-Score Histogram	68
Comparative Bar Chart.....	68
Pie Chart	69
Error Bars.....	70
Box Plots	71
Modifying Graphs	73
Frame.....	74
Point Overlap	74
Subset Specify.....	78
Error Bars.....	79
Standard Deviation Bands	80

Confidence Bands.....	80
Outlier	81
Notch	81
Percentile.....	81
Bar	81
Equal Axes.....	82
No Symbols	82
Composite and Paging Graphs.....	83
Layout of StatView Graphs.....	84
Resizing	84
Preferences.....	86
Colors	87
Drawing Tools.....	87
Selecting	88
Cut, Copy, and Paste.....	89
Rulers and Grids.....	91
Text.....	92
Drawing Objects.....	93
Changing the Axes.....	95
The Legend	96
When Customizations Disappear.....	97
StatView's Drawing Features	97

Chapter 5 — The Describe Menu

Numeric Descriptive Statistics For A Single Variable.....	102
Measures of Central Tendency and Skewness.....	103
Variance and Standard Deviation	103
Maximum, Minimum and Range.....	103
Kurtosis	104
Coefficient of Variation	104
Percentiles.....	105
Standard Error of the Mean	106
Geometric Mean.....	106
Harmonic Mean	107
Graphic Description of Data For A Single Variable.....	107
Standard Deviation Error Bars	107
Univariate Scattergrams	108
Confidence Intervals	108
z-Score Distribution.....	110
Box Plot	111
The Cumulative Frequency Curve.....	113
Summary	117
Graphic Comparisons	117
Making Descriptive Comparisons Around the Mean.....	117
Summary	125
Frequency Distribution	125
Frequency Distribution of Continuous Data Variables	125
Frequency Distribution of Category Data Variables.....	129
Summary	130

Chapter 6 — The Compare Menu

Compare Percentiles.....	133
t-Test	135
One Group t-Test	136

Paired Two Group t-Test.....	137
Unpaired Two Group t-Test.....	138
Correlation Coefficient.....	139
Correlation Coefficient	140
Correlation Matrix	141
Saving a Correlation Matrix as a Datafile	142
Regression	143
Simple Regression	145
Multiple Regression	149
Polynomial Regression	151
Stepwise Regression	157
Factor Analysis.....	162
Parameters.....	163
Tabular Views	166
Graphic Views.....	175
ANOVA	177
Assigning Variables.....	178
Single Factor Factorial, Non-Repeated Measure.....	178
Two Factor Factorial, Non-Repeated Measures — Balanced Model	181
Two Factor Factorial, Non-Repeated Measures — Unbalanced Model	183
Single Factor Factorial — One Repeated Measure	184
Three Factor Factorial — One Repeated Measure	187
Contingency Tables and Cross-tabs.....	189
Chi-Square — One Group	190
Contingency Table.....	190
Nonparametrics.....	192
Mann-Whitney U	193
Wilcoxon Signed-Rank	194
Spearman Rank Correlation Coefficient.....	195
Kendall Rank Correlation Coefficient	196
Kolmogorov-Smirnov Tests	196
Wald-Wolfowitz Runs	197
Kruskal-Wallis Test	198
Friedman Test.....	199

Chapter 7 — Advanced Column Creation

Transform.....	201
Formula	202
Recode	204
Continuous Data to Category Data.....	204
Missing Values to Specified Value.....	206
Range of Values to Specified Value	207
Series.....	208
Splitting Columns.....	209

Appendix A — StatView Memory Limits.....	213
--	-----

Appendix B — Formulae and References.....	215
---	-----

Index.....	229
------------	-----

Introduction

StatView SE+*Graphics* and StatView II are statistical analysis and presentation graphics programs. This manual describes both StatView II and StatView SE+*Graphics*. StatView SE+*Graphics* is designed to run on a Macintosh without a floating-point math coprocessor. StatView II is designed to run on a Macintosh with a floating-point math coprocessor.

Throughout this manual, both products will be called "StatView". The differences you will find between the two programs are in the hardware required for each and in their use of color. Although this manual talks about "StatView", this does not refer to previous versions of the StatView line – our StatView and StatView 512+ products. If you have StatView or StatView 512+, please call Abacus Concepts at (800) 666-STAT for information on how to upgrade to the latest versions of StatView.

With StatView, analyzing data is simple: enter your data, select the variables of interest, choose the statistical test to perform, and view the results of the analysis. You can accomplish this without having to learn a difficult command language or typing in command scripts. Because your analysis is not complete until you present your results, StatView includes a complete graphing and drawing package allowing you to analyze your data and generate publication quality results in a single application.

Because data analysis is an iterative process, StatView allows you to quickly and easily explore your data. You can change data values, eliminate outliers, examine subsets, select new variables, choose new statistics, and the program will automatically recompute your results and redraw your graphs. This process can be accomplished in seconds, because StatView performs its tasks at speeds equal to or greater than mainframe computers.

Since data analysis encompasses more than just producing values, StatView includes a complete presentation graphics package. You can create a variety of graphic views of your data and customize them to produce your final output. StatView contains all the tools you need to add emphasis to your graphs. These tools include text, arrows and shapes, choice of fill patterns, text fonts, text style, text size, and more. StatView SE+*Graphics* and StatView II use the more than 16 million colors available on the Macintosh.

With StatView, one program allows you to manage and analyze your data, and then produce presentation quality output which clearly and concisely expresses your results and conclusions.

System Requirements

There are two versions of StatView: StatView SE+*Graphics* and StatView II. Both have minimum requirements for your Macintosh hardware and software. Both programs require the following:

- a minimum of 1 megabyte of main memory (RAM)
- Macintosh Operating System version 4.2 or later
- a hard disk (strongly preferred) or two 800K floppy drives

The difference between the two is related to the kind of Macintosh you use; StatView II requires a math coprocessor; StatView SE+*Graphics* will not take advantage of the coprocessor. Be sure you have the correct version for your machine.

If you have any questions regarding your Macintosh's ability to run StatView or if you have purchased the incorrect version, please contact Abacus Concepts.

Note: If you plan to use StatView on two separate Macintoshes which require different versions of the StatView software, you can purchase a license to run StatView on both machines. Please contact Abacus Concepts at (415) 540-1949 for more information.

What You Need to Use StatView II

StatView II requires a floating-point math coprocessor (FPU) and consequently will operate only on the following Macintosh systems:

- a Macintosh II, IIfx, IICx, IICI, IIcx, SE/30, or IISI. The IISI must contain the optional FPU available from Apple.
- a Macintosh SE, Macintosh Plus, Macintosh Portable, or any Macintosh with a 3rd party accelerator board containing a 68020 (or later) CPU and 68881 (or later) floating-point math coprocessor.

What You Need to Use StatView SE+Graphics

StatView SE+*Graphics* does not require a floating-point math coprocessor. It is designed for the following Macintosh systems:

- a Macintosh Classic, Macintosh Plus, Macintosh SE, Macintosh Portable or Macintosh LC. It will also run on the IISI if you have not purchased the optional FPU.
- a Macintosh 512k enhanced with 1 or more megabyte of RAM

Recommendations

We recommend that your Macintosh have a hard disk and that you operate StatView from that hard disk.

Terminology

Please note that StatView does not require a large screen display. However, a larger screen is very useful since it allows you to display more data and manipulate larger graphs, thereby increasing your efficiency.

StatView is completely compatible with MultiFinder. Since StatView allows full cut, copy and paste of both data and graphics, using it with other applications within MultiFinder will increase your productivity.

This manual assumes that you are familiar with standard Macintosh terminology and operation. If you are unfamiliar with the Macintosh terms and actions used in this manual consult the your Macintosh owner's guide.

StatView contains several hardware-dependent features. The following StatView terms are used in this manual:

Term	Meaning
color systems	All Macintosh IIs (and future machines) with monitors set to 16 or more colors or shades of gray.
non-color systems	Black-and-white Macintoshes with monitors set to less than 16 colors or grays.
old QuickDraw	The release of QuickDraw that is available on the Macintosh Plus and the Macintosh SE. It supports only 8 colors: black, red, green, yellow, blue, magenta, cyan, and white.
numeric co-processor	A dedicated hardware chip which speeds up numerical computations. The Macintosh uses either the model 68881 or 68882.
small screen system	A system with a screen width less than 640 pixels. This is the size of a Macintosh Plus or SE screen.
large screen system	A system with screen width greater than or equal to 640 pixels. This is the size of the standard Macintosh II monitor.

StatView and Your Hardware

Four items on the Menu Bar will appear differently depending on your Macintosh screen size:

Large screen system	Small screen system
Variables	Vars
Describe	Desc
Compare	Comp
Window	Σ

The Color choice in the Graph menu will change depending on whether the Macintosh you are using is a color or non-color system. On a non-color system, the

Printing

Color choice will contain a list of 8 colors: black, red, green, yellow, blue, magenta, cyan, and white. On a color system, the Color choice will display a palette of 16 or 32 colors, depending on whether your monitor is using 16 or 256 colors. You can customize these palettes to include any of the Macintosh 16 million colors.

On a non-color system, objects will appear in black on your monochrome screen; however, they will appear in color when they are printed. To find out the color of an object, select the object and then choose the Color command from the Graph menu. The object's color will be checked in the Color menu.

StatView will work with any printer, color plotter and slide maker for which Macintosh drivers are available. Printing quality depends on the type of printer and driving software. StatView exploits the full capabilities of the Apple ImageWriter and LaserWriter families.

StatView constructs its tables, text, and graphs in a resolution-independent manner; therefore, printouts are generated to the full resolution of your printer.

There are several options available for getting hard copy output of your StatView color graphs. You can produce color output by printing with an ImageWriter II or ImageWriter LQ using a color ribbon. Since a Macintosh II can generate more colors than most printers, the colors you print will often be an approximation of the colors on your screen. Any color plotter or slide making hardware that is compatible with the Macintosh will also produce color plots of your graphs.

Working with Other Applications

Data generated as ASCII files from other Macintosh applications, as well as from applications on other computers, can be directly imported to StatView. Data from StatView can also be saved as ASCII text files and transferred to other applications. You can also use the Clipboard to transfer data to and from other Macintosh applications. Importing and exporting data is described in Chapter 2.

Graphs created by StatView can be exported to other Macintosh applications. You can copy StatView graphs onto the Clipboard and paste them into other applications (such as Microsoft Word, MacWrite, Pixel Paint, MacDraw and PageMaker) or store them in the Scrapbook.

Your analysis results can also be transferred to other applications. You can copy analysis tables as a PICT into the Clipboard and transfer them as a picture to other applications, or you can copy the values as text to move into spreadsheets, word processors, or other analysis packages.

StatView SE+*Graphics* and StatView II can open and analyze data files built by earlier version of StatView (StatView 512+ and the original StatView). In addition Abacus Concepts' general linear modeling package, SuperANOVA, can read and write StatView data files.

Sample Data

Sample datasets that come with StatView are used throughout this manual. The discussion of the descriptive statistics and several of the comparative statistics uses the Lipid Data dataset. This dataset was provided by Dr. Terence T. Kuske, Professor of Medicine and Associate Dean for Curriculum, Medical College of Georgia, Augusta, GA.

The data are blood lipid screenings of medical students at the Medical College of Georgia. Blood lipid levels and other cardiovascular risk factors (cigarette smoking, hypertension, family history of coronary heart disease) are evaluated in freshman students and later when they are seniors. This program personalizes education in lipid metabolism, prevention of cardiovascular disease, and management of hyperlipidemia by diet and medication.

Lipids include cholesterol and triglycerides and their lipoprotein carriers in blood, very-low, low, and high density lipoproteins (VLDL, LDL, and HDL). Cardiovascular risk is increased proportional to all these parameters except for an inverse relationship to HDL cholesterol. This study measures cholesterol, triglycerides and HDL cholesterol. A factor allows an estimate of VLDL cholesterol from the triglycerides value. Subtraction of VLDL and HDL cholesterol from total cholesterol yields a calculated LDL cholesterol value.

"Healthy" values for adults are:

<200 mg/dl	Total cholesterol
<130 mg/dl	LDL cholesterol
>50 mg/dl	HDL cholesterol
<150 mg/dl	Triglycerides

Dietary modification is indicated when cholesterol values exceed 200. Dietary modification with medication is indicated for cholesterol, LDL and triglycerides that exceed 240, 190 and 350 respectively, especially when other risk factors are present. Exercise can raise HDL.

The discussion on the ANOVA analyzes three datasets from Winer, B. J., *Statistical Principles in Experimental Design*, © 1971, McGraw-Hill, New York, New York and one dataset from Afifi, A. and Azen, S., *Statistical Analysis: A Computer Oriented Approach*, © 1979, Academic Press, Orlando, Florida

The discussion on Factor analysis analyzes the Eight Physical Variables data from Harmon, *Modern Factor Analysis*, © 1967, University of Chicago Press, Chicago, Illinois.

Technical Support

The services of Abacus Concepts' Technical Support staff are available to registered owners of StatView. You must fill out the registration card included with your

manual and mail it to the company in order to be eligible for technical support. Please provide us with your serial number when you contact us by mail or phone.

General Tips

Read through the pertinent sections in this manual if you are having difficulty using StatView. Practice the procedure with one of StatView's sample datasets and run through a few of the examples.

If you think you've found a bug...

If StatView crashes somewhat randomly, try booting off a clean system disk. This will reveal whether the difficulty is simply an incompatibility between StatView and an INIT or CDev in your system.

- Turn off your machine.
- Insert the System Tools disk that came with your Macintosh and restart. (If you are running system 7.0, you can create a minimal start-up disk on a single high-density disk.)
- Check that the top-most icon on the desktop is the System Tools disk.
- Drag the System document from the system folder on your hard disk to the desktop.
- Run the application now using the "clean" system.
- Before you Restart, drag the System document back into the System folder on your hard disk.
- If the problem disappeared using the clean system, you can correct it by removing the incompatible INIT or CDev from your hard disk.

When all else fails...

Call us between 8:30 am and 4:00 pm Pacific time at (415) 540-1949 and ask the receptionist for technical support. Please have the following information available:

1. Your serial number (from the back of your manual or program diskette)
2. The model of Macintosh and version of the operating system you are using
3. The amount of RAM in your Macintosh

Chapter 1 — Quick Start

This chapter shows you how to get started with StatView. You will find that StatView's features are straightforward to use and let you perform complex statistical analyses and generate presentation quality output easily. The ideas you learn in this chapter will enable you to start performing analyses on your own data immediately.

Installation

If you have a hard disk, install StatView by copying the program and the sample data from the master disk to your hard disk. If you are running on a two-floppy system, copy the master disk to an empty floppy disk. Put the master disk in a safe place in case you need to copy from it again in the future. Never run StatView from the master disk. Please make a note of the serial number located on the rear of your master disk. You will need to refer to this number if you need to call for technical support.

Overview

StatView datasets have many features that make statistical analysis easy. Columns can contain simple data or be transformations of other columns. You can import and export data on disk or through the Clipboard.

Each dataset is laid out in rows and columns, similar to spreadsheet programs such as Excel. Each column is a variable that you can analyze. When you set up your dataset, you specify the type of variable that will be in each column and the way you want the data in each column to be displayed. Each row is a record, such as a patient or a run of an experiment.

Most columns are simple data that you enter by hand in each cell. If you wish, you can import data from text files or from the Clipboard. You can also create columns that are transformations of other columns.

To perform an analysis you denote the columns you are interested in analyzing as either an X or a Y. This tells StatView which variables you are interested in and how you plan to use them in your analysis. You then select a statistical test from the **Describe** or **Compare** menus and select a view of your results from the **View** menu.

StatView has two basic types of analyses:

- Descriptive statistics (descriptions of one column)
- Comparative statistics (comparisons of two or more columns)

You can display the results of these statistics either as tables of text or as graphs. To make it easy for you to remember how to perform each analysis, descriptive statistics are all found in the **Describe** menu, and comparative statistics are found

Creating StatView Datasets

in the **Compare** menu. You specify whether you want tables or graphs in the **View** menu.

Your analyses are presented in a *view window* which is a different window than your dataset. Like most other Macintosh applications, you can have many windows open at once. However, each dataset only has one view window.

As you read this chapter, you should perform the steps at the same time on your Macintosh. The steps shown here are the same as those that you will use in your work. As you will see, getting tabular and graphic results from StatView is quick and easy, allowing you to perform more analyses in less time.

Creating a new dataset takes two simple steps:

- Define the structure of the columns
- Enter the data

Assume that you wish to create a dataset with three columns: patient name, gender, and age.

Defining a New Dataset

- Select **New** from the **File** menu. You see the following:

New Data Column Information

Name:

Type: ☒ integer ☐ real ☐ long
 ☐ category ☐ string

Decimal Places:
☐ 0 ☐ 1 ☐ 2 ☒ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

- Type **Patient** in the **Name** choice.
- Click the **string** button to indicate that the patient name is a string of characters.
- Click **More** to indicate that you have more columns to define.
- Type **Gender** for the name of the second column.
- Click the **category** button since you plan to enter categorical information (male and female) to denote the patient's gender.

- Click **More** to indicate that you have more columns to define. Before you go on, however, StatView wants you to define the category you are using for Gender. You see:

Please choose the new column's category.

Select

New

File

Cancel

Untitled-1

- Click **New** to create a new category. You see:

Create a New Category

Category Name:

Element Name:

Add Replace Delete

Done Cancel

File

Untitled-1

- Type **Gender** for the **Category Name**.
- Press the **Tab** key.
- Type **Male** for the first **Element Name**.
- Click **Add** or press the **Return** key.
- Type **Female** for the second **Element Name**.
- Click **Add**.
- Click **Done** to indicate that there are no more elements in the Gender category.

You are returned to the New Data Column Information dialog box.

- Type **Age** for the **Name** of the third column.
- Click **integer** for the type of this column.
- Click **Done** to indicate that you are finished defining the dataset. StatView opens a window for your new dataset:

Opening a Dataset

- Select **Open** from the **File** menu. Choose the **Sample Data** folder, then select **Lipid Data**.
- Select **Zoom Up** from the **Window** (or Σ) menu to fill the entire screen with this dataset:

Lipid Data						
	Name	Gender	Age	Weight	Cholesterol	Trigly
1	J. Suds	male	22	138	197	
2	T. Wilson	female	22	115	181	
3	D.S. Quintent	male	22	190	190	
4	R. Beal	female	22	115	131	
5	R. James	male	25	160	172	
6	S. Kaufman	male	22	150	233	
7	M. Mubroid	male	23	154	194	
8	L. Phote	male	24	185	155	
9	C. Norman	male	23	178	234	
10	R.S. Smith Jr.	male	22	158	201	
11	Walker	male	26	188	258	
12	W. Rogers	male	22	150	212	
13	M. Lumpole	male	22	123	137	
14	D. Fineman	female	27	138	285	
15	R. Smith	male	22	143	218	

This dataset contains blood lipid screenings of medical students at the Medical College of Georgia. This example will compute basic descriptive statistics on the students' cholesterol and weight.

Selecting Columns

To run analyses, you must assign an X or a Y to the columns you want to look at. You can do this from the menu or with the mouse.

From the menu:

- Click on the column heading of the **Cholesterol** column. This selects the column.
- Select **Choose X** from the **Variables** menu or press **Command-U**. X_1 appears under the column name.

With the mouse:

- Double-click on the column heading of the **Cholesterol** column. X_1 appears under the column name.

Lipid Data						
	Name	Gender	Age	Weight	Cholesterol	Trigly
1	J. Suds	male	22	138	197	
2	T. Wilson	female	22	115	181	
3	D.S. Quintent	male	22	190	190	
4	R. Beal	female	22	115	131	
5	R. James	male	25	160	172	
6	S. Kaufman	male	22	150	233	
7	M. Mubroid	male	23	154	194	
8	L. Phote	male	24	185	155	
9	C. Norman	male	23	178	234	
10	R.S. Smith Jr.	male	22	158	201	
11	Walker	male	26	188	258	
12	W. Rogers	male	22	150	212	
13	M. Lumpole	male	22	123	137	
14	D. Fineman	female	27	138	285	
15	R. Smith	male	22	143	218	

Performing an Analysis

- Select Mean, Std. Dev., etc... from the Describe menu. A check mark appears next to the command indicating that this analysis is selected.
- Select Confidence Intervals from the Describe menu. This selection is also checked. StatView can perform many analyses at the same time. A dialog appears that allows you to select the distribution and confidence intervals that will be displayed.
- Click OK to specify the t distribution and 95% confidence interval. The view window appears on top of your dataset window. The table in the window shows the analyses you described:

View of Lipid Data						
X1: Cholesterol						
Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:	
191.232	35.674	3.66	1272.648	18.655	95	
Minimum:	Maximum:	Range:	Sum:	Sum of Sqr.:	Missing:	
115	285	170	18167	3593733	0	
t 95%:	95% Lower:	95% Upper:				
7.267	183.964	198.499				

Analyzing Additional Columns

StatView can perform the same analysis on many columns at the same time. When you specify another column as an X variable, the view window shows the analyses for both X variables.

- Bring the Lipid Data window to the front by clicking on it or by selecting Lipid Data from the window menu.
- Double-click on the column heading of the Weight column. X₂ appears under the column name.

Lipid Data					
	Name	Gender	Age	Weight	Cholesterol
				X2	X1
1	J. Suds	male	22	138	197
2	T. Wilkan	female	22	115	181

Select the view window; it now shows the same table for the Weight column. You can select the view window in two ways:

- Click anywhere on the view window if part of it is visible, or
- Select View of Lipid Data from the Window menu.

View of Lipid Data					
X2: Weight					
Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
158.653	28.389	2.913	805.931	17.894	95
Minimum:	Maximum:	Range:	Sum:	Sum of Sqr.:	# Missing:
107	234	127	15072	2466970	0
t 95%:	95% Lower:	95% Upper:			
5.783	152.87	164.436			

Note that the scroll bar on the right of this window is now active. If you scroll up, you will see the same table for the Cholesterol column.

Graphs

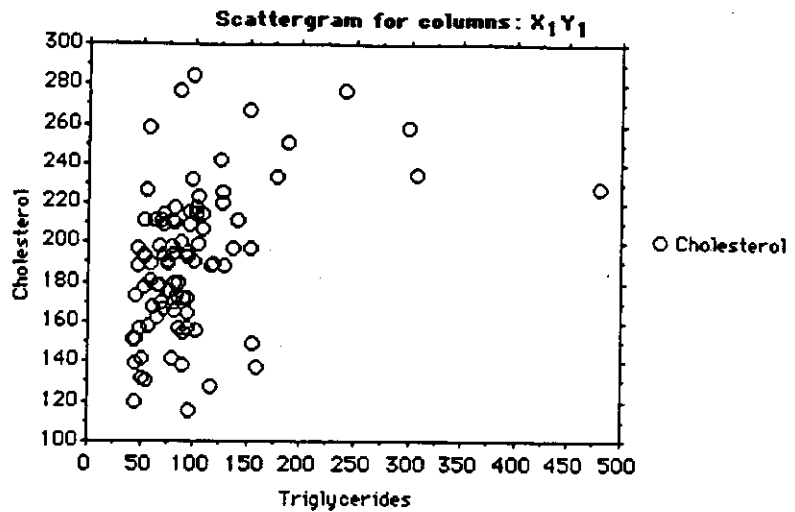
Creating graphs in StatView is as easy as creating statistics tables. Graphs are simply different views of your data. This example examines whether there is a dependency between the Cholesterol and Triglyceride values of the medical students.

Creating a Graph

- Select the dataset window.
- Select both the Weight and Cholesterol columns by clicking and dragging the cursor across the column heads.
- Select Clear X&Y from the Variables menu. This removes the X and Y indications from those columns.
- Select None from the Describe menu.

To first examine the data, you will generate a scattergram visually comparing the Cholesterol and Triglycerides data.

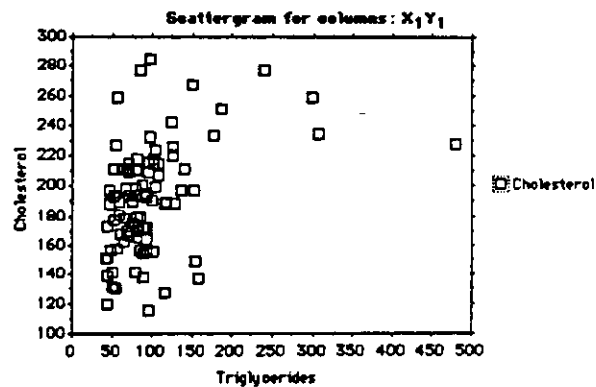
- Select the Cholesterol column.
- Select Choose Y from the Variables menu. Y₁ appears under the column name.
- Double-click on the column heading of the Triglycerides column. X₁ appears under the column name.
- Select Scattergram from the View menu. The view window changes to:



Customizing Parts of Your Graph

StatView gives you many choices in determining how your graph looks. For example, you can change the shape of the points or their color.

- Click on the circle symbol in the legend at the right of the graph.
- Select **Point Type** from the **Graph** menu. On the submenu that appears to the side of this menu, select the square in the first row. The scattergram now looks like:



- Select **Color** from the **Graph** menu. The menu that appears to the side of this menu depends on whether or not you have a color system. If you have a non-color system, the names of the colors are spelled out in words; if you have a color system, the colors are shown in a palette.
- Select a new color for the points. If you have a non-color system, the points will remain black, but they will appear in the selected color if you print your graph on a color printer.

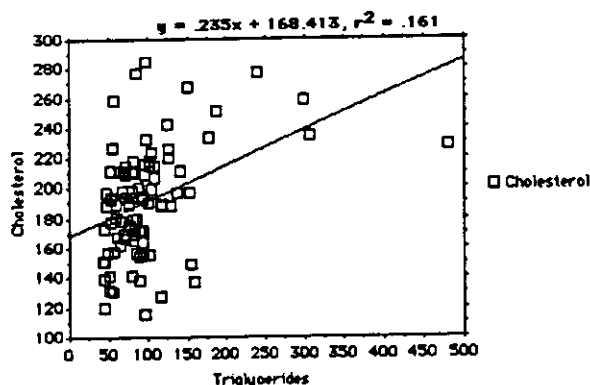
Adding a Regression Line

StatView can quickly add a regression line to your graph. At the same time, StatView shows you the equation for the line.

- Select **Regression** from the **Compare** menu. You see:



Select Regression:
☒ Simple ☐ Multiple ☐ Polynomial
Select the order of the regression:
☒ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9
Confidence intervals: 95% and
Add as a new column to the data window:
☐ Residuals
☐ Standardized Residuals
☐ Fitted Values
☐ Predicted Values
Create new column(s) using values from:
☒ Included rows ☐ all rows
Compute residual statistics: ☒ No ☐ Yes

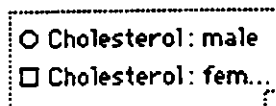
- Click OK. StatView shows:



Make Subset by Gender

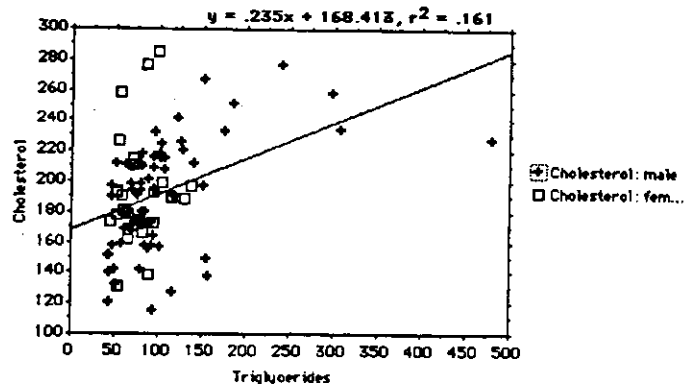
StatView lets you easily distinguish the male and female values in this graph.

- Click the subset specify tool, . (If you do not see the tool, enter composite mode by clicking on the  tool.) You see a list of the categorical columns.
- Select Gender. Click OK. Note that the legend now has two entries, one for Cholesterol: male and one for Cholesterol: female. You want to resize the legend so that you can see the whole names.
- Click in the grow box of the view window and drag it to the right. This makes the window wider and shows more text in the legend at the right of the graph.
- Click on the legend to select it:



- Click on the square in the lower right of the legend and drag it to the right to have it show all the letters in the names.

- Click on the top circle symbol in the legend.
- Select **Point Type** from the **Graph** menu. On the submenu that appears to the side of this menu, select the cross in the second row. The scattergram now looks like:

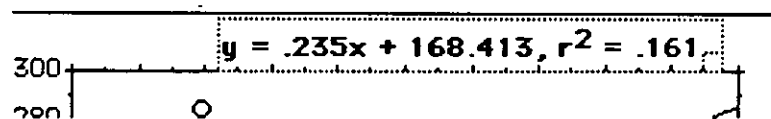


The male values are now displayed with a cross and the female values with a square.

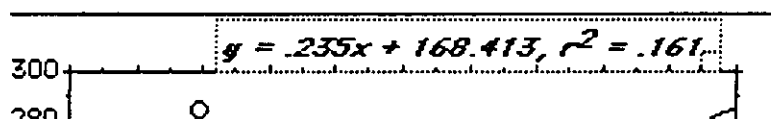
Adding Display Features to Your Graph

To customize parts of graphs, select the desired parts and select the desired changes from the menu bar. You can also add items to your graph with the tools at the left part of the view window.

- Click on the regression line equation in the graph. The text is surrounded with a dotted rectangle:



- Select **Style** from the **Text** menu. On the submenu that appears to the side of this menu, select **Italic**.



- Click on the line tool at the left of the view window.



The tool becomes black and the cursor becomes a cross-hair.

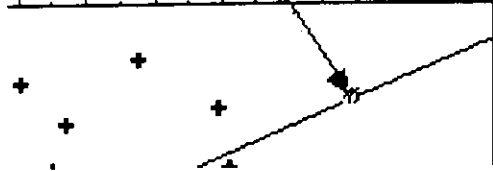
- Move the cursor just below the equation. Click and drag the cursor to the right and down to the regression line. Release the mouse button.

$$y = .235x + 168.413, r^2 = .161$$



- Select **Arrow Head** from the **Graph** menu. On the submenu that appears to the side of this menu, select the second item from the top, the arrow that is pointing to the right. The line becomes an arrow pointing at the regression line.


$$y = .235x + 168.413, r^2 = .161$$



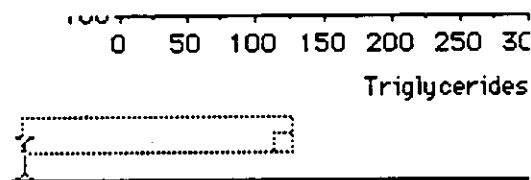
Adding a Caption

You can add your own text to your graphs. This is useful if you are using the graphs in presentations or papers.

- Select the graph area by clicking in it.
- Drag the lower right corner of the graph up half an inch.
- Click on the horizontal graph label that reads "Triglycerides".
- Drag it up half an inch, leaving room at the bottom for a caption.

- Select the text tool, .

- Click in empty area at the bottom left of the graph. StatView shows an empty text box:



- Click in the box at the lower right corner of this text box and drag the box to near the right side of the graph. This will make the box large enough for your text.
- Type **Figure 1: The graph shows the dependence between triglycerides and cholesterol.**
- Select **Figure 1:** by clicking on it and dragging.
- Select **Style** from the **Text** menu. On the submenu that appears to the side of this menu, select **Bold**. Your graph now looks like:

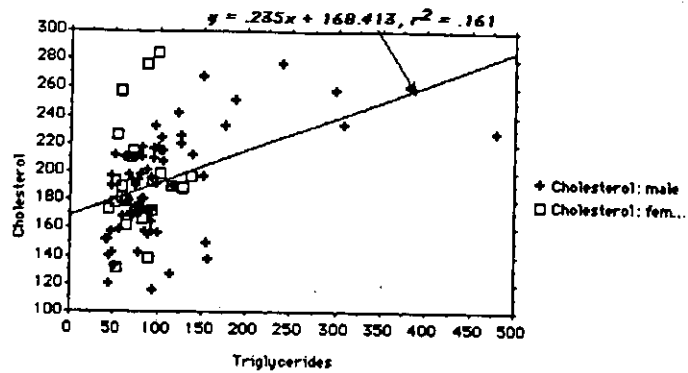


Figure 1: The graph shows the dependence between triglycerides and cholesterol

New Columns

When you create a dataset, you enter your raw data in the columns. It is often useful to create charts and graphs of a column after the data in the column has been transformed mathematically. For example, if one of your columns is population size, you might want to perform analyses on the logarithm of the population.

To perform analyses on transformed columns, you create new columns that are transformations of your data columns. You can treat these new columns just like any other column and use StatView's analysis tools on them. The Transform command creates a new column that is a transformation of a selected column.

- Open Lipid Data, or clear variable assignments if Lipid Data is currently active.
- Select Transform from the Tools menu. You see:

Name: **1/x of Gender**

Select column:

Gender
Age
Weight
Cholesterol
Triglycerides
HDL
LDL

Transformation:

1/x
Sqrt
|x|
x^2
x^n
ln(x)
ln(1+x)

Decimal Places:

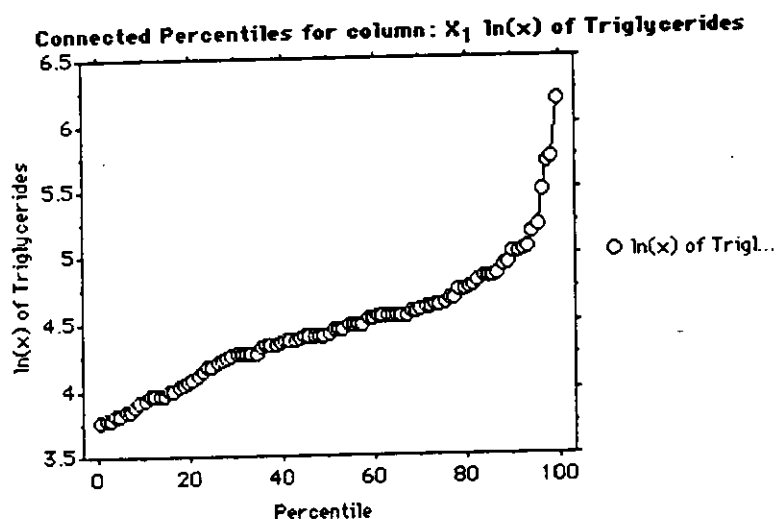
☐ 0
☐ 1
☐ 2
☒ 3
☐ 4
☐ 5
☐ 6
☐ 7
☐ 8
☐ 9

Transform
Exit

The list on the left lets you select which column you want to take the data from, and the list on the right shows all the possible transformations.

- Select Triglycerides from the column on the left.

- Select $\ln(x)$ from the column on the right. Note that the name at the top of the dialog is a description of the transformation. You can change this name to anything you want.
- Click the Transform button. This creates a new column.
- Click the Exit button. If you wanted to create more new columns, you could do so before clicking Exit.
- Scroll horizontally in your dataset until you reach the last column. This is the new column you just added.
- Double-click beneath the name of the column. This makes it the X_1 column.
- Select Percentiles from the Describe menu.
- Select Line chart from the View menu. StatView shows you a line chart of the percentiles of the logarithm of the Triglycerides column of your dataset:



Where To Go Next

These examples have shown you the basic steps of how to use StatView to analyze and graph your data. The rest of the manual provides more detailed information on specific areas of the software. Please refer to the following chapters for more detailed information on specific areas of StatView's operation.

Creating datasets	Chapter 2
Setting up analyses and graphs	Chapter 3
Generating graphs	Chapter 4
Customizing graphs	Chapter 4
Descriptive statistics	Chapter 5
Comparative statistics	Chapter 6
Special column types	Chapter 7

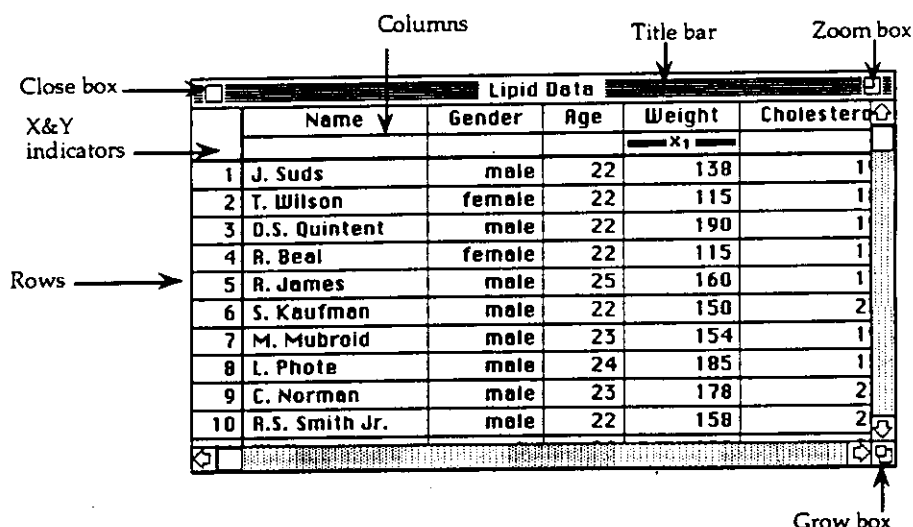
Chapter 2 — Datasets

This chapter describes StatView's datasets. Datasets are very easy to use, especially if you are familiar with spreadsheets such as Excel. This chapter also describes how to import data into your datasets and how to export your data to other Macintosh programs or other computer systems. After you become familiar with datasets, you will be able to quickly generate tables and graphs from your data.

The Data Window

Datasets are shown in the data window. This window looks like common Macintosh windows. You can control the appearance of the data window just as you do with windows in other applications. You can have up to eight StatView datasets open at one time.

A typical data window looks like:



To activate a data window, click anywhere in the window. You can also select the data window from the **Window** menu. You can move the window on your screen by clicking on the window's title bar and dragging it to a new location.

To change the size of the data window, click and drag on the grow box in the lower right corner of the window. This is just like other Macintosh applications.

StatView also lets you make a data window the full size of your Macintosh screen. This works like other Macintosh applications. If your screen is smaller than the full screen size, you can automatically enlarge it by clicking on the zoom box in the upper right corner of the window, by double-clicking on the window's title bar, or by selecting **Zoom Up** from the **Window** menu. If the window is already at full size, clicking in the zoom box, double-clicking in the title bar, or selecting **Zoom Down** from the **Window** menu changes the window back to its previous size and location.

Datasets

If you are on a large-screen Macintosh, you can specify whether you want "full-screen" to mean the full size of your screen or the size of a small-screen Macintosh. Select **Preferences** from the **Tools** menu to tell StatView which you prefer.

Datasets are organized in rows and columns. Each column in a dataset represents a variable in your data and each row represents a record. A record consists of *cells* which contain data. You can also create columns that are mathematical transformations of other columns in the dataset.

Columns

When you create a new dataset, you define what each column will look like and what kind of data will be entered into the cells in the column. You can add new columns to your dataset after you have created it. StatView datasets can contain up to 8,191 columns.

Each column has the following attributes:

- Name
- Type
- Number of decimal places displayed
- Width

The column's name can be up to 37 characters long. It can contain any characters including spaces.


The column type is one of the following:

Type	Description
Integer	Whole numbers in the range $\pm 32,767$
Real	Numbers with fractional parts. The range for real numbers is $\pm 1.1\text{E}4932$. The real number closest to 0 is $\pm 1.9\text{E}-4951$. StatView stores real numbers to 18 decimal places.
Long	Whole numbers in the range $\pm 2,147,483,648$.
Category	Data that falls into groups. Category columns are used to describe the different levels of a treatment or classification under which observations are recorded. For example, you might have a category called "Gender" that has two elements, "Male" and "Female". Categories are described in detail later in this chapter.
String	Text values of up to 80 characters. These values cannot be used in analyses.

If you have selected Real for the type, you can specify the number of decimal places to be displayed in the data window. StatView will display from 0 to 9 places.

StatView has no explicit date data type. You can express dates as long integers (such as "900416" to represent April 16, 1990) if you need to perform date arithmetic.

When you create a column, StatView automatically makes the column wide enough to accommodate the column's name. To change the column width:

- Move the cursor to the right of the column name you want to change, between the two column names at the top of the data window. The cursor becomes .
- Click and drag the separator line to the desired position.

If a column is too narrow for the data in the column, StatView displays the data in a narrower form. Numeric data is displayed by a series of pound signs (#####) and text data is displayed as the leftmost part of the data followed by ellipses (...).

Rows

Each row in a dataset represents a record. You add rows to your dataset by entering a value in the *input row*, the gray record at the bottom of your dataset. As soon as you add a value to a cell in the input row, StatView adds a new row and creates another input row. StatView datasets can contain up to 32,765 rows.

Opening Datasets

To open a dataset that you have stored on disk:

- Select **Open** from the **File** menu.

StatView's **Open** command acts just like the **Open** command in other Macintosh applications.

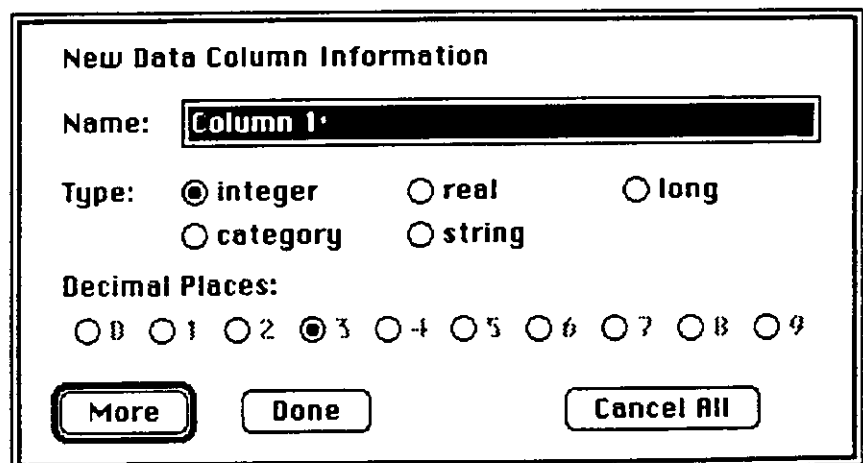
You can also open a dataset in the **Finder** by double-clicking on its icon or by selecting its icon and choosing **Open** in the **File** menu.

If you have a data file that you want to analyze but it is not in StatView format, you must import it into a StatView dataset. This is described later in this chapter.

Creating Datasets

To create a dataset from scratch, you simply define the characteristics for each column in the dataset.

- Select **New** from the **File** menu. You see the following:



The dialog box titled "New Data Column Information" contains the following fields and controls:

- Name:** A text field containing "Column 1".
- Type:** A group of five radio buttons:
• ☒ integer
• ☐ real
• ☐ long
• ☐ category
• ☐ string
- Decimal Places:** A group of ten radio buttons labeled 0 through 9.
• ☐ 0 ☐ 1 ☐ 2 ☒ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9
- Buttons:** Three buttons at the bottom: "More", "Done", and "Cancel All".

Use this dialog to enter the information for each column. For each column in your dataset:

- Enter the column's name in the text edit area. StatView requires that each column name be unique. If you try to create a column which has the same name as an existing column you will get an alert.
- Select the column type from one of the five choices.
- Enter the number of decimal places (if this is a real column).

You then either click **More** to add another column or click **Done** to indicate that there are no more new columns to define. When you click **Done**, StatView shows you the new dataset you have just created.

Note: If you are creating a column whose type is **Category** you will need to enter additional information regarding the column after you click the **More** or **Done** button. Please see the discussion below on **Category** data for detailed information.

Altering Datasets

You can change the structure of a dataset by adding or deleting columns or by changing the attributes of a column.

Adding Columns

- Select **New Column** from the **Tools** menu to add new columns to the end of the dataset.

This command displays the same dialog as the **New** command. To add more than one column at a time, click the **More** button after describing each new column.

Inserting Columns

To insert a new column between two columns in your dataset instead of at the right side:

- Place the cursor between the two columns, hold down the **Command** key, and click on the line between the columns.

You will see the same dialog as the **New** command. To add more than one column at a time, click the **More** button after describing each new column.

Removing Columns

- Select the column by clicking on the column's name at the top of the data window.
- Select **Delete** from the **Edit** menu or hit the delete key on the keyboard

Changing Column Attributes

- Select the column.
- Select **Format** from the **Tools** menu.

You can change a column's name and, if the column has the type "Real," the number of decimal places.

Changing Column Type

You cannot change the type of the column with the **Format** command. To change the type of a column, you have to create a new column with the desired type, then move the values from the original column to the new column:

- Select **New Column** from the **Tools** menu. Enter the desired attributes for the new column.
- Select the original column by clicking on the column's name at the top of the data window.
- Select **Copy** from the **Edit** menu.
- Select the new column.
- Select **Paste** from the **Edit** menu.
- Select the original column.
- Select **Delete** from the **Edit** menu.

Inserting Rows

StatView does not directly support inserting rows between two existing rows. Instead, you must make an empty row in the input row (all missing values), then move the rows below where you plan to insert the new row. For example, to insert a row above row 20 in a dataset:

- Select the first cell in the input row.
- Type **Option-8** or **period**. StatView fills the rest of the row with missing values.
- Select row 20 by clicking on its row number.
- Scroll to the last records in the dataset.
- Hold down the **Shift** key and click on the row number of the row above the new row you just started. This extends the selection from row 20 to the last full record.
- Select **Cut** from the **Edit** menu.
- Click on the gray row number for the input row. This selects the input row.
- Select **Paste** from the **Edit** menu. StatView fills all the records below the new one.

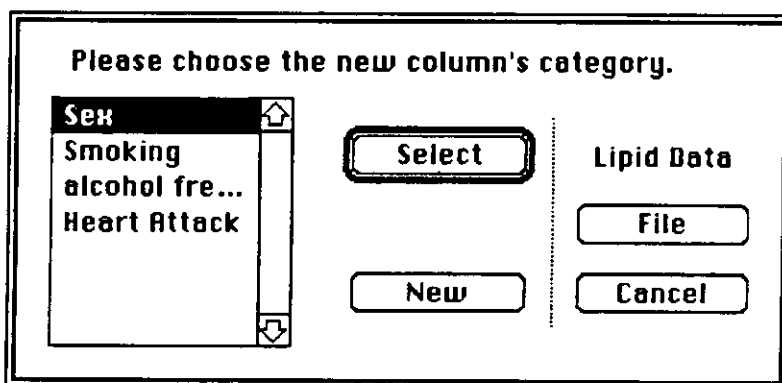
Category Data

StatView's advanced data-handling features let you create categories to handle columns which contain nominal information. A category column contains a number of distinct values that describe the different levels of a treatment or classification under which observations are recorded. When you assign a column the data type Category, you must either create a category and define which elements can be entered into the cells of the category column or select an already created category. You can define up to 255 elements for a category in StatView. For instance, Lipid Data has a column called Smoking History which references the category called "Smoking." The values that appear in the Smoking History column can be only one of the five elements of this category: "no," "quit," "cigarettes," "cigars," and "pipes." Categories can be stored in the StatView library file and used across several datasets.

StatView also makes entering data into category columns easier since you can enter the fewest characters which identify the element. For example, if you use the "Smoking" category described above, to specify that a record has the "pipe" value, you need only type "p".

StatView stores category elements as ordinal numbers, starting with 1 for the first element you define. You can use these ordinal numbers to enter values into category columns as well. Typing a "1" will enter the first element of the category, typing a "2" will enter the second element, and so on. In addition you can use these ordinal numbers for rearranging your categorical data. See the Recode command in Chapter 7 for more information.

When you are adding a column that is a category, you select category as the column type in the New Column Information dialog. When you click **More** or **Done**, StatView prompts you to tell it what category you want to use for that column:



You have many choices here:

- You can use a category that is already defined in the current dataset, in this example Lipid Data is active. Choose a category from the list on the left and click **Select**.
- You can create a new category. Click the **New** button.
- You can use a category that you have previously defined and stored in the StatView Library file. Click the **File** button, select a category from the list for the library, and click **Select**.

- You can use a category from a different StatView dataset that is already open. Click the File button, select a category from the list for that file, and click Select.

New Categories

To create a new category for a column you have added:

- Click New in the dialog which appears after you add the category column. You see:

- Enter Hair color for the category name. Category names and their elements must be less than 20 characters in length. Press the Tab key.
- Enter Brown for the first element of the category.
- Click Add or press the Return key to add that element to the list.
- Enter Blonde for the next element name.
- Click Add again.
- Enter Brunette for the next element name.
- Click Add again.
- Enter Red for the next element name.
- Click Add again.

You have now defined the "Hair color" category as having four levels: brown, blonde, brunette, and red.

You can store this category definition in two places: the dataset you are currently working in or the StatView library. Storing a category definition in the StatView library allows you to easily use this category in several datasets. For example if you do several experiments where you use the category "Hair color" you may wish to store this definition in the StatView library. Then when you need to use this category again in a different dataset you can simply select it from the StatView library as opposed to defining it again. Please note that once you store a category in the library you can not edit or delete that category.

Category names do not appear in a dataset, but the element names do. Category names let you keep track of the different categories you use in your work. You should give a category a clear name which you will recognize if you need to use the category again in a different column of the same dataset or in a different dataset.

Use the **File** button to determine where to store the category definition. Clicking this button toggles between the current dataset and the StatView library.

Once you have determined where to store the category,

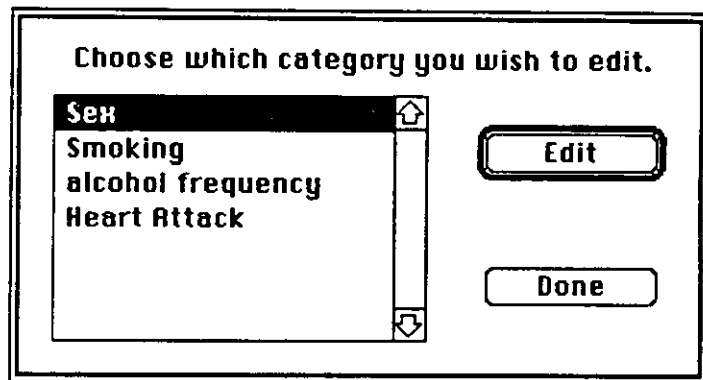
- Click **Done**.

If you mistype a name as you are entering new elements, select the name from the list, enter a new name next to **Element Name**, and click **Replace**. To remove an element from the list, select it and click **Delete**.

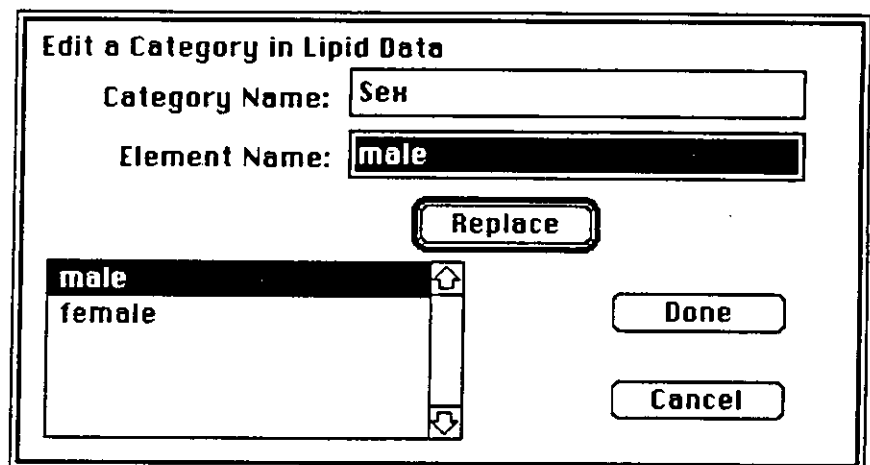
Editing Categories

You can change the names of elements in a category or the category's name.

- Select the dataset in which you want to change the categories. This example uses Lipid Data.
- Select **Edit Categories** from the **Tools** menu. You see:



- Select the category you want to edit.
- Click **Edit**. For example, if you select "Sex," you see:



Entering and Editing Data

Enter a new name for the category or select an element in the list and enter a new name for it.

The **Edit Categories** command is also useful for finding the ordinal values of the elements. In the command's dialog, the first element in the list has the ordinal value 1, the second has ordinal value 2, and so on.

You cannot add elements to a category once you have defined it. You can perform this action by copying the data to a new column that has a different (expanded) category.

To add elements to a category data column:

- Select **New Column** from the **Tools** menu to add a new column.
- Click **category** for the column type. Click **Done**.
- Click **New** to create a new category.
- Enter the same elements in the same order at the category that you want to expand. Add the new elements to the end of the elements list.
- Click **Done**.
- Select the original column.
- Select **Copy** from the **Edit** menu.
- Select the new column.
- Select **Paste** from the **Edit** menu.

You cannot delete elements from a category. If you need to remove elements from a category you will need to recode the information. See the **Recode** command in Chapter 7 for more information on how to do this easily.

You add rows to your dataset by entering a value in the *input row*, the gray row at the bottom of your dataset. As soon as you add a value in any cell in the input row, StatView makes that a new row and creates another input row. You can edit any cell in your dataset by selecting it and typing a new value.

You can enter data either from the main part of the keyboard or from the numeric keypad (if you have one). When entering category data, you only need to enter the least number of characters necessary to differentiate one element from another or enter the ordinal number for the element.

Missing Values

It is common to have *missing values* in your dataset. Missing values are placeholders that tell StatView that the data for that cell is not specified. A missing value is different than a zero or blank text field. StatView handles missing values intelligently, such as when it calculates averages for a column.

To enter a missing value in a numeric data column, type a period or press Option-8. When you move to a new cell, StatView denotes the missing data as a large dot

(•). For example, the following shows the cell in the Weight column as a missing value:

Gender	Age	Weight	Cholesterol
male	22	•	197

Selecting Data with the Mouse

To select a cell with the mouse, simply click in that cell. To select a group of cells, drag through that group of cells while holding down the mouse button. You can also extend a selection by holding down the Shift key and clicking in the cell at the end of the range. This is useful when you are selecting many cells at once since you can select the cell in one corner of the dataset, scroll to the diagonally opposite corner of the dataset using the scroll bars, hold down the Shift key, and click in the other corner.

You can select all the cells in a record by clicking once on the row number. To select a group of rows, select the first row and drag down to the other rows, or hold down the Shift key and select the last row you want. Selecting columns is similar: click once on the column names. Be careful not to double-click on a row number or column name, since these actions cause StatView to perform actions other than selecting.

You can choose **Select All Columns** or **Select All Rows** in the Edit menu to select your whole dataset. **Select All Columns** is useful for assigning variables while **Select All Rows** is useful for row inclusion. Either command is useful for selecting the entire dataset for copying to the Clipboard.

Use the **Show Selection** command in the Edit menu to display the selected cells if you have scrolled away from them.

Selecting Data with the Keyboard

You can move the selection in your dataset without the mouse. The following keys move the selection:

Key	Direction
Tab	Right
Shift-Tab	Left
Return	Down (except in the input row)
Shift-Return	Up
Arrow keys	The specified direction

Pressing the Tab key when you are in the last cell in a record moves to the first cell in the next record.

You can move by more than one cell at a time with other keys and combinations (some keyboards do not have these keys):

Cutting, Clearing, and Deleting Data

Key	Movement
Page Up	Page up
Page Down	Page down
Command-Page Up	Page left
Command-Page Down	Page right
Home	Upper left corner of dataset
End	Lower right corner of dataset
Command-Home	Upper right corner of dataset
Command-End	Lower left corner of dataset

Holding down the Command key while pressing an arrow key causes StatView to scroll the window by a page in the specified direction without moving the selection. This is useful for seeing other parts of your dataset without losing your selection.

Pressing Enter and Shift-Enter move in different directions depending on the selection you make in the Preferences command in the Tools menu. Your three choices for how the selection moves when the Enter key is pressed are:

- Move right
- Move down
- Do not move

Pressing Shift-Enter moves in the opposite direction of pressing Enter.

Use the Cut command to remove values from your dataset and place them in the Clipboard similar to other Macintosh applications. The Clear command clears cells in your dataset and the Delete command removes entire rows or columns. Note that Clear and Delete commands do not copy the removed data to the Clipboard. Unless you use the Undo command immediately after giving one of the commands, the data is permanently lost.

To cut cells, columns, or rows to the Clipboard:

- Select the cells, columns, or rows.
- Select Cut from the Edit menu.

If you have selected cells, they are filled with missing values; if you have selected columns or rows, they are removed from the dataset structure.

To clear cells:

- Select the cells you want to clear.
- Select Clear from the Edit menu.

Cleared cells are filled with missing values.

Pasting into StatView

To remove columns or rows:

- Select the columns or rows you want to remove from the dataset.
- Select **Delete** from the **Edit** menu.

You can also use the **Delete** command to remove the contents of the Clipboard. It is unlikely that you would want to do this unless you are low on memory and want to reclaim the Clipboard's memory. To clear the Clipboard:

- Activate the Clipboard window. If the window is already open, select it; if it is not open, select **Clipboard** from the **Window** menu or select **Show Clipboard** from the **Edit** menu.
- Select **Delete Clipboard** from the **Edit** menu.

Copying to the Clipboard

The **Copy** command in StatView acts just like in other Macintosh applications. You can copy cells, columns, or rows to the Clipboard.

After you have cut or copied data to the Clipboard, you can paste it in other places in your dataset. The data in the Clipboard could have been put there by StatView or by another program such as Excel; StatView pastes the same way in either case. The **Paste Transposed** command in the **Edit** menu changes the rows to columns and columns to rows when you paste.

Pasting Records

The most common task you use pasting for is to add new records to your dataset. To add one or more records from the Clipboard, you must select the entire input row, not just the cells in the input row.

To paste whole records into your dataset:

- Copy or cut the desired records to the Clipboard as described above.
- Select the entire input row by clicking on the gray box in the row number column (not on the cells themselves).
- Select **Paste** from the **Edit** menu.

If the record in the Clipboard does not match the records in the dataset, StatView will do its best to paste the data in. If there are fewer fields in the records in Clipboard than there are in the dataset, StatView fills the cells on the right with missing values. If there are more fields in the records in the Clipboard than there are in the dataset, StatView ignores the extra columns.

Pasting Cells

Another common task is to paste groups of cells into your dataset. It is important to note that StatView *overlays* data as it pastes; it does not displace existing data to

another location in the dataset. This overlaying process removes the current values of the cells and replaces them with the values from the Clipboard.

For the examples in this section, assume that the Clipboard has a six cells with the following data and arrangement:

194	65	52
173	130	70

To see how StatView overlays data, assume that you have the following cells in your dataset selected:

Weight	Cholesterol	Triglycerides	HDL	LDL
138	197	152	43	151.6
115	181	59	60	120.1
190	190	117	41	147.1
115	131	54	58	72.1

After giving the Paste command, the new data will look like:

Weight	Cholesterol	Triglycerides	HDL	LDL
138	197	152	43	151.6
115	194	65	52	120.1
190	173	130	70	147.1
115	131	54	58	72.1

As you can see, the previous contents of the selected cells were replaced with those from the Clipboard; they did not "move" to other parts of the dataset.

Pasting Different Sizes

In general, you will want to paste into a selection that is the same size and shape as the cells that are in the Clipboard. To check on the Clipboard contents, open the Clipboard window by selecting Clipboard from the Window menu. The Clipboard window tells you how many columns and rows are in the Clipboard:

Clipboard		
3 Columns with 2 Rows; 6 Total.		
194	65	52
173	130	70

If you select the same size and shape as the cells on the Clipboard, pasting will act just as you expect. If you select a different size or shape, StatView will adjust what is pasted in the following ways:

- If the selected area has fewer rows or columns than the Clipboard, StatView does not paste the cells outside the selection.
- If the selected area is an exact multiple of the size of the Clipboard, StatView will duplicate the data throughout the selection. For instance, if the

Clipboard has two rows by three columns and you select four rows by three columns, StatView will paste two copies of the Clipboard.

- If the selected area is not an exact multiple, StatView pastes only one copy of the data into the selection and fills the other cells with missing values.

Pasting Different Types

You should pay attention to the type of data that is being pasted into your dataset. StatView uses the following rules for converting data from one format to another:

		Copying from		
		Numeric	Category	String
Copying to	Numeric	Converted to column type (integer, real, or long real)	Converted to ordinal value of element	Converted to number if pasted text is a number
	Category	Number becomes element of category	Converted to new element based on both ordinal values	Converted to element if names match
	String	Text of pasted number	Text of pasted element name	No conversion

If StatView cannot make the conversion (such as from string to numeric when the string contains letters), it pastes missing values.

Importing Data

Of course, StatView is not the only program you use in your work. You will want to use StatView with applications such as word processors, spreadsheets, and so on. StatView can easily import data in text files created by other programs, even those on other computers. It can also paste data into an existing dataset from the Clipboard, as described earlier in this chapter.

Methods of Importing Text Files

Almost every Macintosh application can save data in text files (also called *ASCII files*). Thus, you can use almost any program to prepare data for StatView. Follow the directions in your applications' manuals for creating text files. In general, StatView will be able to read the output from most applications fairly easily.

If you are using programs on other computers to prepare your data, you must be able to access that data through files on your Macintosh. The three common methods for importing text files from other computers to your Macintosh are:

- Over a network that contains Macintoshes and other brands of computers (such as PCs and UNIX systems)
- Through a modem and a communications program (such as Microphone or White Knight)

- From a diskette that can be read on your Macintosh

Text File Format

All text files imported to StatView must have certain characteristics so that StatView can recognize your data. These required characteristics are:

- The text must be organized in rows and columns as they appear in the data window.
- Each row must be terminated by either a carriage return character, a line feed character, or a carriage return/line feed pair. Blank lines in your text file are ignored.
- The first row of your imported dataset may be the column names, as described later in this section. If you do not include column names, StatView will give the columns names such as "Column 1," "Column 2," and so on. You can, of course, rename the columns later.
- Your data points may optionally be enclosed by quote (") characters.
- A data point in your imported dataset that is either a single period (.) or a bullet (•, ASCII value 165) is read as a missing value.
- Each data point must be separated from the next data point by one or more *separator characters*. The most common separator character is the tab character. You can also use space characters, commas, carriage return characters, or any other character that you specify.
- If you use spaces for your separator, StatView compresses multiple spaces into a single space. This is useful when you to capture columnar data from the screen in a communications program and import it into StatView.
- When you use any separator character other than a space, StatView will put a missing value in the dataset if it sees two separator characters together.
- If you use the carriage return as the separator character, you must tell StatView how many columns of data you have in each row.
- You can import category data by coding that data as integers in your imported dataset. When importing, you can specify that any column that contains only small integers (between 0 and 255) whose lowest value is between 0 and 6 is to be imported as a category column. After importing, you can then specify new names for each category element.

Although this list of requirements is long, most Macintosh applications save text files according to these rules. Thus, you can generally assume that exporting from another Macintosh program will work on the first try.

The Import Command

To import a text file:

- Select **Import** from the **File** menu. You see a list of the text files on your disk.
- Select the desired text file from the list.

- Click **Import**. You see:

Please specify how this text file looks.

Items may be separated with tabs and:

☐ spaces ☐ commas ☐ returns ☐

Number of data columns:

☐ Convert small integers to Categories.

OK **Cancel**

Select the type of separator character in your text file; if your file uses tab characters, do not check any of the choices. If you are using carriage returns for separators, you must also specify the number of data columns. If you want to convert columns of small integers to categories, check that option.

- Click **OK**.

After you import your dataset, examine each row to make sure that the data is consistent with what you expect. It is common for some rows to have unexpected missing values at the end of the row; this is usually caused by two data points having no separator between them. If the data points appear to be accidentally shifted to the right, it is likely that you have doubled separator characters.

How StatView Imports Datasets

When importing, StatView checks whether the first row is column names. If it is, it uses those names for the columns. In order for the first row to be interpreted as column names, every data point in that row must be non-numeric. Thus, you cannot have column names such as "1990". In the unlikely case that your whole dataset is string data and there are repeated values in it, StatView will check whether the data points in the first row are unique within their columns. If they are unique, the first row is used as column names; if any are not unique or there are no repeated values in the dataset, the first row is used as data.

As StatView imports the file, it determines the data type of each column. In columns with all numbers, it assumes that the column is integers until it sees the first real number (one with a decimal point); it then assumes that the column is all real numbers. In any column, when StatView sees an alphabetic character, it assumes that it is a string or category column. In such a column, if none of the text values are repeated, it becomes a string column. If some text values are repeated but there are more than 256 types of text values, it will also become a string column. If there are repeated values and fewer than 256 text types, StatView makes the column a category. If there are exactly 256 text types, the type of the resulting column depends on whether StatView uses the first data point as a column name.

Some columns may have both text and number values. StatView counts the number of each and creates the column based on which type has more values. For instance, if a column has 25 numeric values and four string values, it becomes a numeric column and the four string values become missing values. Note that numbers with currency marks (such as "\$") or commas within values are

considered text values. If a column becomes a category column, any small integers found in the column will be mapped to the elements with those values.

After you have imported your data, you may want to modify the dataset before saving it to disk. Use the techniques described earlier in this chapter to change things such as the column names, column width, and so on. Save your dataset with the **Save As** command in the **File** menu.

Note that if you are importing a dataset that has only one column, StatView will not find missing values that are simply empty lines. In this case, missing values must be entered as a period (.) or Option-8 (•) character in the original data.

Example of Importing

You can see how some of these rules are applied by looking at a real example. The import example file in the **Sample Data** folder is a tab-delimited text file created in Microsoft Word, on the Macintosh. It looks like:

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
1	12/1/87	\$1	1	1	red	Charlie	
2	12/2/87	\$10,001	2		blue	Parker	
3	12/3/87	\$20,001	3	3	1	Miles	
2	12/4/87	\$30,001	4	4	yellow	Davis	
3	12/5/87	\$40,001	5.1	5	red	John	
2	12/6/87	\$50,001	5.2		red	Coltrane	
1	12/7/87	\$60,001	5.3	7	blue	Roscoe	
2	12/8/87	\$70,001	5.4	8	green	Mitchell	
3	12/9/87	\$80,001	5.5	9	10	23.7	
1	12/10/87	\$90,001	5.6		red	Ra	

Import the data with the following steps:

- Select **Import** from the **File** menu. You see a list of the text files on your disk.
- Open the **Sample Data** folder and select **import example**.
- Click **Import**. Do not change the default choices (tab-delimited data, no integer-to-category conversion).
- Click **OK**.

The new dataset looks like:

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
1	1	12/1/87	\$1	1.0	1	red	Charlie
2	2	12/2/87	\$10,001	2.0	•	blue	Parker
3	3	12/3/87	\$20,001	3.0	3	red	Miles
4	2	12/4/87	\$30,001	4.0	4	yellow	Davis
5	3	12/5/87	\$40,001	5.1	5	red	John
6	2	12/6/87	\$50,001	5.2	•	red	Coltrane
7	1	12/7/87	\$60,001	5.3	7	blue	Roscoe
8	2	12/8/87	\$70,001	5.4	8	green	Mitchell
9	3	12/9/87	\$80,001	5.5	9	•	23.7
10	1	12/10/87	\$90,001	5.6	•	red	Ra

Note how each column is imported:

Column	Type	Notes
1	Integer	All values were integers.
2	String	Even though the values appear like dates, StatView reads them as string values since they have non-numeric characters (the "/" characters).
3	String	Even though the values appear like numbers in currency format, StatView reads them as string values since they have non-numeric characters (the "\$" and "," characters).
4	Real	All values are numeric and the fourth value has a decimal point, so the column becomes real.
5	Integer	Note the three missing values for the data items that were missing.
6	Category	Since there are repeated strings, StatView makes this a category column. The numeric value "1" is mapped to the first category element, "red". The numeric value "10" becomes a missing value because there are fewer than ten elements in the category created by StatView.
7	String	The value 23.7 is a string of four characters in this column.

Note that if you had selected **Convert small integers to categories** in the **Import** dialog, columns 1 and 5 would have become category columns since they contain all small integers.

Importing from Excel

Microsoft Excel saves formatted values in text files. For example, if you have cells formatted with currency marks such as "\$", those will appear in the exported text file. Also, dates are saved with formatting characters such as "/" and "-". Be sure to remove any such formatting before exporting your dataset.

To save an Excel worksheet as a text file while running Excel:

- Select **Save As** from the **File** menu.
- Click **Options**.
- Click **Text** in the dialog. Click **OK**.
- Type a new name for the exported file. Click **OK**.

Importing from Other Applications

Each application saves text-only documents in slightly different ways. See your program's manual for more information. If you have a choice, select tab-delimited text files since those are least prone to importing mistakes. Remember that you can also import data into an existing StatView dataset through the Clipboard.

Exporting Data and Pictures

StatView can export data both in text files and through the Clipboard. This makes it easy for you to copy your data and the results of your analyses to other Macintosh applications.

Exporting Data with Text Files

To save your dataset as a text file:

- Select **Save As** from the **File** menu. The **Save As** dialog is the same as for other Macintosh applications except that there is a choice at the bottom for the file format.
- Click **Text** for the file format. You see:

Please specify how to save this text file.

Separate items with:

☒ tabs ☐ commas ☐ returns

☐ Save column names.

☐ Enclose text items with quotes.

☐ Save Category columns as small integers.

OK **Cancel**

Many of the choices in this dialog are the same as those in the **Import** command. You can specify what character will be put between each field in a record, whether to include the column names as the first row, whether to put text items in quotes, and whether to save category data as small integers.

Most Macintosh applications assume that data being imported has tab characters between columns, so you should select this unless your application's manual specifies otherwise. Some database programs expect to see column names in the first record, others cannot accept column names. Similarly, some applications insist that text fields be quoted, while other insist that they are not.

When StatView saves the text file, it uses the dataset's settings for each column to determine how many decimal points each real number should have.

Exporting Data Through the Clipboard

You can pass data to other Macintosh programs through the Clipboard. See the sections on using the Clipboard earlier in this chapter for information on how to copy records and cells to the Clipboard.

Exporting Results Through the Clipboard

Once you have performed an analysis on your data, you may want to export that analysis to another program. For example, you might make a scattergram, embellish it with StatView's drawing features, then want to incorporate it into a Microsoft Word document. Any results in the view window are always exported through the Clipboard.

You can export two types of results in the Clipboard: text and pictures. If the view window has a table in it, you can select some or all of the text and save that text to the Clipboard. If the view window has a graph, you can select some or all of the graph and save it as a picture in the Clipboard. You may also want to save a table as a picture in the Clipboard.


StatView saves pictures in the Macintosh-standard PICT format with each item in the picture being its own element. When you import those pictures in a drawing program such as MacDraw or SuperPaint, you can manipulate each element separately.

To export from the view window to the Clipboard:

- Select the items you want to export. As a short-cut to select all the items in a view you can use the **Select All** command in the Edit menu.
- Select **Copy** from the Edit menu.

You can see the contents of the Clipboard by selecting **Clipboard** from the **Window** menu. You can, of course, transfer items from the Clipboard to the Scrapbook desk accessory in the same fashion you do with other Macintosh programs.

There are three ways to select items:

- Click on individual items
- Hold down the Shift key and click on other items to include them in the selection
- Click and drag over a group of items to select them all. When you click outside of an item in the window, the cursor becomes  and a rectangular marquee forms as you drag the cursor.

To select and copy all the items in the view window as a picture:

- Select **Copy View** from the Edit menu.

If you have a table in the view window, copying selected items to the Clipboard will automatically copy them as text.

To copy all the text results of an analysis:

- Select **Select All** from the Edit menu.
- Select **Copy** from the Edit menu.

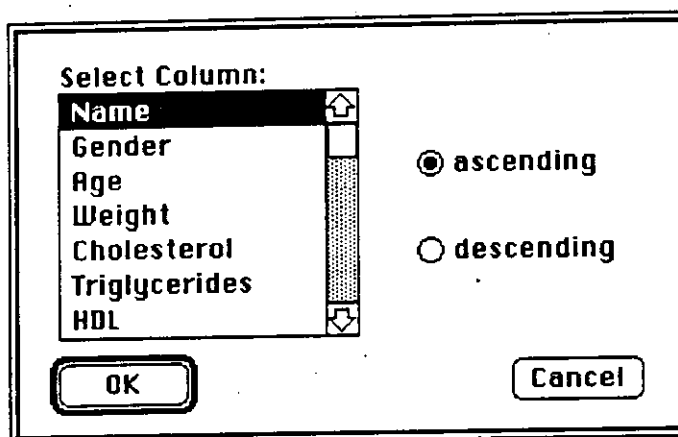
If you are displaying a Table view, using the **Copy View** command will copy out the entire table as a picture.

Sorting Datasets

You may want the records in your dataset sorted by a particular field. For example, you might want to sort all the records so that you can compare the values against another dataset that is sorted. You can sort on any one column in either ascending or descending order. Note that you cannot undo a Sort command.

To sort your dataset:

- Select Sort from the Tools menu. You see:



- Select the desired column from the list.
- Select an order to sort: ascending (A to Z) or descending (Z to A).

If you sort a dataset in ascending order, records with missing values in the selected column will be placed at the end of the dataset.

If you want to sort a dataset and later go back to the original order, you must first create a new column that contains the record number for each record using the Time Series command in the Tools menu. See Chapter 7 for more information on this command. After you have created such a column, you can later sort using that column to restore the original order.

Printing

You can print both the data and view windows. StatView can print on any Macintosh printer and works very much like other programs. You use the Print command in the File menu to print StatView windows. When you print a dataset, the entire dataset is printed.

When you print a view, you can add titles and footers to your printed page. For view windows, you can also select whether or not to print the view frame around the picture and whether to print as large as your pages. When you give the Print command, the standard Print dialog has these choices added to the bottom:

Page Title:

Page Footer:

☐ Don't print Page Frames

☐ Use whole sheet of paper for View

If you want text in the title and/or footer on every page, enter it in the boxes. If you do not want the frame around your graphics, select **Don't print Page Frames**.

You may want to print your graphs as large as possible so that you can see more detail. In that case, select **Use whole sheet of paper for View**. This option is not always what you want since it can sometimes distort graphs, especially pie charts. This option stretches the graph in one dimension, so it is more common to use this when you are printing in landscape (horizontal) mode. StatView cannot expand user-added elements such as arrows when printing in this mode so printing graphs with these elements is not recommended. Instead, resize your graph on a large screen and move the elements yourself before printing.

Before printing, be sure to select the correct printer with the **Chooser** desk accessory. Also, be sure that it is set up correctly with the **Page Setup** command in the **File** menu.

Printing on a LaserWriter

If you are printing on a LaserWriter, you should select a LaserWriter font such as Times or Helvetica for your text. If you use a non-LaserWriter font such as Geneva, your printed output will look jagged since StatView does not perform font substitution. You can set the default font for non-tabular views by selecting **Preferences** from the **Graph** menu. These choices are discussed in Chapter 4.

Chapter 3 — Introduction to Graphs and Analyses

Once you have created a dataset and have entered data, it is easy to graph the data and create statistical graphs and tables using StatView. Performing an analysis is very easy: you simply tell StatView what columns you want to analyze, the types of analysis you want, and in what format you want to display the output.

This chapter shows you how to graph data and explains the basics of creating graphs and tables of statistical results. This chapter also tells you how to create criteria which allow you to examine specific subsets of your dataset. Chapter 5 and Chapter 6 give in-depth information about the specific tests you can perform and how to use those tests in your work.

Assigning Variables

Before you create graphs and tables, StatView must know what columns you are interested in. To tell StatView which columns you want to analyze, you assign an X or a Y to the columns. A column which has an X assigned to it is called an X variable; a column which has a Y assigned to it is called a Y variable. Descriptive statistics and their associated graphs all use X variables; comparative statistics and their associated graphs use many X variables or a combination of X and Y variables.

Your dataset can have more than one X and Y variable, in which case each variable has a subscript such as X_1 , X_2 , and so on. StatView automatically assigns a higher-numbered subscript as you assign new variables and rennumbers subscripts as you remove assignments. The importance of the subscript numbers is described in the section on Creating Graphs and Tables later in this chapter.

When you assign an X or Y to a column, the variable assignment appears under the column name. For example, if you assigned an X to the HDL column, the screen would show:

Triglycerides	HDL	LDL
	X ₁	
152	43	151.6

There are four ways to assign variables:

- Double-click on the variable name
- Command-click on the variable name
- Select one of the Choose X, Choose Y, and Clear X&Y commands from the Variables menu
- Select Quick Assignment from the Variables menu

Assigning with the Mouse

The fastest method for making a column an X variable is to double-click on the column name. To make a column a Y variable, hold down the Option key as you double-click on the column name. Double-clicking and Option-double-clicking also clear a variable assignment if the column is already assigned.

Holding down the Command key while single-clicking on a column name has the same effect as double-clicking.

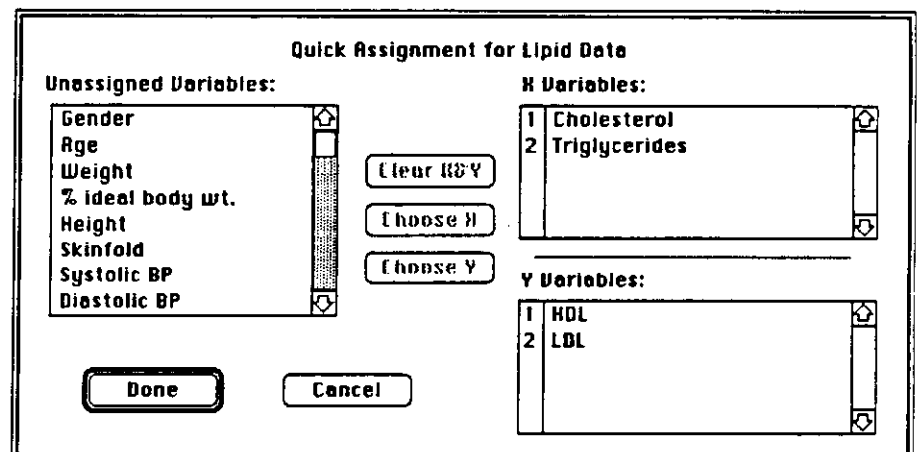
Assigning with the Menus

To assign variables using menu commands, select the desired columns and select **Choose X** or **Choose Y** from the **Variables** menu. This allows you to assign an X or Y to several columns at a time. To clear the variable assignments from selected columns, select **Clear X&Y** from the **Variables** menu. An easy way to clear all variable assignments is to choose **Select All Columns** from the **Edit** menu (or use Command-A) and then select **Clear X&Y** from the **Variables** menu.

Quick Assignment Command

The Quick Assignment command in the **Variables** menu is a general way of assigning variables through a single dialog. It allows you to make variable assignments without scrolling through your dataset to find a specific column.

You can display the Quick Assignment dialog box when either a dataset, a table or a graph is the active window. If a table or graph is active, selecting Quick Assignment allows you to assign variables for the dataset associated with the view. With Lipid Data active, the Quick Assignment dialog box looks like:



The scrolling list on the left shows the columns in the dataset that have not been assigned an X or a Y. The lists on the right show the variables that have been assigned and their usage.

To make any unassigned variables in the left-hand list an X or Y variable, select the column name in the list and click on **Choose X** or **Choose Y**.

To clear assignments, select the items from the X and Y lists and click **Clear X&Y**.

Note: You can select discontinuous columns in any list by holding down the command key and clicking additional column names.

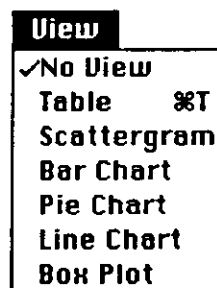
Creating Graphs and Tables

You can assign variables in the Quick Assignment dialog without clicking on buttons. If you double-click on a name in the Unassigned Variables list, it is assigned an X variable; if you hold down the Option key as you double-click, it is assigned a Y variable. To clear variable assignments, you can double-click in either of the lists on the right. To change an X variable to a Y variable or vice versa, hold down the Option key while double-clicking in one of the lists on the right.

Click Done in the Quick Assignment dialog, the dialog disappears and the variable assignments take effect.

Once you have assigned your variables, you can instantly produce graphs and tables. Select how you want to view your results from the View menu and the type of analysis you want from the Describe or Compare menus. If you haven't chosen any statistical tests from the Describe or Compare menus, StatView graphs your data; if you have chosen one or more statistics, StatView displays the results of the statistical test in tabular or graphic form.

The View menu has the following options:



Some analyses limit the graphic format choices. For example, you cannot display confidence intervals as a pie chart, but you can display confidence intervals as a scattergram, bar chart, or line chart. Additionally, some analyses require particular mixes of X and Y variables; those are described later in this chapter.

Automatic Updating

Views always reflect the contents of the data window. They are automatically updated when you change anything in the data window. The most common actions you take that update the view window are changing variable assignments and changing your data. Two other actions change the view window: selecting a new statistic from the Describe and Compare menus and changing the data selected for analysis (described later in this chapter).

Modifying Views

You can tell StatView the number of decimal places for results in the view window as well as the default size of the view window. Select Preferences from the Tools menu to specify either of these options.

You can select a default font and font size for graphical output as well as setting the default plotting symbols and symbol colors. Select Preferences from the Graph menu to specify these options.

Graphing Data

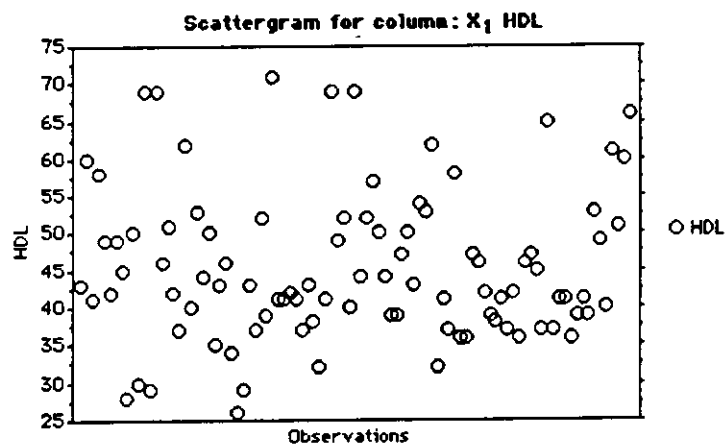
You can change the size of the view window if it has a graph in it. Drag the grow box in the lower right corner of the window or use the **Zoom Up** command in the **Window** menu. You cannot change the size of the view window when you are viewing a table.

To graph your data, assign X or Y variables to the columns you wish to view and select the desired graph from the **View** menu. Graphs can display "raw data" — that is data without any additional statistical output, or computed information — such as a regression line fitted to a scattergram or a histogram of a frequency distribution. The type of graph which is displayed depends on the statistical choices chosen from the **Describe** and **Compare** menu. If you have a question about how to display a particular graph, Chapter 4 describes all the graphs available in StatView and how to generate them.

If you are plotting just raw data, you can view your columns as a **Scattergram**, **Bar** or **Line** charts. The type of graph you see and the number of variables displayed on the graph depend on your X and Y variable assignments.

If you have assigned one or more X columns but no Y columns, StatView plots your data along the horizontal axis as observations in the order they appear in your dataset. This is useful for seeing the dispersion of data or any record-order trends in a column. For instance:

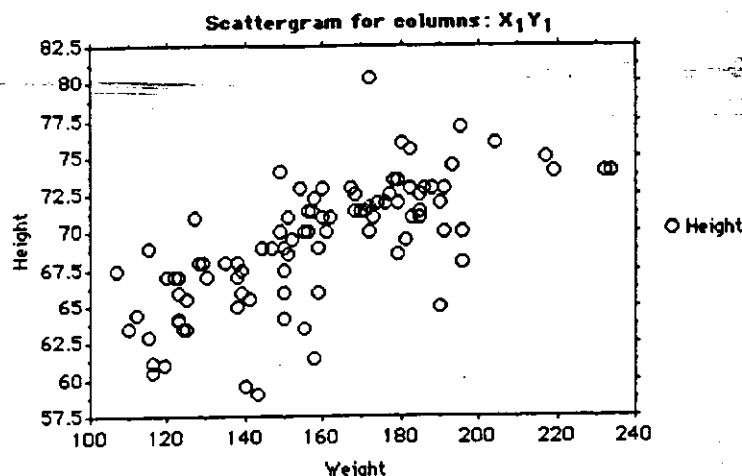
- Open Lipid Data.
- Assign X to HDL.
- Select Scattergram from the **View** menu.



It is more common to plot one column against another to see the relationship between variables.

- Activate the dataset window.
- Double-click on the HDL column name to clear the X_1 variable. No variables are assigned, so the view window is empty.
- Assign X to Weight.
- Assign Y to Height. Activate the view window, this graph appears:

Statistical Analysis



Chapters 5 and 6 describe each StatView analysis in detail. There are some features common to all analyses, and some general guidelines in arranging data and specifying views.

When you select a statistic from the **Describe** or **Compare** menu, a check mark appears next to the statistic. Several descriptive statistics can be viewed simultaneously, with one exception. When you select **Frequency Distribution**, all other choices in the **Describe** menu are unavailable. Selecting any statistic in the **Compare** menu unselects all other choices in that menu; only one type of comparison shows at a time. To clear all selections in a menu, choose **None** (the first choice) in that menu.

When you analyze a dataset, StatView only looks at rows that have not been *excluded*. Excluding rows is described later in this chapter.

Variable Assignments

The **Describe** and **Compare** statistics expect different variable assignments. The descriptive statistics located in the **Describe** menu are used with X variables. The comparative statistics located in the **Compare** menu are used to analyze several X variables or combinations of X and Y variables. The table at the beginning of Chapter 6 provides a list of the variable requirements for each type of statistical test. When multiple variables are assigned StatView uses the follow rules to determine which variables to look at:

- If you have assigned more than one X variable but no Y variables, StatView analyzes (or graphs) each X variable.
- If you have assigned multiple X and multiple Y variables, StatView pairs the variables with the same subscripts. For instance, if you have assigned X₁, X₂, Y₁, and Y₂, StatView analyzes X₁ vs. Y₁ and X₂ vs. Y₂.
- If you have assigned only one X variable and many Y variables (or one Y and many X variables), StatView pairs each of the many variables with the single variable. Thus, if you have assigned X₁, X₂, X₃, and Y₁, StatView analyzes X₁ vs. Y₁, X₂ vs. Y₁, and X₃ vs. Y₁.

Note that if you have more than one of both X and Y variables but do not have an equal number of each, the unmatched variables will not be analyzed. Thus, if you have X_1 , X_2 , Y_1 , Y_2 , Y_3 , and Y_4 , you will only get analyses for X_1 vs Y_1 and X_2 vs Y_2 .

Grouped Data

Several statistics analyze grouped data. For these tests the grouping variable is designated as an X column and the dependent data as a Y column. A grouping variable must be either a category column or an integer column. For a category column StatView automatically determines the number of distinct values in the group. For integer data, StatView calculates the number of groups as being the column's Maximum Value - Minimum Value + 1. We recommend that you use a category variable as a grouping variable because the category variable allows you to clearly label the elements of a group.

Views of Analyses

Many analyses can be viewed in several different forms, while others can only be viewed as tables. The available views are:

Statistic	Table	Scatter	Bar	Line	Box	Pie
Mean, Std. Dev., etc.	√	√	√	√		
Confidence intervals	√	√	√	√		
Percentiles	√	√	√	√	√	
Mode	√					
Geometric mean	√					
Harmonic mean	√					
Kurtosis & skewness	√					
Frequency distribution	√		√			√
Comparative percentiles	√	√		√		
t-test	√					
Correlation coefficient	√	√				
Regression	√	√				
Stepwise regression	√	√				
Factor analysis	√	√				
ANOVA	√					
Chi-square	√					
Nonparametrics	√					

Multiple Analyses

For several statistics you can set up more than one analysis at a time by assigning multiple X and Y variables. For instance, you might want to perform an unpaired t-test comparing the age, weight, and cholesterol of male and female patients. Instead of performing three separate t-tests, assign all the variables at once and page through the results.

- Open Lipid Data.
- Assign X to Gender.
- Assign Y to Age, Weight, and Cholesterol, in that order.
- Select t-Test from the Compare menu.
- Click Unpaired t-Test in the dialog, then click OK.

StatView displays the table for X_1 vs. Y_1 :

Unpaired t-Test X_1 : Gender Y_1 : Age				
DF:		Unpaired t Value:		Prob. (2-tail):
93		-.824		.4123
Group:	Count:	Mean:	Std. Dev.:	Std. Error:
male	71	24.155	3.183	.378
female	24	24.792	3.538	.722

- Click the arrow at the bottom of the scroll bar in the view window. StatView displays the table for X_1 vs. Y_2 :

Unpaired t-Test X_1 : Gender Y_2 : Weight				
DF:		Unpaired t Value:		Prob. (2-tail):
93		8.192		.0001
Group:	Count:	Mean:	Std. Dev.:	Std. Error:
male	71	169.282	23.288	2.764
female	24	127.208	16.208	3.308

- Click the the arrow at the bottom of the scroll bar in the view window. StatView displays the table for X_1 vs. Y_3 :

Unpaired t-Test X_1 : Gender Y_3 : Cholesterol				
DF:		Unpaired t Value:		Prob. (2-tail):
93		-.537		.5926
Group:	Count:	Mean:	Std. Dev.:	Std. Error:
male	71	190.085	35.299	4.189
female	24	194.625	37.322	7.618

Missing Values

StatView ignores cases that contain missing values. All statistics are calculated using non-missing values only. Statistics which use grouped data exclude any missing values from the group. Statistics that use paired data require that values be non-missing across all records. Any records where one or more values are missing

Interactive Analysis

are deleted from the analysis. For each StatView analysis you will be notified if cases have been deleted due to missing values.

If your dataset contains missing values you may wish to recode them. Chapter 7 describes the types of recoding available in StatView.

Calculation

As StatView calculates statistics, it changes the cursor to a rotating yin-yang shape:



This indicates that StatView is calculating results. To stop a calculation, press Command-Period. This both stops the calculation and closes the view window. By closing the view window, you are guaranteed that StatView will never display partially-calculated results.

So far, all of the analyses shown have been performed on all the records in a dataset. This is the most common case when you begin to analyze a dataset. However, after you have looked at the entire dataset, you may want to perform analyses on a subset of your data; this is easily done with StatView. When you specify that you are analyzing a subset of your dataset, StatView instantly updates the view window to reflect only that subset.

There are two methods for specifying a subset of a dataset: excluding rows and using ranges. Excluding rows is a task that you do by hand when you want to narrow the dataset based on your own criteria. Using ranges allows you to set automatic restrictions on the records you want included in the analysis, such as all records with Cholesterol greater than 140 or all records with gender equal to Female.

Excluding Rows

Initially, all records are included in an analysis. To exclude a record, double-click on its record number. To include a record that is excluded, double-click on its number again. When a row is excluded, its row number is gray:

6	S. Kaufman
---	------------

Included

6	S. Kaufman
---	------------

Excluded

You can exclude or include many rows at once by selecting the rows and choosing **Exclude Rows** or **Include Rows** from the **Variables** menu. To quickly include all excluded rows, choose **Select All Rows** from the **Edit** menu then **Include Rows** from the **Variables** menu.

Ranges

When you want to systematically exclude rows from an analysis based on a criterion, use ranges. A range is based on criteria that you set for one or more columns. For instance, you might define a range that excludes all males and anyone who weighs less than 95 pounds. After you have specified a range, you can edit the range to change the criteria.

To create a range, select the columns which you will use as criteria and choose **Select Range** from the **Tools** menu. StatView prompts you for the criteria for each selected column from left to right. If the columns you wish to use are not contiguous you must select and specify the range for each column separately. For category columns, the dialog lets you select the elements to include; for numerical columns, the dialog lets you specify a range.

For each column, you can specify whether to combine that criterion with the other criterion using a logical AND or logical OR. Logical AND means that both criteria must be met for the record to be included; logical OR means that either criterion may be met for it to be included. For instance, to find all patients who are males and weigh over 150 pounds, you would use AND. To find all people who either smoke cigarettes or weigh less than 100 pounds, you would use OR.

After you have created a range, you can add additional restrictions by selecting more columns (or the same columns again) and choosing **Select Range** from the **Tools** menu. This adds the criteria to the previous ones. At any time, you can change the criteria with the **Edit Range** command. To remove all range restrictions, select **Clear Range** from the **Tools** menu. Editing any cell also clears the range. Thus, a range only exists as long as you do not change any of the data in your dataset.

For example, to analyze only those records of tall smokers in Lipid Data:

- Open Lipid Data and select the Height column.
- Choose **Select Range** from the **Tools** menu. StatView prompts you for the criteria for the column:

Include rows based on the range of values in column:
--Height

lower bound:

59

☒ ≤
☐ <

H

☒ ≤
☐ <

upper bound:

80.25

Done

Cancel

Currently included data ☒ **AND** the above restriction
☐ **OR**

- Enter 73 for the lower bound.
- Click **Done**.

You have just excluded those records whose height values do not fall between 73 and 80.25 inches.

- Select the Smoking History column.
- Choose **Select Range** from the **Tools** menu. StatView prompts you for the criteria for the column:

Include rows based on the range of values in column:
--Smoking History

no	↑
quit	
cigarettes	
cigars	
pipes	↓

Currently included data ☒ **AND** the above restriction
☐ **OR**

Done
Cancel

- Click on cigarettes and drag down to pipes to select these three elements. Note that you can select discontinuous elements by holding down the Command key while you click.
- Click Done.

StatView now excludes all rows that do not meet both of these criteria. You can add another criterion to the range by selecting the desired column and using **Select Range** again. For instance, if you want to examine statistics only for tall heavy smokers:

- Select the Weight column.
- Choose **Select Range** from the Tools menu. StatView prompts you for the criteria.
- Enter 180 for the lower bound.
- Click Done.

You can modify your range with the **Edit Range** command. This command prompts you for each of the criteria you have selected, allowing you to change each. Assume that you now want to investigate only short, heavy smokers:

- Select **Edit Range** from the Tools menu. StatView prompts you for the first criterion, Height:

Range Restriction #1

Column: Height

Boolean: AND

Restriction:
 $73 \leq H \leq 80.25$

Edit
Next
Exit

- Click Edit. The dialog is the same as when you made the criteria:

Include rows based on the range of values in column:
--Height

lower bound: ☒ \leq ☐ $<$ H ☒ \leq ☐ $<$ upper bound:

Currently included data ☒ AND ☐ OR the above restriction

- Enter 0 for the lower bound. Tab over to the upper bound box.
- Enter 67 for the upper bound.
- Click OK.
- Click Exit to stop editing the criteria.

The logical OR operator is useful if you want to analyze data in two ranges in one column. For instance, you might want to look at all people whose height is greater than or equal to 73 inches or whose height is less than or equal to 63 inches. Instead of clicking AND when you create your criteria, you would click OR.

- Select Clear Range from the Tools menu.
- Click OK to clear the previous range.
- Select the Height column.
- Choose Select Range from the Tools menu.
- Enter 73 for the lower bound.
- Click Done. This sets one criterion, tall people.
- Choose Select Range from the Tools menu. Since the Height column is still selected, this allows you to set a second criterion for height.
- Enter 63 for the upper bound.
- Click OR.
- Click Done.

If you investigate the records, you can see that both tall and short people are included while others are excluded.

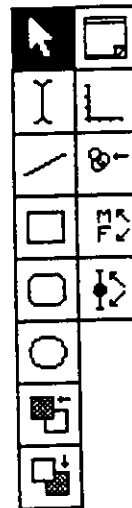
Chapter 4 — Graphs and Drawing with StatView

So far, you have only briefly seen StatView's powerful graphing and drawing capabilities. This chapter shows you how you can use StatView to generate presentation quality graphs and to modify your graphs to make them clearer. StatView's capabilities allow you to create your entire presentation inside your statistical analysis software.

As you read this chapter, you should note the difference between the two types of drawing:

- Creating a graph from your statistics and data is a mechanical routine that StatView does for you. When you select a graph, StatView automatically draws the axis, tick marks, data points, legend, and so on. You will find that most of the graphs that StatView creates are so complete that you do not need to modify them at all.
- Modifying a graph is easy. You can change features of the graph (such as the colors used, the point types, the range of the axes, and so on) and can add information to the graph (such as pictures, additional text, arrows to highlight important data, and so on). The tools you use to add information to the graph are almost identical to those used in popular programs such as MacDraw.

The tools palette on the left side of the screen has two columns.



The left column contains *drawing tools* for selecting and adding information to your graph. The right column contains *view controls* for changing graph attributes and modifying what you see in your graph, such as the axes and points.


Graphs You Can Make

Note that this "palette" is available at all times when a graph is shown. As you read this chapter, remember that the drawing tools and view controls appear in the view window, not in the Macintosh's pull-down menus.

Note: The tools palette is not copied or printed with the graph.

StatView gives you a very wide variety of graphs. Since graphs are related to the types of data you are using, there are often many relevant graphs for a particular set of data. This section lists all the types of graphs you can produce with StatView.

The following sections tell you what you need to do in order to create graphs. The elements you must specify are:

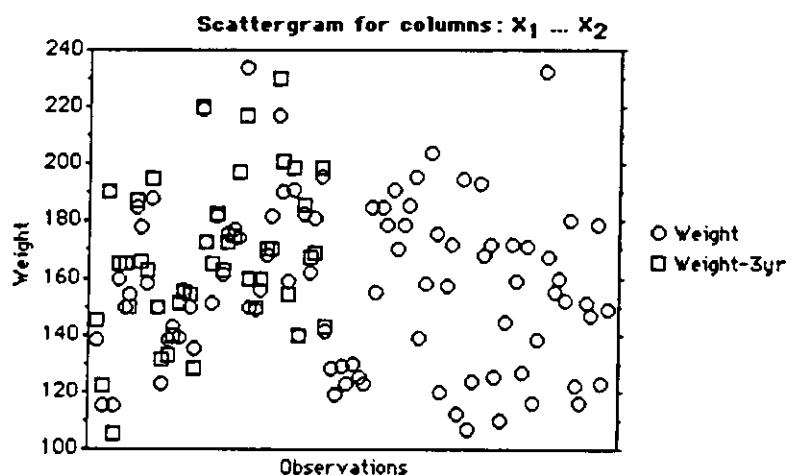
Variables	The number of columns selected, and whether they are specified as X or Y variables.
Statistics	Choices from the Describe or Compare menu, such as Percentiles or Regression .
View	Specify the type of graph you want to see from the View menu, such as Scattergram or Box Plot .
Composite/paging tool 	This is an important tool for most views. It is described later in the chapter.

Univariate Chart

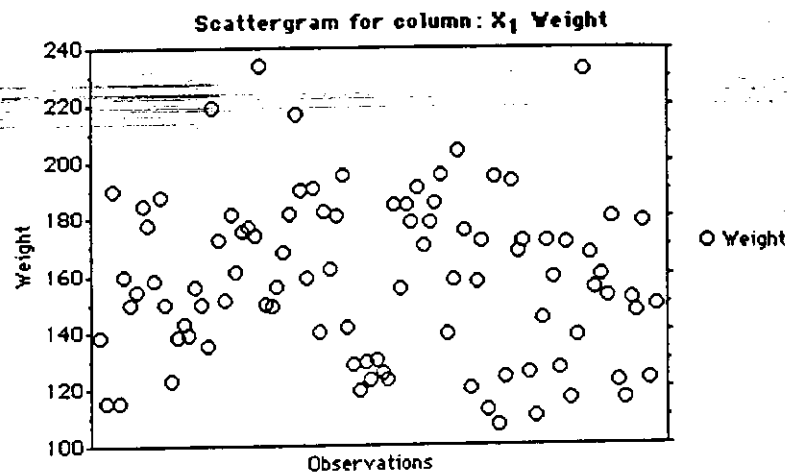
Univariate plots present one-dimensional data. They have only an ordinate (Y axis), as there are no values for the abscissa. Each individual observation is plotted.

To create a univariate chart, you need to assign at least one X column and no Y columns. If you assign more than one X column, you may overlay all the X variables on one composite graph or view each individual X column on a paging graph. Select **None** in the **Describe** and **Compare** menus. In the **View** menu, you may select either **Scattergram**, **Bar Chart**, or **Line Chart**.

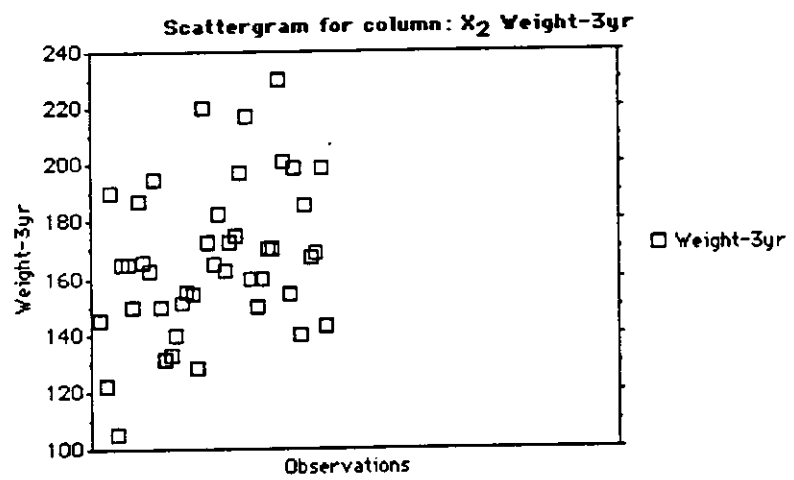
A univariate scattergram comparing two X variables looks like:



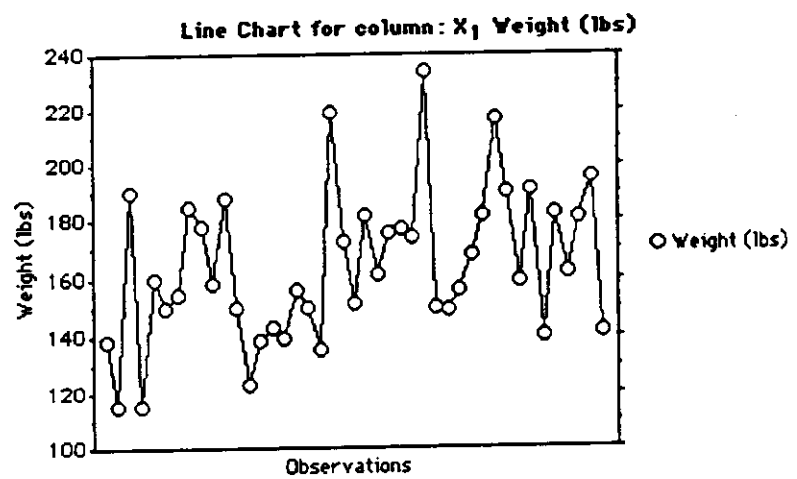
If you click the composite/paging tool, each variable will appear in a separate page. The first page for the previous graph looks like:



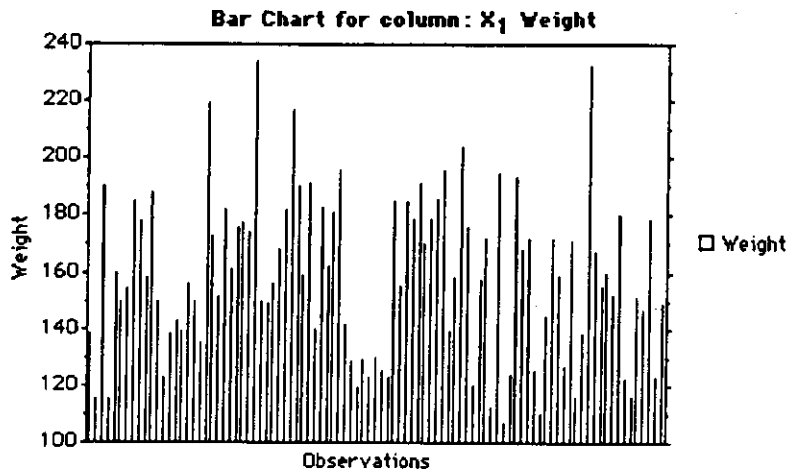
The second page looks like:



A univariate line chart looks like:

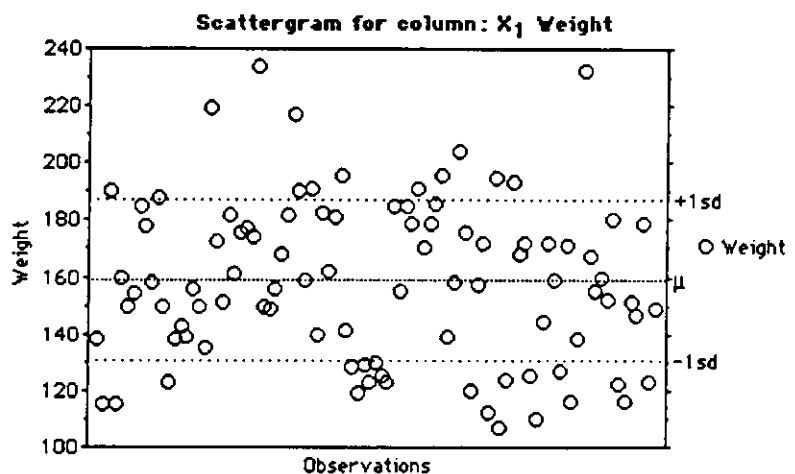


A univariate bar chart looks like:



There is no composite view available for the bar chart.

If you select **Mean**, **Std. Dev.**, etc. from the **Describe** menu while in scattergram/paging view a standard deviation error band and the mean are noted by lines and marked on the right ordinate. A univariate scattergram with mean and standard deviation lines looks like:



±sd

The standard deviation control is unique to this display. This control, **±sd**, allows you to specify the width (in standard deviations) of the band displayed by the standard deviation lines.

If you set the composite/paging tool to composite you will produce a plot of means with error bars around them (see the section on Error Bars in this chapter for further information).

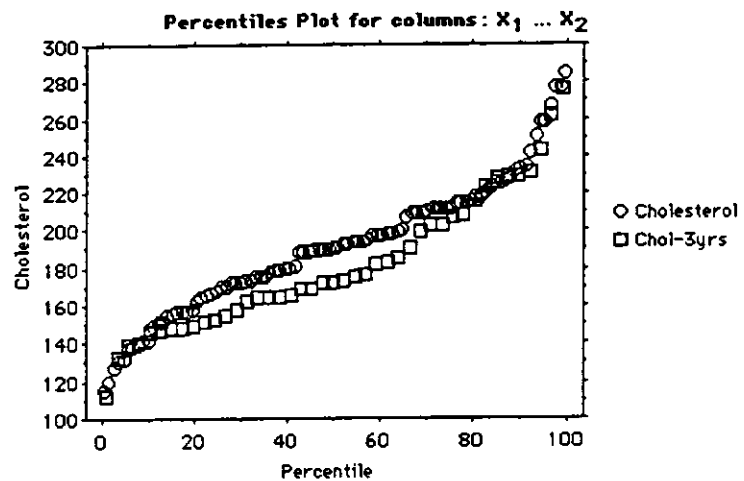
Percentile Plot (Cumulative Frequency Curve)

A percentile plot plots observed values against their percentiles. It allows you to quickly estimate the percentile associated with any observed value in a distribution. There are several different views available.

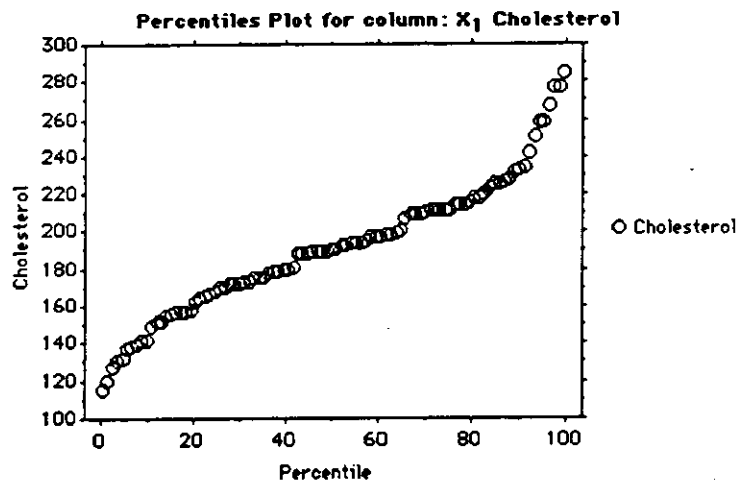
To create a percentile graph, assign X to one or more columns. If you assign X to more than one column, you may overlay all the X variables on one composite graph or view each individual X variable on a paging graph. Select **Percentiles** in

the Describe menu. In the View menu, select Scattergram, Bar Chart, or Line Chart.

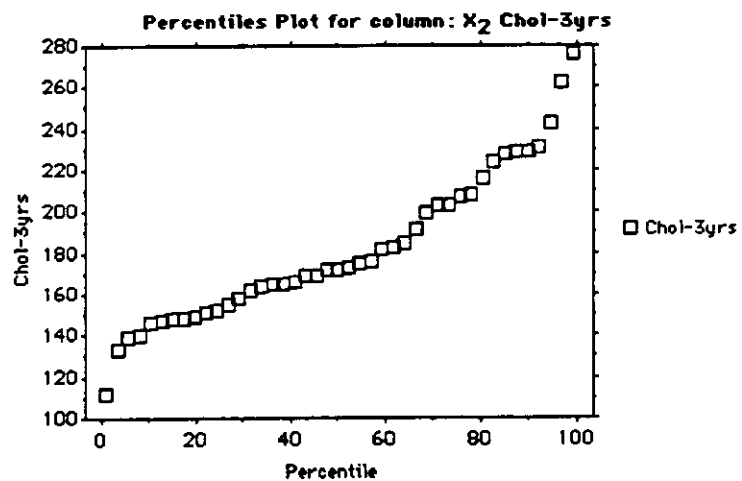
A composite percentile graph comparing two X variables looks like:



If you click the composite/paging tool, each variable will appear in a separate page. The first page for the previous graph looks like:

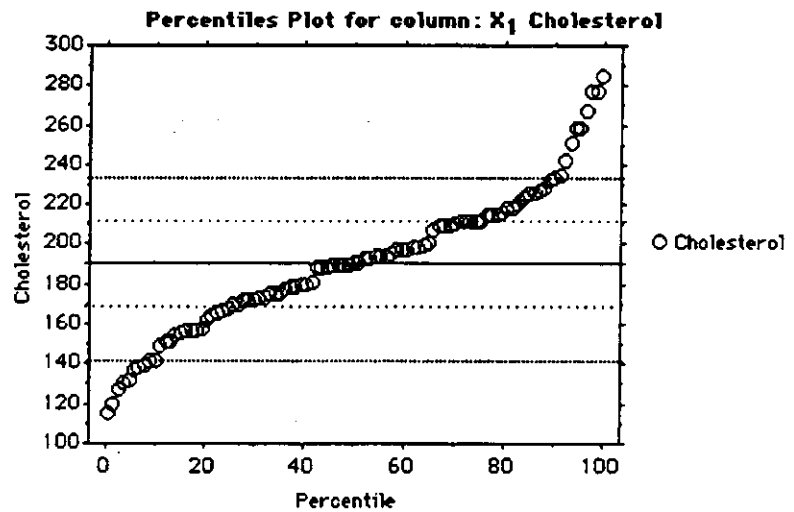


The second page looks like:

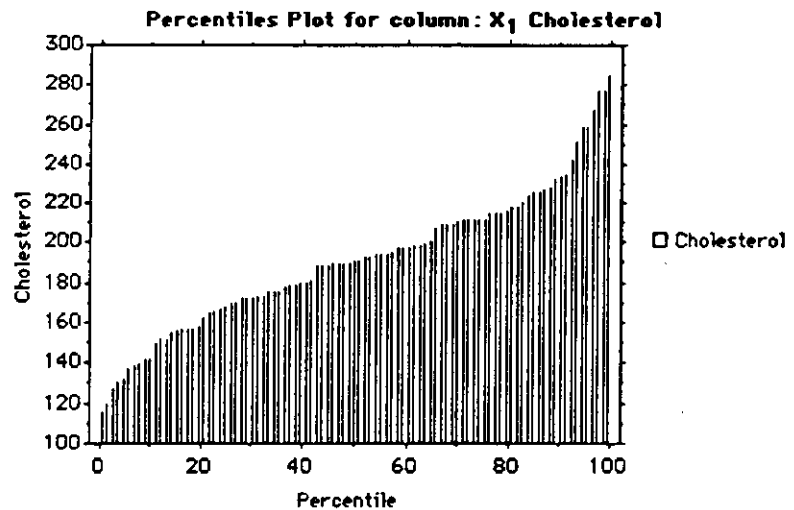




Clicking the percentile control forces horizontal lines to be displayed on the plot representing the five percentile values (10th, 25th, 50th, 75th, and 90th). This control, which is unique to percentile plots, is only available on the paging view.



You can view this graph as a line chart or as a bar chart:



There is no composite view available for bar charts.

Scattergram

There are two types of scattergrams. A univariate scattergram plots one variable as observations and is discussed earlier in this chapter. A bivariate scattergram plots the relationship between two variables, X and Y. The types of bivariate scattergrams are:

- regular scattergram
- scattergram with fitted simple regression
- scattergram with fitted polynomial regression

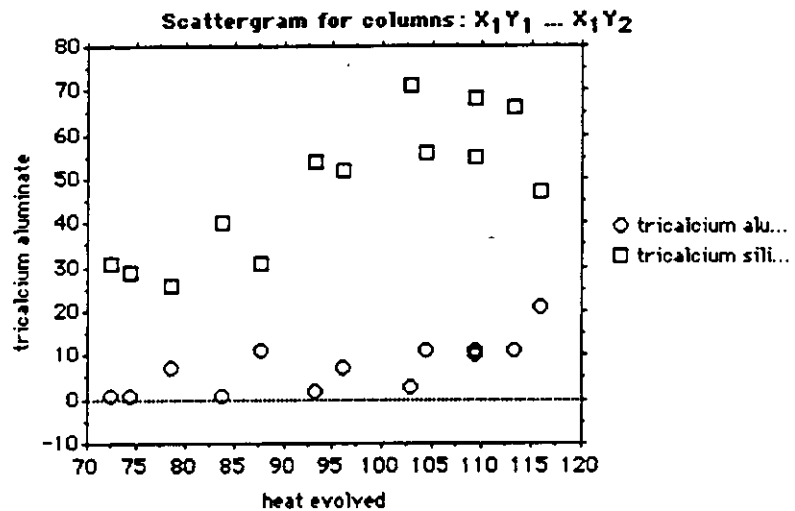
To create a scattergram, assign X to one or more columns and assign Y to one or more columns.

If more than one X and more than one Y variable are assigned, Y_1 is plotted against X_1 , Y_2 against X_2 , and so on. If three X variables and four Y variables are assigned, the Y_4 variable (with no matching X variable) is not graphed. If there is a single X variable and more than one Y variable, each Y variable is plotted against the X variable. The same situation occurs if there is a single Y variable and more than one X variable.

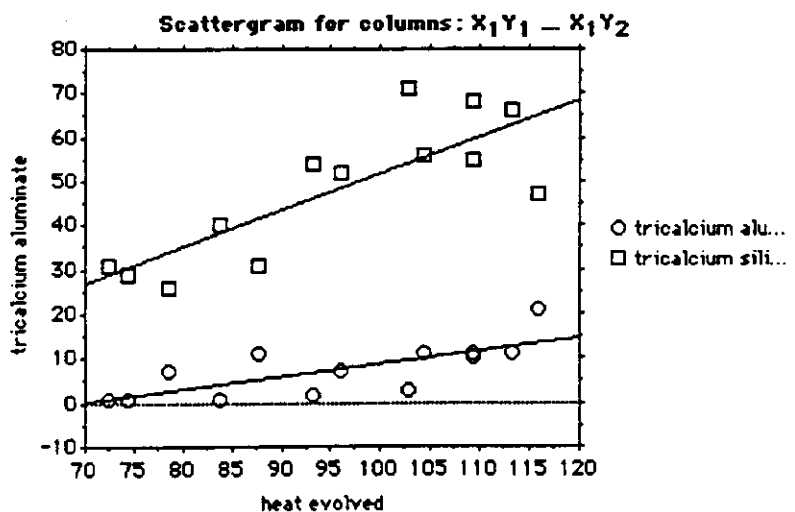
If you have assigned more than one X-Y pair, you may overlay all the variable pairs on one composite graph or view each individual X-Y pair on a paging graph. The following sample graphs were created using the Hald sample data:

Select **None** in the **Describe** menu. In the **Compare** menu select **None** or select **Regression** if you wish a regression line fitted to your data. In the **Regression** dialog, select **Simple** or **Polynomial**. In the **View** menu, select **Scattergram**.

A regular scattergram with two Y variables plotted against one X variable looks like:




A simple regression fitted to this data looks like:

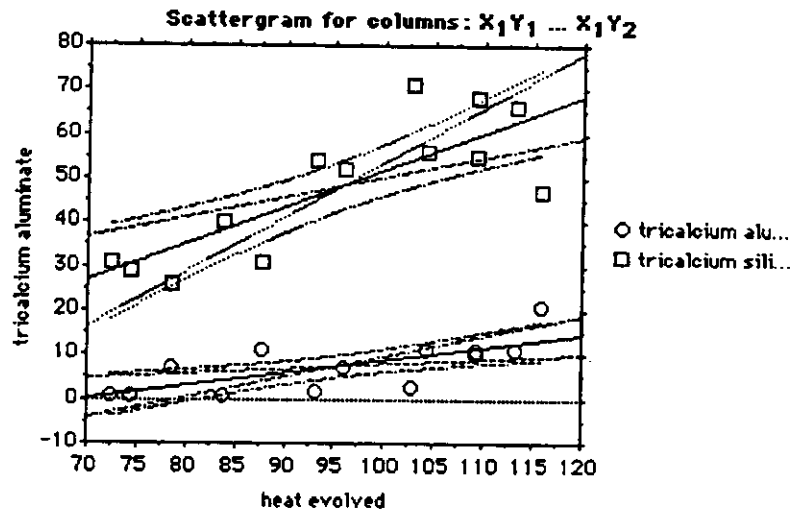


The coefficients of the equations for these regression lines are located in the table views and are discussed in Chapter 6. The equation of each of these lines is shown above the graph in the scattergram/paging view

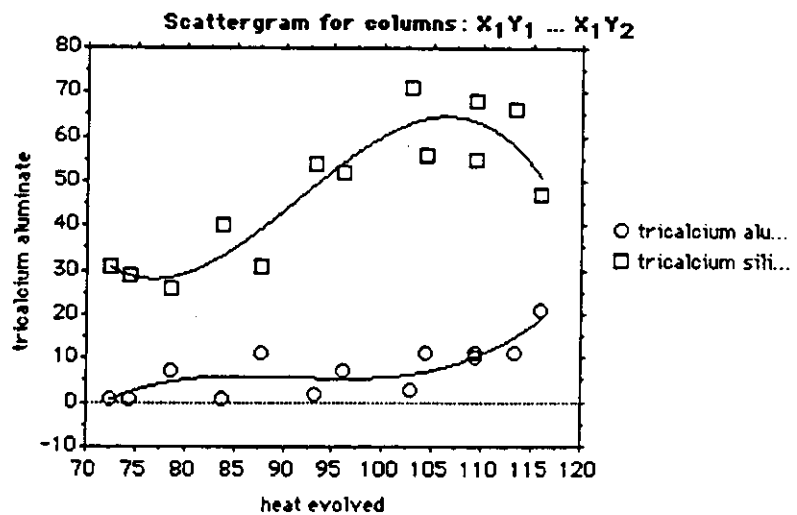
The confidence bands control is unique to a scattergram with fitted simple



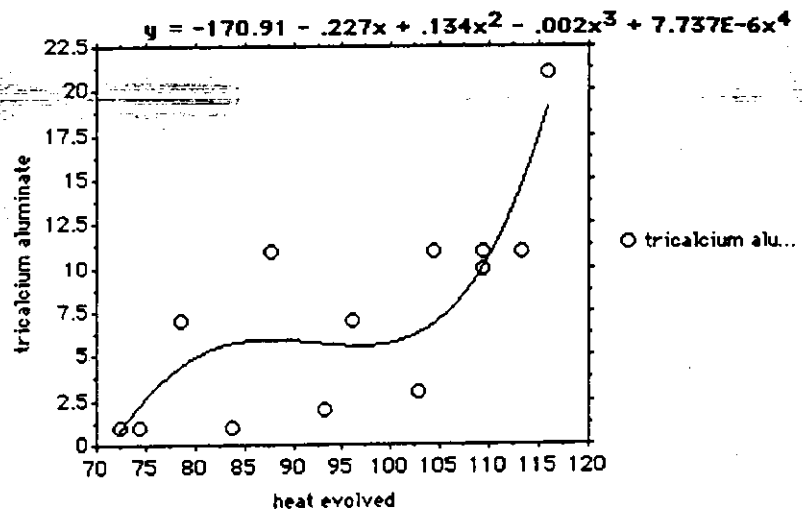
regression. This control, , allows you to display confidence bands for the mean and slope of the fitted regression.



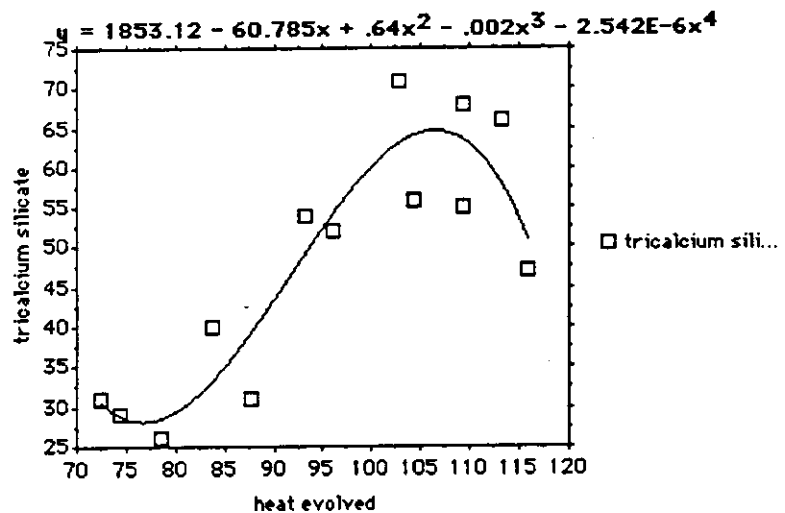
A fourth-order polynomial regression fitted to this data looks like:



If you click the composite/paging tool, each variable pair will appear on a separate page. The first page shows:



The second page shows:



Comparison Percentile Chart

A comparison percentile chart compares 19 corresponding percentiles of two variables. The percentiles compared are: 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 96, 97, 98, 99. This chart is extremely effective for comparing two data distributions.

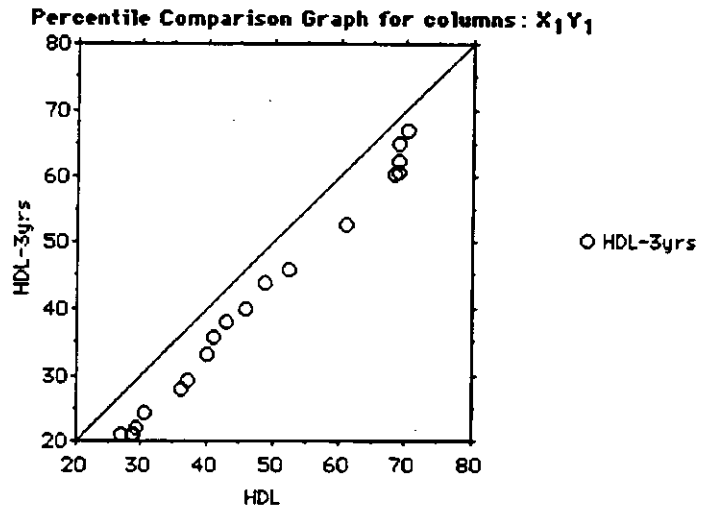
To create a comparison percentile chart, select one or more columns as X variables and one or more column as Y variables.

If more than one X and more than one Y variable are assigned, Y_1 is plotted against X_1 , Y_2 against X_2 , and so on. If three X variables and four Y variables are assigned, the Y_4 variable (with no matching X variable) is not graphed. If there is a single X variable and more than one Y variable, each Y variable is plotted against the X variable. The same situation occurs if there is a single Y variable and more than one X variable.

If you have assigned more than one X-Y pair, you may overlay all the variable pairs on one composite graph or view each individual X-Y pair on a paging graph.

In the Describe menu you should have None selected. In the Compare menu select Compare Percentiles. In the View menu, select Scattergram or Line Chart.

A comparison percentile scattergram with one X variable compared to one Y variable looks like:

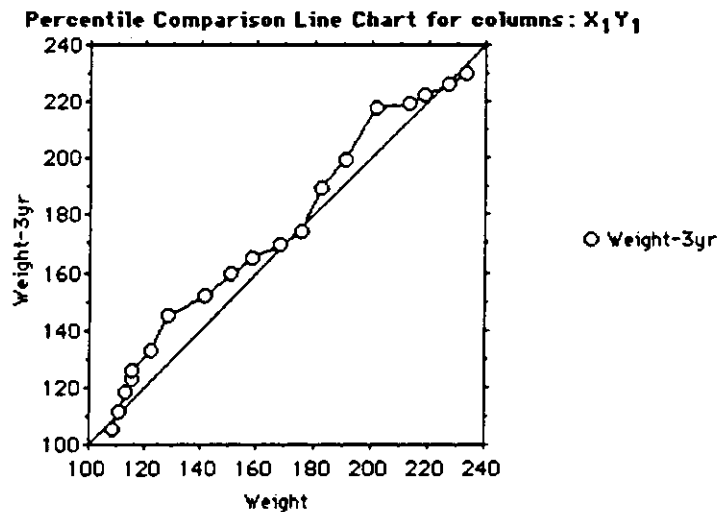


The equal axes control is unique to a comparison percentile chart. This control,



, allows you to cause the comparison percentile chart to be a square chart displaying the line $y = x$.

A comparison percentile line chart with one X variable compared to one Y variable looks like:



Line Chart

A line chart plots the relationship between two variables, X and Y. It is one of the best displays of a set of measurements of a variable through time. The line chart can be drawn with plotting symbols to clearly distinguish individual data points, or without plotting symbols

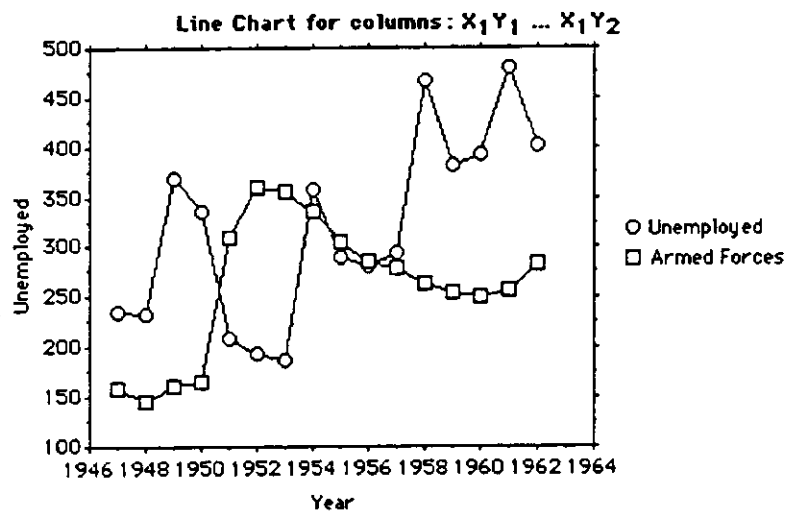
To create a line chart, assign one or more columns as X variables and one or more columns as Y variables.

If more than one X and more than one Y variable are assigned, Y_1 is plotted against X_1 , Y_2 against X_2 , and so on. If three X variables and four Y variables are assigned, the Y_4 variable (with no matching X variable) is not graphed. If there is a single X variable and more than one Y variable, each Y variable is plotted against the X variable. The same situation occurs if there is a single Y variable and more than one X variable.

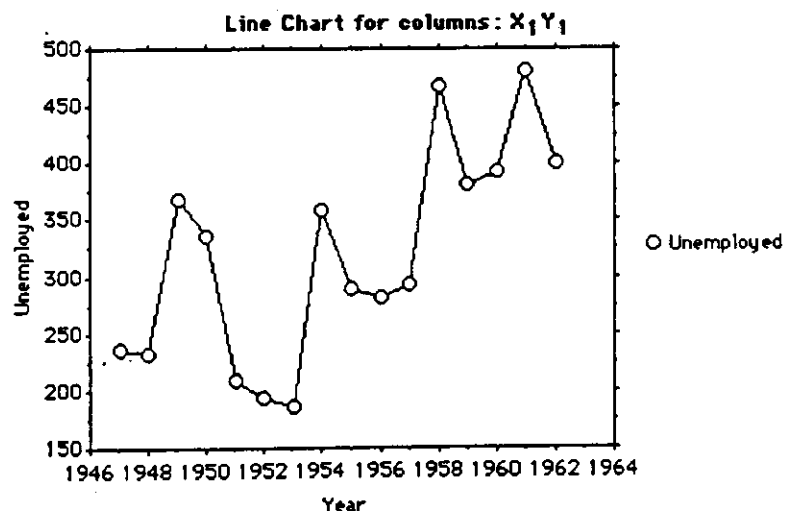
If you have assigned more than one X-Y pair, you may overlay all the variable pairs on one composite graph or view each individual X-Y pair on a paging graph.

The following sample graphs were created using the Longley sample data. In the **Describe** and **Compare** menus, you should have **None** selected. In the **View** menu, select **Line Chart**.

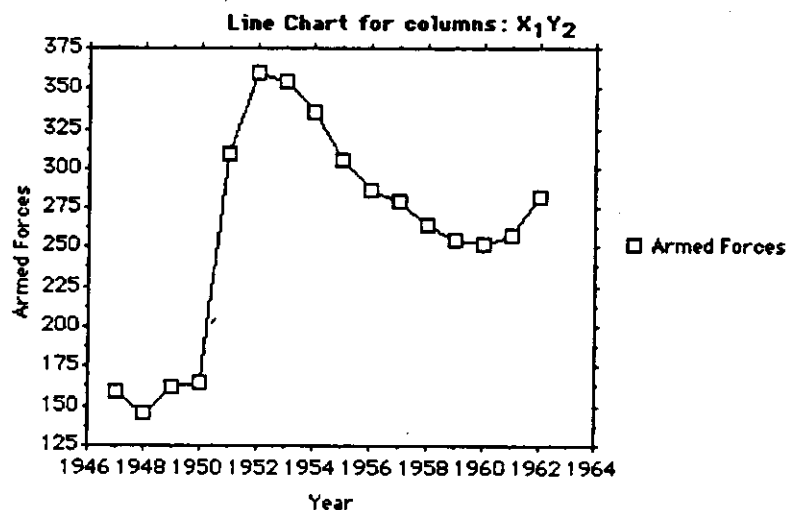
A regular line chart with two Y variables plotted against one X variable looks like:




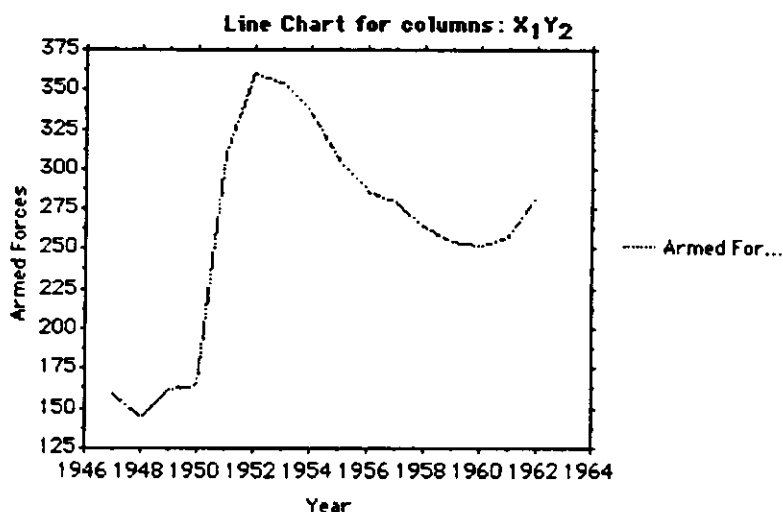
If you click the composite/paging tool, each variable pair will appear on a separate page. The first page looks like:



The second page looks like:



The no symbols control is unique to a line chart. This control, , allows you to specify whether or not you wish the line chart to contain plotting symbols. A line chart with one Y variable plotted against one X variable and no plotting symbols looks like:



Bar Chart

A bar chart plots the relationship between two variables, X and Y. Like the line chart, it best displays a set of measurements of a variable through time.

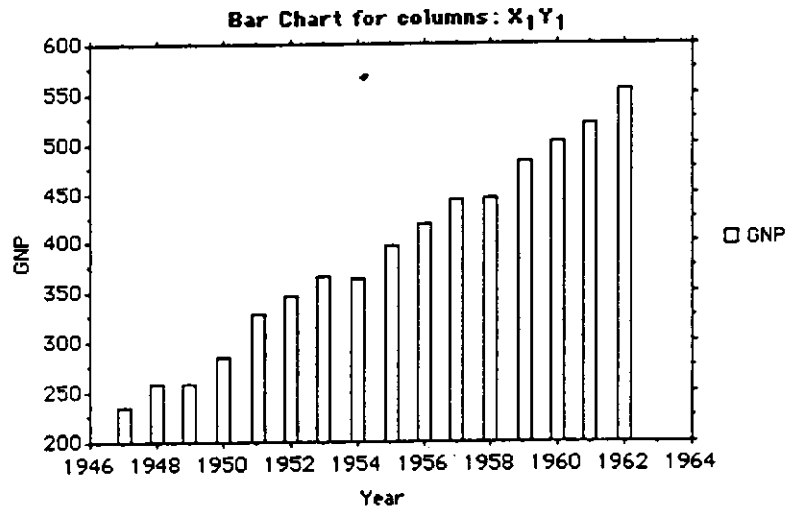
To create a bar chart, select one or more columns as X variables and one or more columns as Y variables.

If more than one X and more than one Y variable are assigned, Y₁ is plotted against X₁, Y₂ against X₂, and so on. If three X variables and four Y variables are assigned, the Y₄ variable (with no matching X variable) is not graphed. If there is a single X variable and more than one Y variable, each Y variable is plotted against the X variable. The same situation occurs if there is a single Y variable and more than one X variable.

The bar chart can only display one individual X-Y pair per page; there is no composite view.

The following sample graphs were created using the Longley sample data. In the **Describe** and **Compare** menus, you should have **None** selected. In the **View** menu, select the **Bar Chart** command.

A bar chart with one Y variable plotted against one X variable looks like:



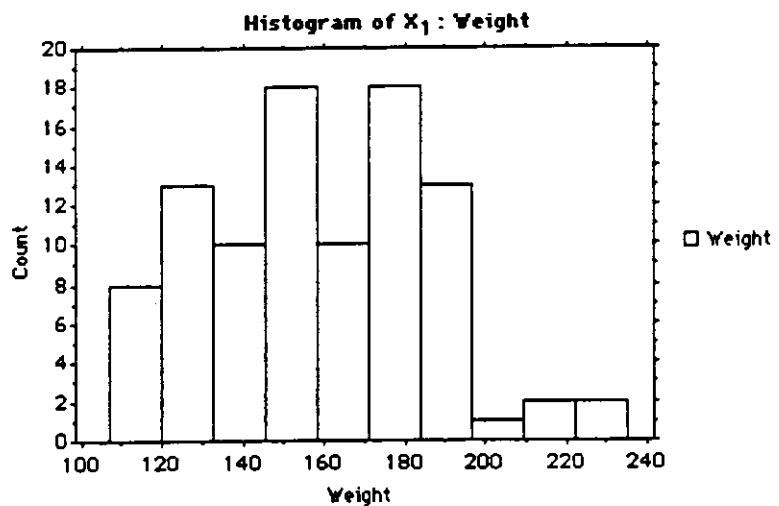
To get a bar chart with error bars, you must first start with a scattergram, then use the error bar tool. See the section titled "Error Bars" which appears later in this chapter for more information.

Histogram

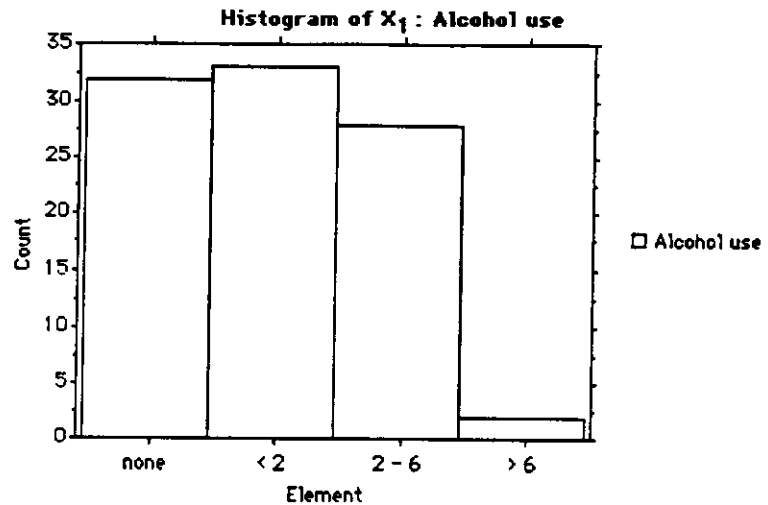
A histogram displays the frequency distribution of a variable.

To create a histogram, assign one or more columns as X variables. In the **Describe** menu, select **Frequency Distribution**. In the **View** menu, select **Bar Chart**. The histogram can only display one individual X variable per page; there is no composite view.

A histogram of continuous data looks like:



A histogram of category data looks like:

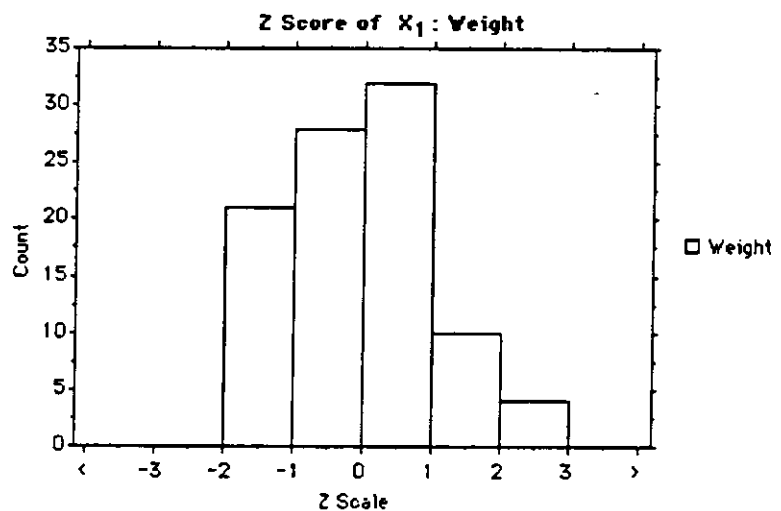


z-Score Histogram

A z-score histogram displays the z-score frequency distribution of a variable.

To create a z-score histogram, select one or more columns as X variables. In the Describe menu, select **Mean**, **Std. Dev.**, etc., or **Confidence Intervals**. In the View menu, select **Bar Chart**. The composite/paging tool should be set for paging.

A z-score chart looks like:



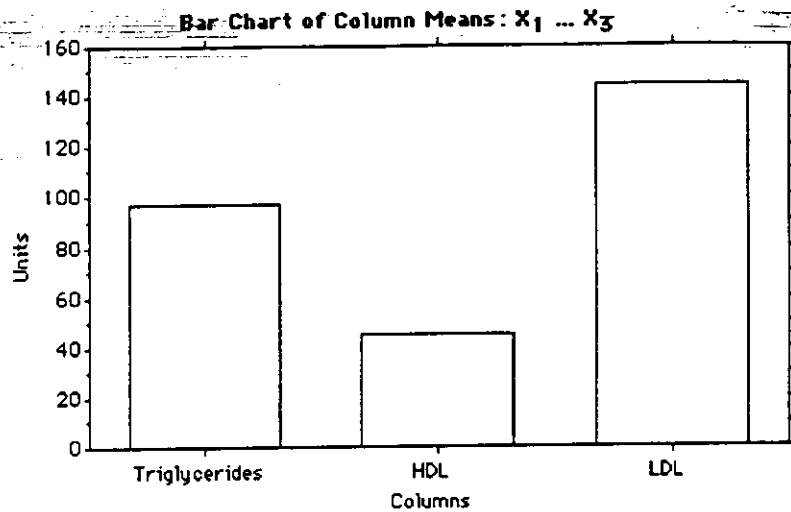
If you set the composite/paging tool to composite, you will produce a comparative bar chart, described below.

Comparative Bar Chart

A comparative bar chart compares the mean values from similar variables. The X axis identifies the columns while the Y axis displays numerical values.

To create a comparative bar chart, select one or more columns as X variables. In the Describe menu, select **Mean**, **Std. Dev.**, etc. or **Confidence Intervals**. In the View menu, select **Bar Chart**. The composite/paging tool should be set for composite.

A comparative bar chart looks like:



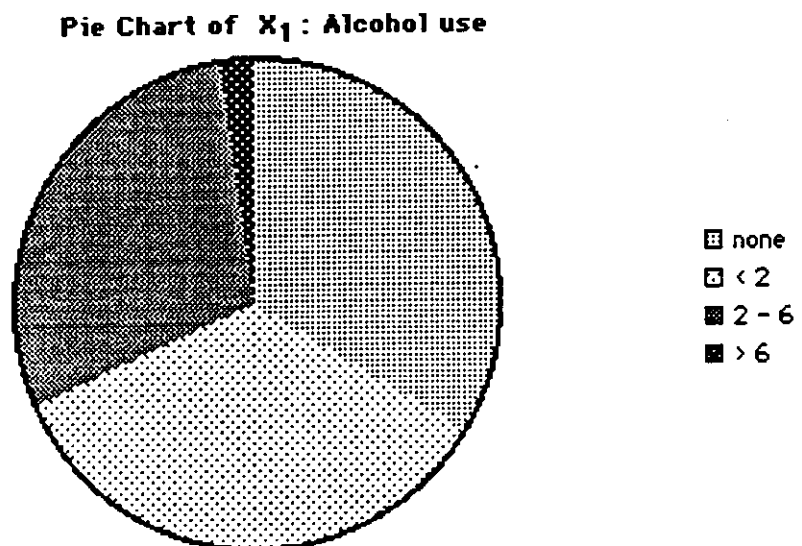
If you set the composite/paging tool to paging you will produce a z-score histogram of each X variable (described earlier).

Pie Chart

Pie charts allow you to visually compare categories to each other. They are ideal for displaying information about the category variables of a StatView dataset.

To create a pie chart, select one or more columns as X variables. In the Describe menu, select **Frequency Distribution**. In the View menu, select **Pie Chart**. The pie chart can only display one individual X variable per page; there is no composite view.

A pie chart of category data looks like:

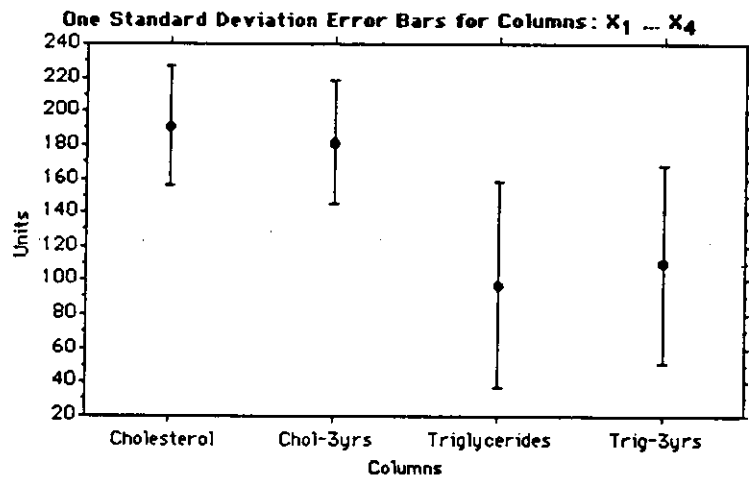



Error Bars

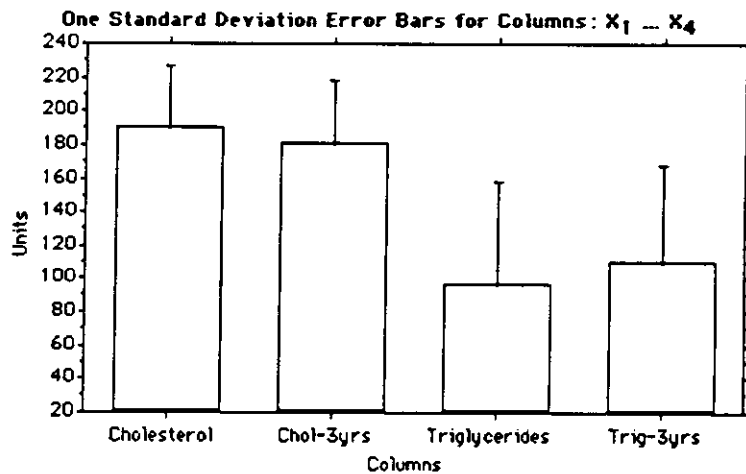
Error bars compare the distribution of variables by displaying variation around sample means. StatView creates one-, two-, or three-tiered error bars portraying confidence intervals or the standard deviation of a variable.

To create error bars, select one or more columns as X variables. In the **Describe** menu, select **Mean**, **Std. Dev.**, etc. or **Confidence Intervals**. In the **View** menu, select **Scattergram** or **Line Chart**. The composite/paging tool should be set for composite.

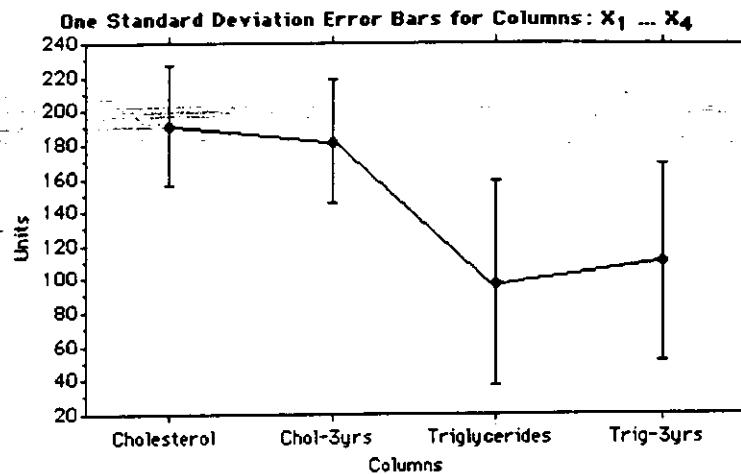
Error bars look like:



The bar control, , is unique to error bars displayed in a scattergram view. It allows you to add a bar to the bottom of your error bars to create a bar chart with error bars attached. Error bars with bars added look like:



Connected error bars in a line chart look like:



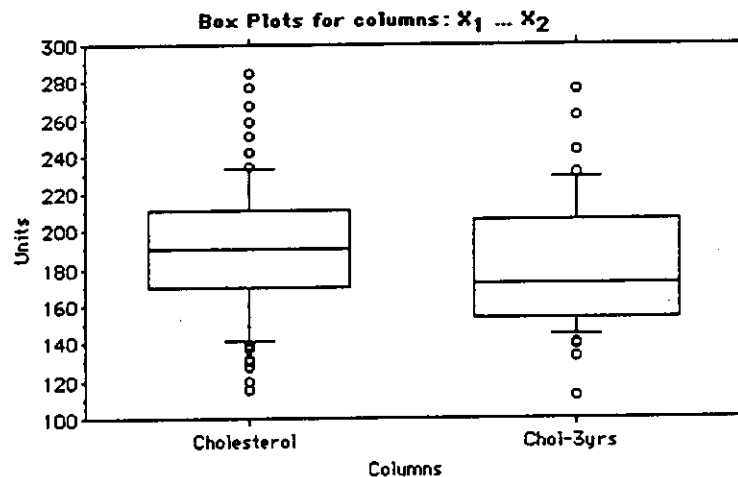
If you select a scattergram view and click the composite/paging tool to paging you will produce a univariate scattergram for each X variable (described earlier).

Box Plots

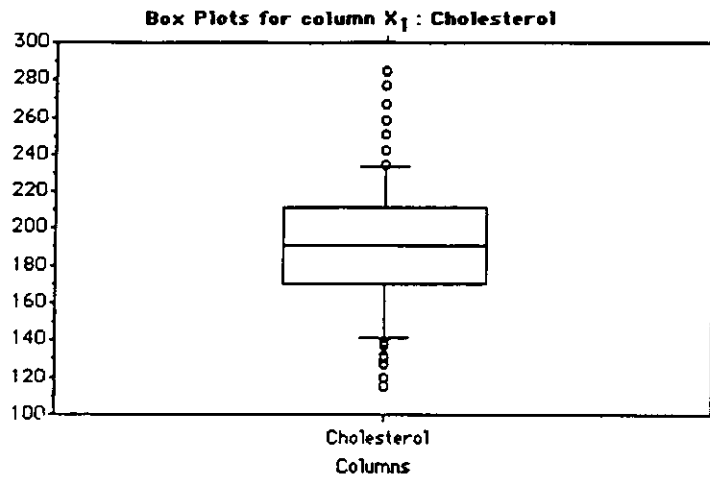
A box plot is a graphic method for displaying the 10th, 25th, 50th, 75th, and 90th percentiles of a variable. It is often used for comparing variable distributions.

To create a box plot, select one or more columns as X variables. If you have more than one X column assigned, you may overlay all the box plots on one composite graph or view each individual box plot on a paging graph. In the Describe menu, select Percentiles. In the View menu, select Box Plot.

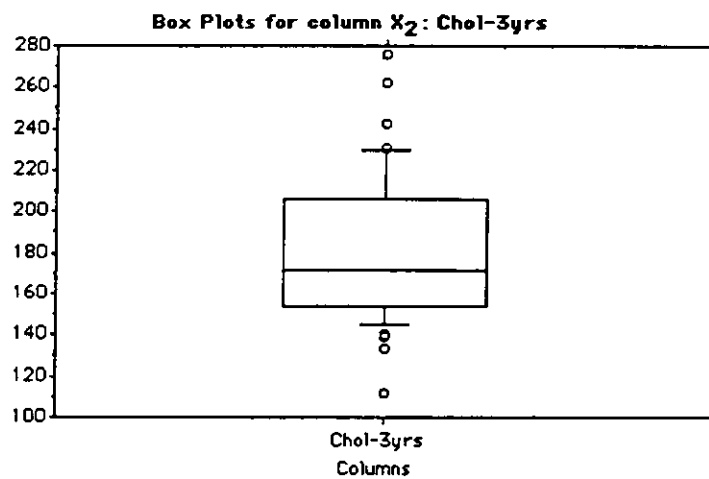
A box plot comparing two variables looks like:




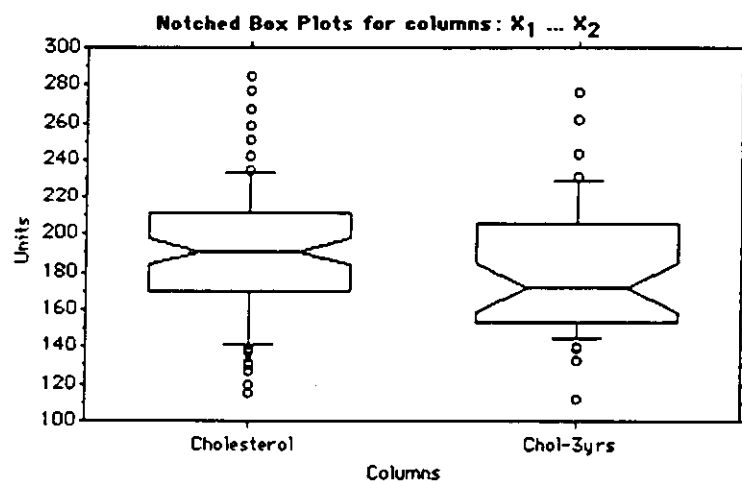
If you click the composite/paging tool, each variable will appear on a separate page. The first page looks like:




The second page looks like:























Two controls appear on this view which are unique to box plots. The notch control, , changes the box plot to a notched box plot, where the notches represent 95% confidence bands about the median.



Modifying Graphs

The outlier control, , eliminates the representation of the extreme twenty percent of the observed values, ten percent below the 10th percentile and ten percent above the 90th percentile.

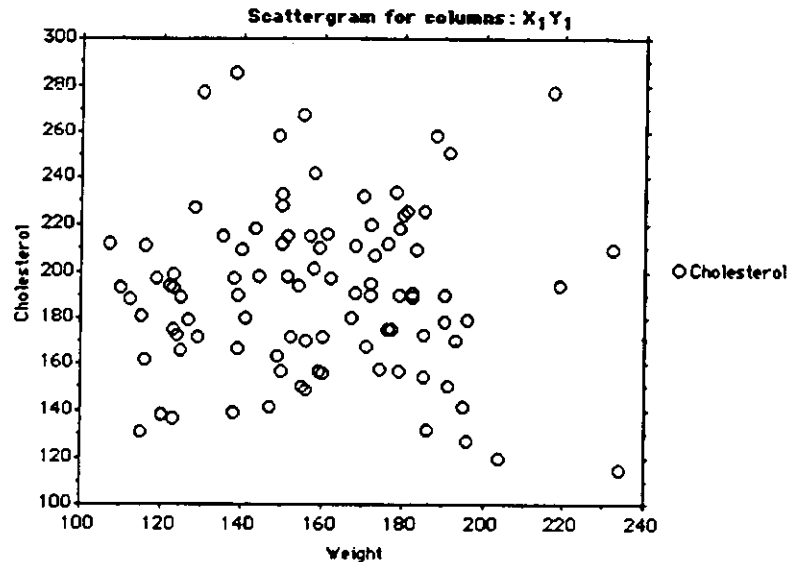
You can modify your graphs using the tools palette on the left side of the screen. Different tools are available depending on the view and analysis you select. These tools are *view controls*, they change graph attributes and modify what you see in your graph. Note that this “palette” is available at all times when a graph is shown, but not copied or printed with the graph. The tools in the palette are:

		Composite/paging tool
		Frame/Unframe control
		Bar chart/No bar chart control
		Percentile control
		Symbols control
		Outlier control
		Equal axes control
		Point overlap control
		Subset specify control
		Add error bar control
		Standard deviation control
		Confidence band control
		Box plot notch control

Many of the view controls are *toggles*; they switch between two settings. For these controls, the visible icon specifies which of the two possible modes you are in. When you click the control, it toggles to the other mode. The other controls are



parameter controls which let you enter values for how the control should act. To see how these controls work:

- Open Lipid Data.
- Select **Quick Assignment** from the **Variables** menu, assign X to Weight and Y to Cholesterol.
- Select **Scattergram** from the **View** menu and **Zoom Up** the view.




Frame

The frame control specifies whether the graph is framed on all four sides or just the left and bottom. Graphs are initially framed. When you are in frame mode, the icon

is . Clicking on the control switches you to unframed mode. The icon in unframed mode is .

Point Overlap

The point overlap control looks like  and is available on scattergrams displaying X-Y pairs. These graphs often contain overlapping points. Overlap occurs when the locations of different data points are either identical or very close to each other.

For example, open and modify Lipid Data so that values overlap.

- Select the first row of the dataset by clicking on its row number.
- Select the **Copy** command in the **Edit** menu.
- Select the input row at the bottom of the dataset by clicking in its leftmost cell.
- Select the **Paste** command in the **Edit** menu.

This creates two data points at the same location. To experiment with handling overlapping points:

- Assign X to Weight and Y to Cholesterol.
- Select Scattergram from the View menu and Zoom Up the view.
- Click on the point overlap tool and this dialog box appears:

Select method of handling point overlap

<p>Determine overlap with:</p> <p><input checked="" type="radio"/> Don't handle overlap</p> <p><input type="radio"/> Exact Coincidence</p> <p><input type="radio"/> Celluation</p> <p>Resolution: <input type="radio"/> Coarse <input checked="" type="radio"/> Medium <input type="radio"/> Fine</p>	<p>Show overlap with:</p> <p><input checked="" type="radio"/> Sunflowers</p> <p><input type="radio"/> Bigger Points</p>
---	--

OK
Cancel

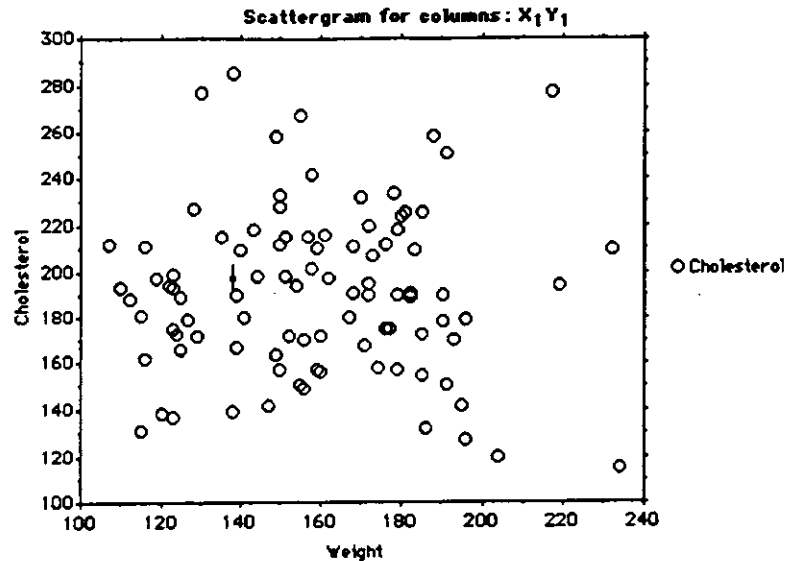
This box offers three choices for displaying overlapping points and two styles for representing those points. **Don't handle overlap** means that StatView does not do anything to indicate hidden points caused by overlap. Points which exactly coincide appear as a single point and points which partially coincide appear as overlapping points.

The two selections below this one are different ways for displaying overlapping points. If you select either of these methods, you must specify how to display the overlap. There are two methods for displaying the points:

Sunflowers	Points with petals (lines) emanating from them. Each petal represents one point at that location. A point with no petal represents a single data point.
Bigger Points	Geometrically enlarged points with enlargement size based on the number of points that coincide at that location. The size of points representing single data points is determined by the point size control. The bigger the starting point, the larger are the enlarged points.

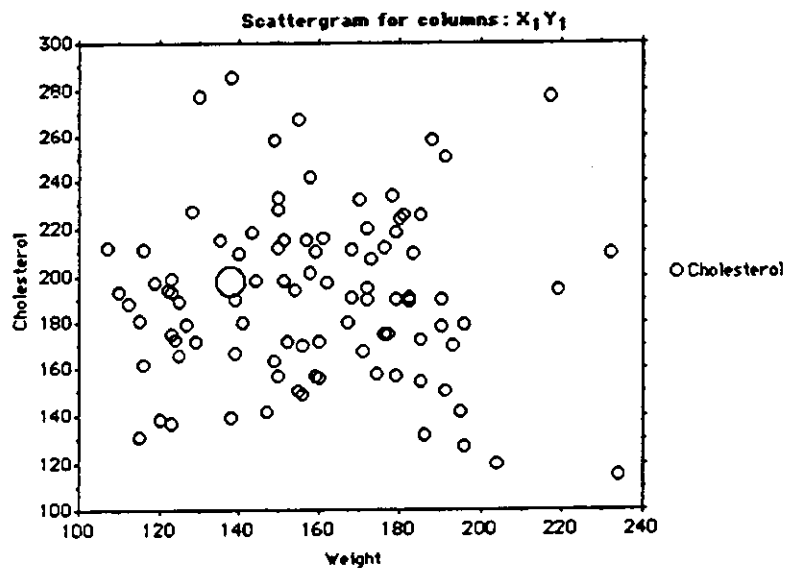
When **Exact Coincidence** is selected, points that overlap exactly are indicated with sunflowers or bigger points.

- Select **Exact Coincidence**.
- Click **OK**, choosing the default, **Sunflowers**, to display the points. This view appears:



There are two data points that overlap exactly, since the first and last records are identical. They are represented by a sunflower with two petals. To see a three point overlap, add the record again.

- Activate the data window.
- Copy the last record and paste it into the gray input row at the bottom of the window.
- Activate the view window; notice that the sunflower now has three petals.
- Click on the point overlap tool.
- Select representation by **Bigger Points**, and click OK.



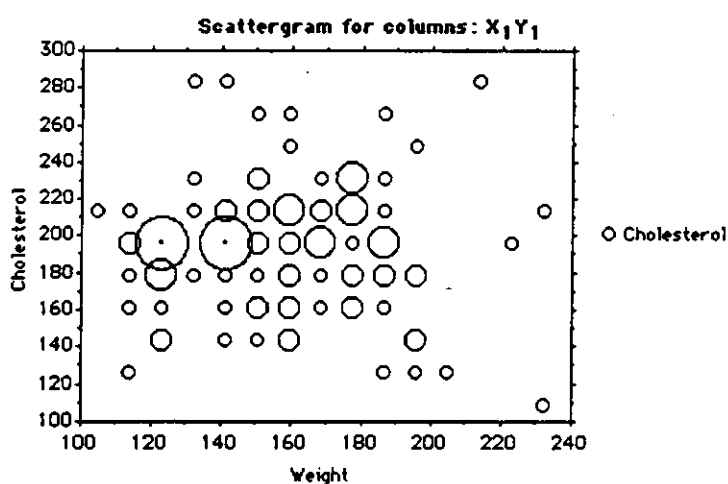
The three points are now represented by a larger data point.

The third choice under the heading **Determine overlap with** is **Cellulation**. This method is best for handling large datasets. When this choice is selected, StatView:

- Divides the scattergram into invisible grid regions.
- Counts the points that occur in each region.
- Represents the population of each region as either a sunflower point (with each petal representing one point in the grid square) or as a geometrically enlarged point. The point is centered in the middle of the square whose population it represents.
- Click on the point overlap tool, select **Cellulation** and leave **Bigger Points** selected.

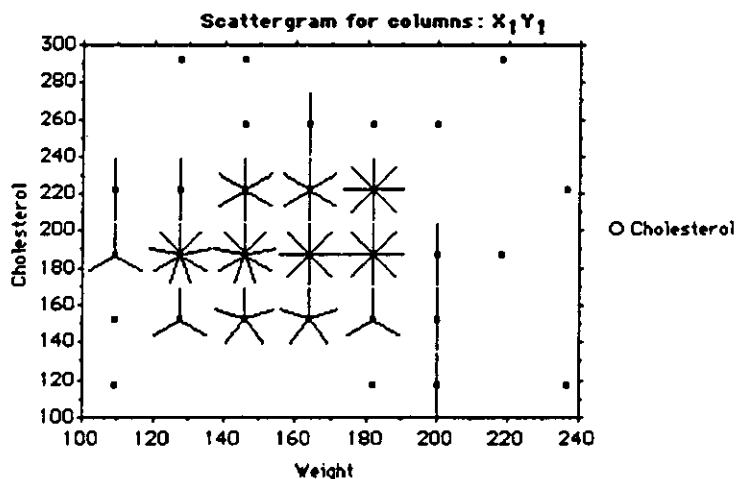
The bottom radio button array titled **Resolution** is activated. It contains the choices: **Coarse**, **Medium**, and **Fine**. These selections determine the size of the region used in cellulation. **Fine** divides the scattergram into small, numerous regions, **Medium** divides the scattergram into fewer medium sized regions, and **Coarse** divides the scattergram into even fewer large regions.

- Click OK, using the default resolution, **Medium**, and this view appears:



The view shows various blocks of circles. You can see how the graph is divided into a grid. The size of the circle shows the number of points within each grid square.

- Click on the overlap tool, select **Coarse**, and use **Sunflowers** to show overlap. Click OK. This view appears:




The graph is now divided into larger grids. The number of petals in the sunflower is the number of points counted in the region.

- When you close Lipid Data do not save the changes made to the dataset in this example. This sample dataset is used in other examples throughout the manual.

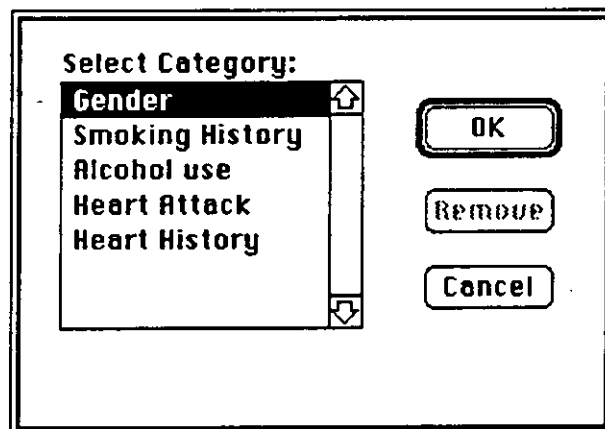
The parameters which you set using the point overlap control remain in effect for all successive scattergrams. Note that if you are handling overlap, the subset specify and error bar controls are no longer active.

Subset Specify



The subset specify control, , allows you to differentiate subsets of values in a scattergram or line chart by a different plotting symbol or color. These subsets are identified by the category variables of the dataset. The subset specify tool is only available when you are in composite mode and no special point overlap techniques are in effect.

- Open Lipid Data.
- Assign X to Weight and Y to Cholesterol.
- Select Scattergram from the View menu
- Click on the subset specify control, and this dialog box appears:



The scrolling list contains the names of all the category variables in your dataset. You can select the category containing the subsets you wish to identify in your graph.

- Click OK, using Gender as the category to determine subsets.


Notice that different points (or color) now distinguish the male and the female data. The legend has also been updated to indicate the subsets in your data.

To remove a subset click the Remove button in the dialog box.

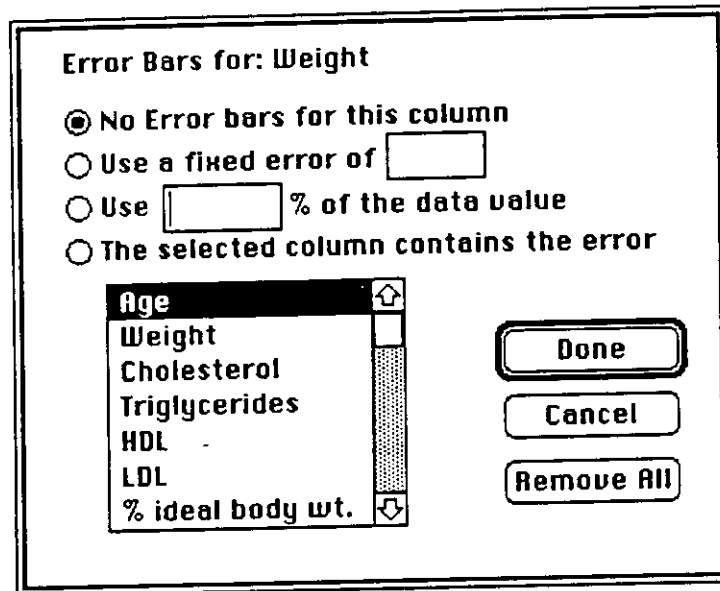
Note that changing or deleting the Y variable assignment, selecting a new statistic or a different view, or toggling from composite to paging will remove any subset specifications.

Error Bars



The error bars control, , allows you to place customized error bars on Y variable observations in scattergrams or on X variable observations in univariate scattergrams. It is available only for composite views.

- Open Lipid Data.
- Assign X to Weight and Y to Cholesterol.
- Select **Scattergram** from the View menu.
- Click on the error bars tool, this dialog box appears:



You may specify error bars values for each Y variable on the graph. The variable currently referenced is noted in the title.

You may specify:

- **No Error Bars** for a column
- **Fixed Error Value** for each data point (value is entered in the text box)
- **% of the Data Value** as the error value (percentage is entered in the text rectangle and applied to each observation)
- **Selected Column Contains the Error Values** (select a column from the scrolling list to contain the error values for the data column. The error value is taken from the same row as the data value with which it is associated.)

Click **Next** to move to the next Y variable. If there are no more columns, this button changes to **Done**.


- Click **Remove All** to remove all error bars from the graph.
- Click **Cancel** to abandon the operation.

If you have chosen a data column as containing the error values, changing values in this column causes the graph to be redrawn. If this column is cut or deleted, no error bars are drawn for the associated data column.

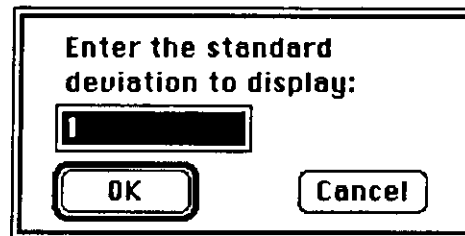
Note that changing or deleting the Y variable assignment, selecting a new statistic or a different view, or toggling from composite to paging will remove error bar specifications.

Error bars for means and confidence bands for X variables are discussed earlier in this chapter in the discussion of Error Bar graphs. There is an example of a line chart with error bars in the Chapter 7 section on Splitting Columns.

Standard Deviation Bands


The standard deviation control, , is unique to univariate scattergrams displaying Mean, Std. Dev., etc in paging mode. This control allows you to specify the width (in standard deviations) of the band displayed by the standard deviation lines.

- Click on the standard deviation control, this dialog appears:

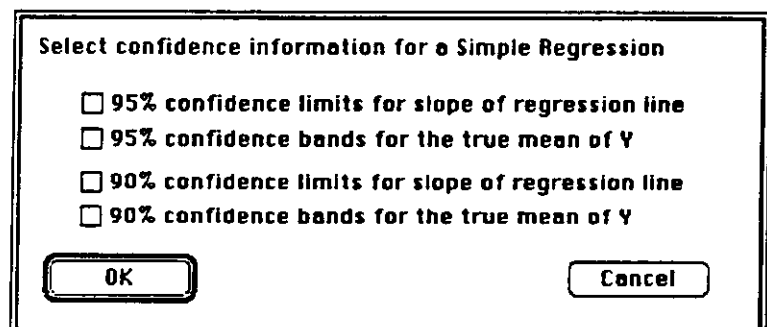


- Enter the width of the band you wish to display and click OK.

Confidence Bands

The confidence bands control is unique to a scattergram with fitted simple regression. This control, , allows you to display confidence bands for the mean and slope of the fitted regression.

- Click on the confidence bands tool; the following dialog box appears:




For both confidence intervals you can plot:


- the confidence limits for the slope of the regression line.

- the confidence bands for the true mean of Y.


Outlier

If you have a box plot, you can specify whether or not to show outliers. Outliers are those value below the 10th percentile and above the 90th percentile. The default


is to show outliers. The outlier control, , eliminates the representation of the the extreme twenty percent of the observed values, ten percent below the 10th percentile and ten percent above the 90th percentile.


After you click this control, it changes to . Clicking this tool adds the outliers again.

Notch

The notch control, , changes the box plot to a notched box plot, where the notches represent 95% confidence bands about the median.



Percentile

The percentile control, , lets you add to or remove from your percentile graph lines representing the 10th, 25th, 50th, 75th, and 90th percentiles. It is available only for paging views. The default is not to draw the lines. Clicking this tool


changes the tool to the no percentile line control, .


Bar

If you have an error bar chart, this control lets you add a bar to the bottom of the

error bar. The bar control, , changes to the no bar control, , after you click it.


Equal Axes


The equal axes control, , lets you change the dimensions of the axes when displaying a scattergram or line chart of a comparative percentile. The default is to



have the axes be equal. Clicking this control changes the tool to .

Composite and Paging Graphs

No Symbols

The no symbols control, , lets you remove plotting symbols from line charts. Each line will then be differentiated by different line pens. Clicking on this tool

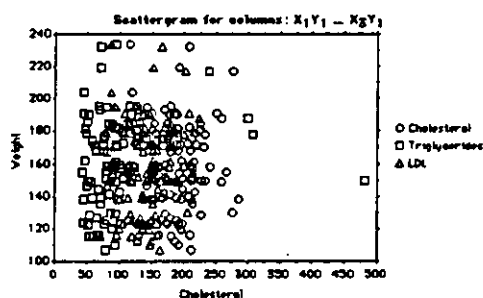
changes it to the add symbols control, .

The composite/paging control is in the upper right corner of the tool palette. When it is in composite mode, the icon looks like . When it is in paging mode, it looks like .


A graph displayed in composite mode has all the variables overlaid on one page. In paging mode, the graph portrays only one variable or X-Y combination per page. The scroll bar on the right side of the screen becomes active; clicking in the scroll bar advances to the next variable or combination. If a single X variable or X-Y pair is being graphed there is no difference between these two modes. If several X variables or X-Y variable pairs are being graphed you can toggle between paging and composite.

Note that toggling from paging to composite mode can change the type of graph you see. If you have a univariate scattergram with Mean, Std. Dev., etc. selected in paging mode, toggling to composite mode will change the graph to error bars. Also note that the error bars view control is only available in composite mode. Even if you have only one X-Y pair, you can put your graph in composite mode to access these view controls.

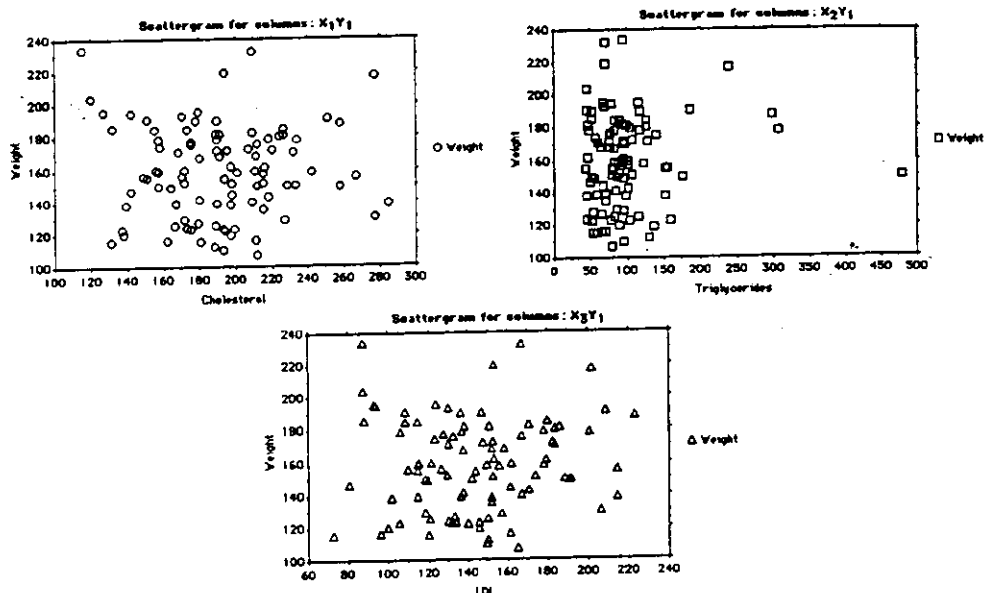
- Open Lipid Data.
- Assign X to Cholesterol, Triglycerides and LDL.
- Assign Y to Weight.
- Choose scattergram from the View menu.
- The composite view graphs the Y variable against each X variables as follows:



Notice that all three X variables are plotted versus the Y variable on a single page.

- Click on the paging tool, , to separate the X,Y pairs into three separate graphs.

- Move between pages by clicking on the vertical scroll bar at the right of the window. The paging views plot the Y variable against each X variable individually as follows:

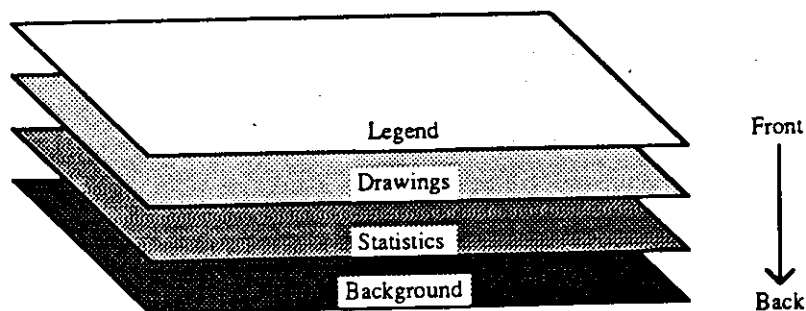


Layout of StatView Graphs

Before you start using StatView's drawing tools, it is useful to see how to modify graphs. The view window shows your graph with any changes you have made in it.

Layers

The window has four *layers* that are manipulated separately:



Anything on a higher layer will have precedence of display. For example, if you add a solid black box to the drawings layer, it will cover up whatever is behind it on the statistics layer. The legend is always seen on top.

The lowest layer, the background, is a single plane of color which you can control.

The next higher layer is the statistics layer. This is the graph that StatView makes. Objects which appear in this layer are the axes and graph contents area. For example, the contents area for a Box Plot would include all the box plots, all the

outliers, and the frame around the graph. Scattergram would include all the points as well as the frame and any regression lines or confidence bands.

The next layer higher is the drawings layer. The elements of this layer are the shapes and text you add with the drawing tools as well as any text created by StatView. Within this layer, elements most recently drawn will cover earlier elements. For example, if you make a small gray box then a large black box in the same location, you will only see the black box. Note that you cannot put shapes behind the statistics or background.

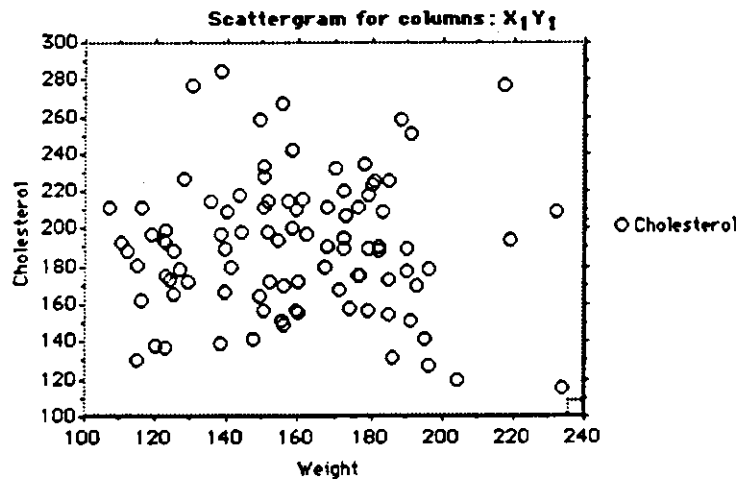
The highest layer is the legend. This means that you can always see your legend, regardless of what you draw in the drawings layer. This is useful since the symbols in the legend are also StatView controls for selecting the shape and color of the points in your drawing.

Resizing

The view window can be easily resized. By dragging the size box in the lower right corner of the window, you can shrink or enlarge the window. Clicking on the zoom box will make the graph the size of the full screen. When the window is the full size, clicking the zoom box again will bring it down to the size it was before you zoomed it up.

When you resize the view window, StatView adjusts the relative positions of objects to fit inside the new window size. Any object you have drawn is kept the same size in the new view window. All other objects are adjusted for the new window size.

To resize a graph, click the mouse inside the graph's frame; a dotted outline of the frame will appear.



Dragging the grow box in the lower right corner of the graph frame changes the size of the frame.

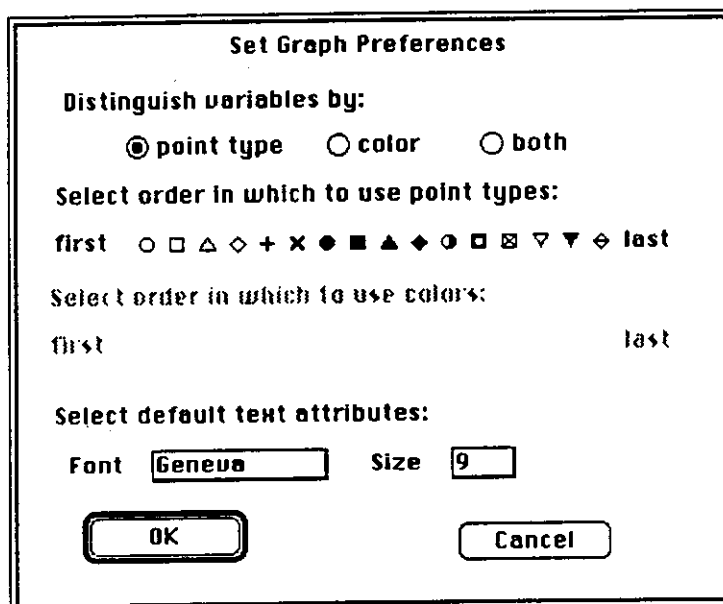
Preferences

The Preferences command in the Graph menu lets you specify many of the graph's general appearances. These settings are saved when you leave StatView and are used each time you run the program. StatView uses these settings when opening a

view window. If you have the view window open when you select these settings it will not be affected.

Note that these settings are saved in the StatView Library file; you should be sure that this file is in the same folder as StatView so that your preferred settings are used each time you run the program.

The command's dialog box looks like:



The dialog box is titled "Set Graph Preferences". It contains several sections: "Distinguish variables by:" with three radio buttons: "point type" (selected), "color", and "both"; "Select order in which to use point types:" with a row of 16 symbols (circle, square, triangle, diamond, plus, cross, solid circle, solid square, solid triangle, solid diamond, open circle, open square, open triangle, open diamond, inverted triangle, inverted square) and "first" and "last" labels; "Select order in which to use colors:" with "first" and "last" labels; "Select default text attributes:" with "Font" set to "Geneva" and "Size" set to "9"; and "OK" and "Cancel" buttons at the bottom.

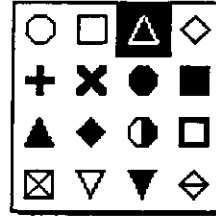
For most graphs, the most important selection in the dialog is the Distinguish Variables By choice. Generally, you will want to:

- select Point Type if you are displaying your results on a non-color system or using a monochrome monitor
- select Color if you are using a color monitor or a monochrome monitor with gray scales

Selecting Point Type causes StatView to use different point shapes for each variable; selecting Color causes StatView to use the same point shape but different colors. Selecting both lets StatView change both point type and color for each variable.

If you are on a non-color system but wish to differentiate your variables by color, they will still appear on your screen as black and white. When you print your graph on a color output device, the objects will appear in color. To find out the color assigned to a point, select the point in the legend and choose the Color command in the Graph menu. Color is described in greater detail in the next section.

One or both of the next two selections are available, depending on what type of system you have. To change the order of points, click and hold down the first point you want to change. For instance, if you selected Point Type and want the third point to be a solid circle, click on the current third element (the hollow triangle). The following menu appears under the pointer:



While still holding the mouse button down, move the pointer to the solid circle and release the button. StatView rearranges the other point types to match your request.

If you are using color, the actions are the same. When you select the color you want to change, the color menu appears. Drag to the color you want and release the mouse button.

The default setting is to use point type to distinguish variables. If you choose to distinguish variables by color only, StatView uses the first point in the point order list as the plotting symbol.

You can also use this dialog box to set the default font and text size of text drawn in the graph view windows. To change either the font or the size, click on the box, and a menu appears under the pointer. Drag to the font or size you want and release the mouse button. If you are printing on a LaserWriter, it is likely that you will want to select a LaserWriter font such as Times or Helvetica.

Colors

StatView operates on systems that support color as well as those that do not. Non-color systems include black-and-white Macintoshes (SE, Plus and Portable) as well as Macintoshes with monitors set to less than 16 colors or grays. Color systems include all Macintoshes (or future machines) with monitors set to 16 or more colors or grays.

If your system is using a monochrome monitor, StatView supports different gray scales. If you are using a non-color system, StatView supports the eight old QuickDraw colors: black, red, green, yellow, blue, magenta, cyan, and white. However, if you are using a color monitor, it is likely that you want to use the Macintosh's color capabilities to their fullest. Even if you don't have color or are running on a non-color Macintosh, you can still take advantage of the eight old QuickDraw colors (and thereby use color output devices).

StatView takes advantage of color in many ways. It can use either 8, 16, or 32 colors. StatView can use any of the 16 million colors in each palette slot. If your monitor can display 16 colors, the StatView color palette has 16 colors; if your monitor can display 256 colors, the StatView color palette has 32 colors. If you use a monochrome monitor, you can have 16 or 32 shades of gray.

To tell StatView which colors you want in the palette, select **Edit Palette** from the **Graph** menu. Click the color you want to modify, and Apple's Color Picker dialog box appears. The Edit Palette command is only available if you are operating on a color system.

To change the color in the Color Picker, drag the pointer around the color wheel. The top portion of the color box on the left shows the new color. The bottom part shows the original color. You can also change colors by clicking on the numerical

values associated with the colors. To add black or white to a color, drag the scroll bar.

When you are finished selecting colors for your palette, click **OK** to save the new colors or **Cancel** to leave the previous selection alone. The **Default** button returns the palette to StatView's default palette.

The colors you select are saved in the StatView Library file. You get the same colors each time you start StatView, even if you changed your palette when running other programs.

You can change the color of almost any object in a graph. Simply select the object, then choose the **Color** command in the **Graph** menu.

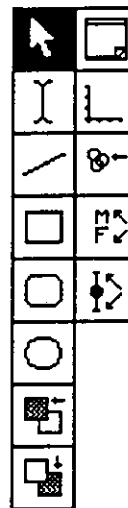
Drawing Tools

So far, you have seen how to create charts and make statistics graphs with StatView. This section shows you how to use the drawing tools to embellish your graphs with text, drawings, and other art.

If you are drawing inside the plotting area of a graph, StatView updates the view window. If you are displaying a large amount of data, this can take a long time. To speed up the drawing process, use row exclusion (described in Chapter 3) to only display a few points. When you are finished embellishing your drawing, include all rows again.

Selecting

The first tool in the drawing tools menu is the selection tool:



You use this tool to select parts of your drawing when you want to modify or change it. Clicking on the selection tool changes the pointer to an arrow. When selecting parts of graphs, you can only select entire groups. For instance, you cannot select only one bar in a bar chart or part of a line in a line chart.

When you select an object, a frame is put around it. For instance, if you select a line of text (such as the title), it becomes framed:

Scattergram for columns: $X_1 Y_1$

To select more than one object:

- Select the first object.
- Hold down the Shift key.
- Select the next object.

You can also select multiple objects by dragging a selection rectangle around them. When you have specified the selection tool and begin a selection somewhere other than on an object, the pointer becomes a finger pointing. Drag that to enclose the objects you want to select. When you let go of the button, all the enclosed objects will be selected.

When you give commands and one or more objects have been selected, the command affects the selected object(s).

Selected objects can be moved by clicking inside the frame and dragging the frame to a new location. When you move a graph, the axes move as well (although the labels do not).

If two objects in the drawing layer overlap, you can move the front one behind the back one with the move to back and move to front tools. The move to back tool is



The move to front tool is



For example, if you draw a box then add some text, you may want to put the text behind the box. Select the text and click the move to back tool.

Resize objects by selecting them and dragging their controls or the grow box in the lower right corner. You can modify the features of a selected item with the various commands in the **Graph** menu.

To select all the objects in a graph window, use the **Select All** command in the **Edit** menu. The **Select Background** command selects just the background so that you can change the color.

Cut, Copy, and Paste

StatView uses the Macintosh standard **Cut**, **Copy**, **Paste**, and **Clear** commands in the **Edit** menu. These act on the object or objects that are selected when you give the commands.

The view window of StatView presents a graph as a group of distinct objects such as text, lines, axes, and the legend; any of these objects may be individually copied out and pasted back to the view window. Also, many of these objects can be cut and cleared from the view window with the **Cut** command. The objects that may not be cut or cleared from the view window are:

- the legend
- axes
- the content area of the graph (such as the plotted points of a scattergram)

If necessary, you can remove the legend by giving the **Hide Legend** command in the **Graph** menu. To remove an axis or the axis lines of the content area, select the

item and make its color the same color as the background. This effectively hides the item.

When any graph objects are copied in the view window, StatView copies a picture that looks exactly like the object, making it static. As a result, if you copy an axis, paste it back in with the Paste command, and then change the original axis' bounds with the Open Axis command, the pasted axis will not match either the graph or the original axis.

You can paste pictures copied from either StatView or any other program into a view window. You can then resize the picture or move it around as you wish. In order to maintain the aspect ratio of pictures pasted into a view window, first select the object, then hold down the shift key and shrink or enlarge the pasted object using its grow box. Double clicking on a pasted picture whose size has been changed will return the picture to its original size. Pasting into the View window will not replace a selected object. The newly pasted object will appear on the top of the graph in the drawing layer.

You have full cut, copy, and paste control over text items and items you draw in the view window. When you cut or copy text with StatView, any style information you may have specified is remembered along with the text. Some applications ignore the style information when text from StatView is pasted into them. This is often true for the subscript and superscript styles.

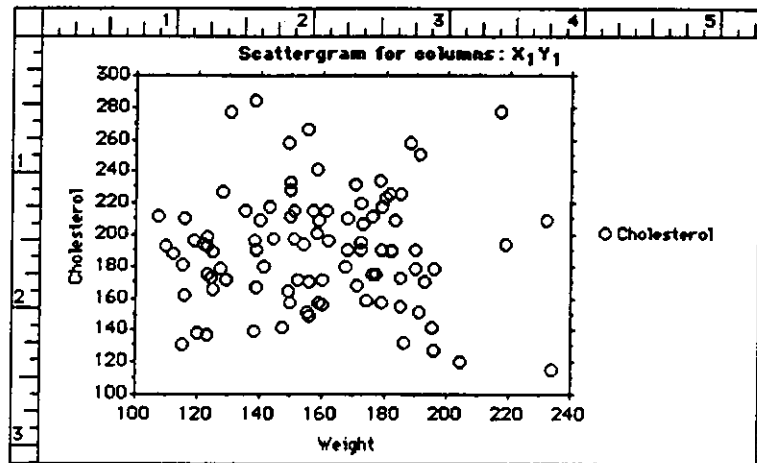
Note: If you want to copy the whole graph (including all the drawing items you have added) to the Clipboard, use the Copy View command in the Edit menu. This command saves the graph in Macintosh PICT format; this allows other object-based programs like MacDraw to manipulate the parts of the view separately. Remember if you use the Copy command, you will only copy out selected objects.

If you are displaying a table view, you are not allowed to cut, paste, or clear items to and from the table view. You are, however, allowed to copy the entire table with the Copy View command. If you want to copy numeric values or table titles of an analysis result, you can select one or more values from the table and copy those numbers out using the Copy command. This is particularly useful for pasting numeric results into spreadsheets, word processors, and data base programs.

Rulers and Grids

StatView's ruler and grid allows you to precisely place any text and drawings you have added to your graph.

The Show Rulers command in the Graph menu turns on the rulers around your drawing:



As you move the pointer around in the graph, the position of the pointer is indicated on the ruler by the moving lines.

For example, if you want to line up the tops of two objects:

- Select the first object.
- Move the pointer to the top of the object and note the position of the indicator line in the ruler in the left margin.
- Select the second object. Move the cursor to the line at the top of the object and drag it to the position you want.
- Note the position of the indicator line on the ruler. When it is the same as the first object, the tops are lined up.

You can change the zero point of your ruler, if you wish. Click in either ruler or in the intersection of the rulers and drag the gray line horizontally or vertically to where you want the zero point to be. You can reset the zero point by clicking in the same rectangle.

You can change the markings on the ruler with the Custom Rulers command in the Graph menu. This command's dialog box is:

Choose Ruler's Units:

☒ Inches ☐ Centimeters

Choose number of divisions per Unit:

☐ 1 ☐ 2 ☐ 4 ☐ 5 ☒ 8 ☐ 10


Select the number of gradations you want and whether you want the rulers to be in centimeters or inches.

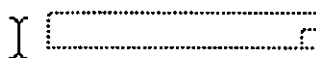
The Turn Grid On command in the Graph menu constrains your movement when you draw or move objects. This forces you to draw objects at intervals of your ruler gradation. It assures that if you place something while the grid is on, it will line up with other objects that are also placed when the grid is on. To stop this action, use the Turn Grid Off command.

Text

If you have objects that you have already drawn and later decide you want them to be aligned with the grid, select them and give the **Align to Grid** command in the **Graph** menu. This can only be done after you have given the **Turn Grid On** command.

It is likely that you will want to add text to some of your graphs. StatView includes sophisticated text editing capabilities including the ability to specify subscripts and superscripts. In the **Text** menu, use the **Style** pull-across menu to add text containing strings such as H_2O and r^2 . The text you add might be something as simple as a short label, or might be a long description of the statistics shown in the graph.

To add text to a graph, select the text tool: . The pointer becomes an I-beam, similar to the one you see in word processing programs. When you click in the graph, StatView draws a box to indicate where the text will go:



Type in whatever text you want. If your text is more than one line, it will justify itself within the size of the enclosing rectangle.

You can easily change the size of the text box. Click on the selection tool (the arrow at the top of the drawing tools), then click and drag the small square in the lower right corner of the text box. This allows you to make the text box as large or small as you want. You can also resize the rectangle with the text tool: when you move over the size box, it turns to the selection tool so that you can drag the size box.

Changing the look of the text is also easy. In the **Text** menu, use the pull-across menus for the **Font**, **Size**, and **Style** commands. You can change the text as a whole or select only part of the text before you change the attributes. Simply select the characters you wish to modify, using the text tool, then choose commands in the **Text** and **Graph** menus.

You can also change how the letters are aligned within their text box with the **Left Justify**, **Center Justify**, and **Right Justify** commands. You can display your text rotated with the **Rotate Left** and **Rotate Right** commands. Note that if you have rotated your text, you must return it to horizontal orientation before editing it. You can change the color of the text by selecting the text box or the color of a character or a group of characters by selecting these characters and choosing the color with the **Color** command in the **Graph** menu.

Note that rotated text with characters having different colors will appear on color printouts in the color of the text's first letter. Multiple lines of rotated text will not print correctly on a LaserWriter.

Although this discussion is about text you add yourself, it also applies to text created by the program. StatView provides a default title for each graph, giving information about the variables in the graph and the graph's type. You may customize this title and change the text in the legend and in the axis labels to fit your needs. Select the text and edit using the text tool or the text menu as explained above.

Drawing Objects

Like the text tool, you will also find the other drawing tools useful for adding information to graphs to make them more understandable or more attractive. The tools are for lines, rectangles, rounded rectangles, and ellipses. To add a line, rectangle, and so on, select the appropriate tool.



For example, to add a box around a group of points:

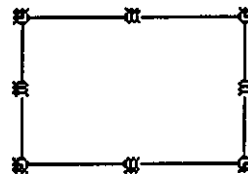
- Select the rectangle tool. The pointer becomes a cross-hair:



- Move the pointer to where you want one corner of the box.
- Click and drag the pointer to where you want the opposite corner.

To draw squares and circles, use the rectangle and ellipse controls. If you hold down the Shift key before you click for the first corner, StatView will restrict your drawing movements to make the object a square or circle. Holding down the Shift key also restricts lines to horizontal, vertical, or to 45° between horizontal and vertical.

When you add an object, StatView shows you its *controls*. The controls are the small squares around the object:



You can resize the object by dragging one of its controls. For instance:

- Select the selection tool.
- Click and drag the lower left control away from the rectangle. When you release the corner, the rectangle grows.

To make an object grow or shrink in both directions by a proportional amount, select the object, hold down the Shift key, then drag one of the corners. StatView constrains the growing and shrinking.

If you want to only stretch or shrink the object in one direction, select a control on the side of the object instead of in the corner. These controls only let you move in one direction.

To move the object, select the middle of the object and drag it around the graph.

Once you have finished drawing an object, the cursor will return to an arrow. By holding down the command key and then clicking in a graphic view window, the arrow cursor will change to the last used drawing tool.

Objects drawn by you or by StatView have attributes that you can change with the following commands in the **Graph** menu:

Command	Description
Color	The color for the object.
Pen	The pattern of the line in the object.
Fill	The pattern of the interior of the object. Selecting "None" makes the object transparent. If you are using color and want a solid color, select black.
Line Width	The thickness of the lines used in the object.
Arrow Head	You can add arrow heads to lines.
Point Type	Changes the selected point to a different shape. The point must be selected in the legend.
Point Size	Changes all points to a different size.

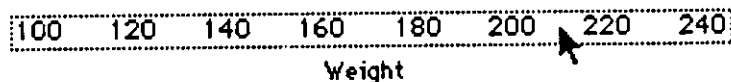
Feel free to experiment with these drawing attributes. Although the standard choices may suit you well, it is likely that you will find use for many of the attributes as you become more proficient with StatView.

Changing the Axes

The axes are often very important in the representations of graphs. It is important that they convey the correct range at the correct intervals to show the meaning of the data. StatView gives you a great deal of flexibility in the way that the axes are displayed.

You can not edit the text in the axes themselves, only in the axis labels. That is, you can not select some of the text labelling points on the axes and change it. However, you can select the axes and apply all of the formatting and styles you have seen above.

To change an axis, select it by clicking in one of the values with the selection tool. StatView draws a rectangle around the axis:



Give the **Open Axis** command in the **Graph** menu. (If you give the **Open Axis** command without selecting an axis, it will let you modify both axes.) The dialog is:

Vertical Axis Information

Bounds:

From: To: ☐ Lock bounds

Length: (inches)

Scale: ☒ linear ☐ log

Ticmarks:

Lie: ☒ outside ☐ inside ☐ both ☐ neither

Per Major Interval: ☒ auto ☐ none ☐ 1 ☐ 3 ☐ 7 ☐ 9

Grid Lines at: ☐ no lines ☒ zero ☐ major tics ☐ minor tics

Option	Description
Bounds	<p>You can set the lower and upper numeric bounds for the axis. When StatView prepares the graph, it chooses the bounds so that the entire range of points will appear. You can, however, stretch or shrink these bounds by entering values into the From and To text boxes.</p> <p>If you specify new bounds, the check box besides Lock bounds will automatically be checked signifying that these values are locked as the bounds for your graph. The bounds will stay in effect for both the paging and composite view of this graph no matter what values are added to or removed from the graph. If Lock bounds is checked and you add a value outside the bounds, it will not appear on the graph.</p> <p>If you wish to unlock your bounds, simply remove the check box from the lock bounds, and StatView will once again automatically set the from and to boundaries.</p> <p>You can also specify how long the axis should be. There are two methods for changing the length of an axis. You can change the number in this dialog box, or, when viewing the graph, select the outline of the graph with the selection tool and stretch it by dragging a control.</p> <p>The axis scale can be linear or logarithmic.</p>
Tick Marks	<p>The tick marks can be shown outside or inside the axis, or both. If you want no tick marks, select Neither. You can specify how many tick marks appear for each major interval.</p>
Grid Lines	<p>StatView normally only puts a grid line at 0. If you want to add other grid lines, specify which you want here. You may also want to turn off the zero line if it appears in the middle of your graph; in this case, select no lines.</p>

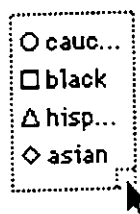
Note that, when you resize a graph, the axes are automatically resized with it.

The Legend

The topmost layer in a graph is the legend. Like the other parts of your graph, the legend is easy to modify. When StatView creates the legend, it assigns it a default width. If the legend text does not fit, the text is followed by ellipses.

☐ caucasian
☐ black
 hispanic
 asian

To change the width of the legend, select it with the selection tool and drag on the bottom corner.



When the legend is selected, you can change the color, font, size, and style of the text items with the various commands in the **Graph** and **Text** menus. Note that, although you can edit legend text individually, you cannot change the attributes (font, style, size, color) of each item in the legend, only of the legend as a whole.

You can switch the orientation of the legend with the **Horizontal Legend** and **Vertical Legend** commands in the **Graph** menu. Making a horizontal legend is convenient if you want to run the legend across the top or bottom of the graph:

☐ caucasian ☐ black hispanic asian

If you don't want the legend to show at all, use the **Hide Legend** command in the **Graph** menu. To bring back a hidden legend, select **Show Legend**.

The legend is the control area which is used to change the points, fills, and patterns in a graph after it is drawn. To change the shape or color of a set of points, the fill or color of a bar chart or pie slice, or the pen or color of a line, select that control in the legend.

☒ caucasian
☐ black
 hispanic
 asian

Then give the appropriate command in the **Graph** menu.

The box plot, error bar, and comparative bar charts do not have legends; instead, each variable is labelled individually on the X axis. To change the attribute of a box or bar, select it in the graph, then give the desired command in the **Graph** menu.

When Customizations Disappear

Your graph customizations can be lost if you change to a different view or select a new statistic. Your choice of point size will remain in effect throughout all changes to the view window.

The following actions WILL cause your graphic customizations to be lost:

- Changing from one statistic to another. This includes switching between simple, polynomial, and multiple regression, and changing any parameters of a factor analysis or a stepwise regression.
- Changing from one view type to a different view type (except for those noted below).
- Removing a variable which appears on a paging view. This will clear any customizations associated with that page of the view.

The following actions WILL NOT cause your graphic customizations to be lost:

- Causing a statistic recalculation or graph redrawing by including/excluding rows, adding or editing a range restriction, or changing a value in a data column.
- Switching to and from a table view.
- Changing between scattergram view and line chart view.
- Toggling between paging and composite modes.
- Changing a statistic parameter. This includes changing parameters of statistics in the **Describe** menu or changing parameters of a simple, polynomial or multiple regression.
- Removing a variable which appears on a *composite* view.

StatView updates the default graph titles and axis labels automatically when you change any variables in your graph unless you have changed the text in the title or label. The only exception to this occurs when the graph title is an equation for a regression. These equations will always update if the regression is recalculated.

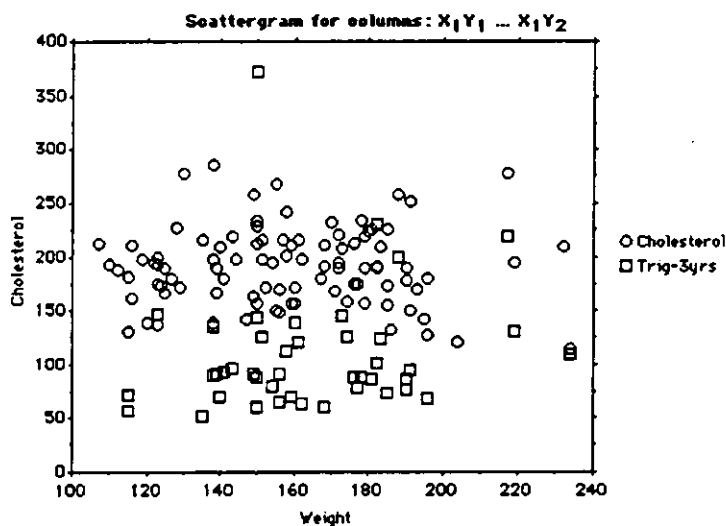
To avoid losing your customizations, be sure you have fully experimented with your data before you begin customizing your graph.

Customizations to scattergram views of multiple regressions and factor analysis factor plots will disappear when you turn to a new page.

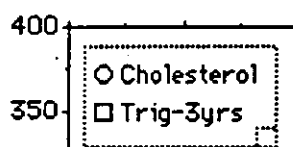
StatView's Drawing Features

This section gives you an extended example of how StatView's drawing features are used. The first example customizes a scattergram.

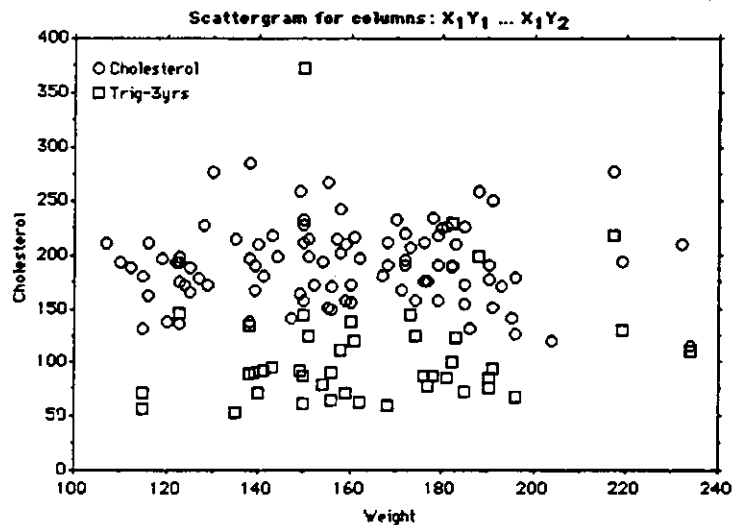
- Open Lipid Data.
- Select **Quick Assignment** in the Variables menu. Assign X to Weight, assign Y to Cholesterol and Trig-3yrs. Click **Done**.
- Select **Scattergram** in the View menu and zoom the window to full size.



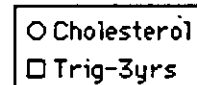
- Select the legend and drag it until it is the upper left hand corner of the graph frame.



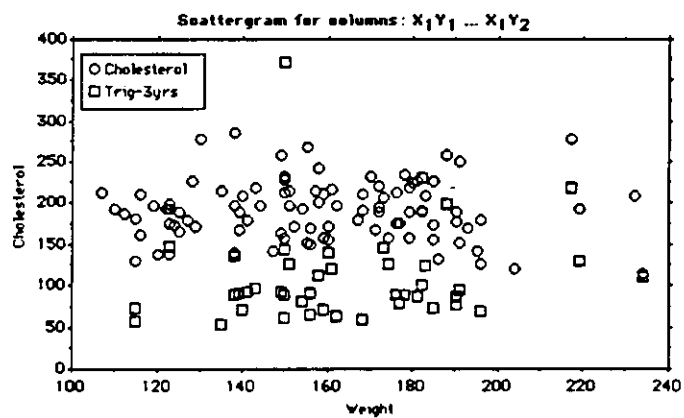
- Select the graph frame and drag the grow box until the frame rests near the right edge of the window.



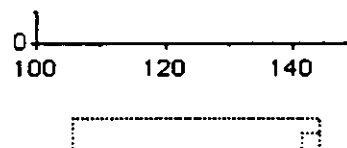
- Draw a box around the legend.



- Select the graph frame and move bottom of the graph up. Move the X and Y axis labels so they are placed evenly with the new frame size:



- Select the text tool. Click under the bottom of the graph:



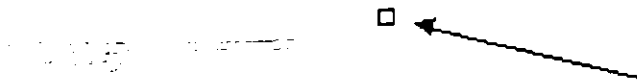
- Type "Illustration 1: Lipid measurements for medical students" and resize the text box to show all the text.

Illustration 1: Lipid measurements for medical students

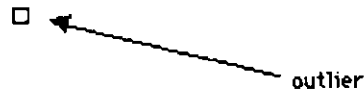
- Select "Illustration 1" and from the text menu change its font size to 10 pt and its font style to bold.

Illustration 1 : Lipid measurements for medical students

- Select the line tool.
- Click the line tool near the top extreme Trig-3yrs value and drag to the right.
- Select **Arrow Heads** in the **Graph** menu. Drag to the single head on the left:

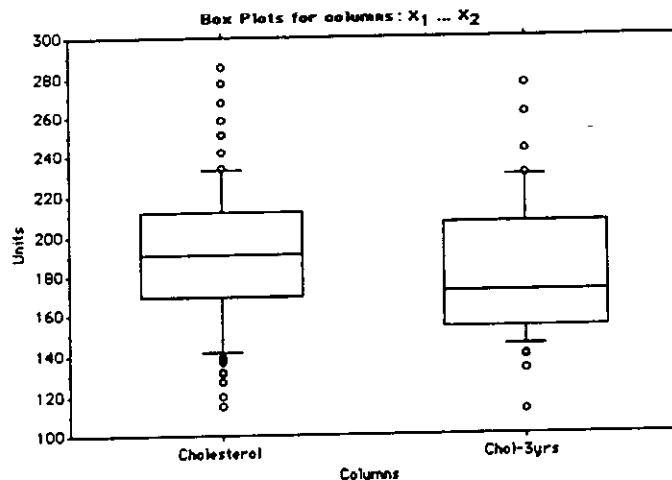


- Select the text tool. Click under the bottom of the arrow.
- Type "outlier".

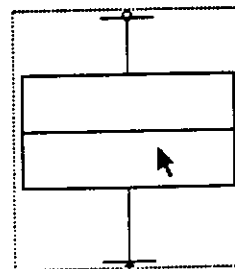


The next example customizes a box plot.

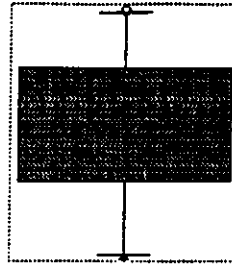
- Select **Quick Assignment** in the **Variables** menu. Clear any X or Y variables. Assign X to Cholesterol and Chol-3yrs.
- Select **Percentiles** in the **Describe** menu.
- Select **Box Plot** in the **View** menu and zoom the window to full size.



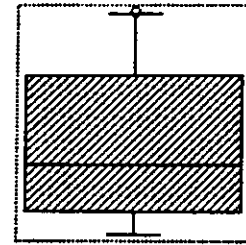
- Select the Cholesterol box plot by clicking in its middle. A frame will appear around the box.



- Select **Fill Pattern** in the **Graph** menu. Drag to the medium gray and release the pointer. This fills the selected box plot with the chosen pattern.
- Repeat this procedure for the Chol-3yrs box. This time fill the box plot with the diagonal black line fill pattern.



Cholesterol



Cholesterol-3yrs.

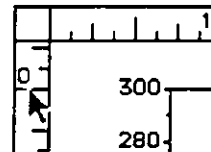
- Select the vertical axis label "Units".
- Select **Rotate Right** in the Text menu. You will now be able to edit this text.
- Select the text tool. Double click in the text field to highlight the entire word.

Units

- Type "Cholesterol (mg/dl)" and resize the text box until all the text is visible.
- Select **Rotate Left** in the Text menu.

Cholesterol (mg/dl)

- Select **Show Rulers** in the Graph menu.
- Click in the top left corner of the ruler and drag down the gray line until it lines up with the top of the graph. The zero point is now the top of the Y axis.



- Select the vertical axis label and drag it until it is centered.

Cholesterol (mg/dl)

Explore the other drawing features available on the Text and Graph menus. If you have a color system, experiment with assigning colors to various objects.

Chapter 5 — The Describe Menu

StatView computes descriptive statistics for any dataset column assigned as an X variable. The descriptive statistics calculated are:

- mean
- standard deviation
- standard error of the mean
- variance
- coefficient of variation
- count
- minimum
- maximum
- range
- sum
- sum of squares
- number of missing values
- t or normal distribution confidence intervals
- number of values below 10th percentile
- 10th percentile
- 25th percentile
- 50th percentile (or median)
- 75th percentile
- 90th percentile
- number of values above 90th percentile
- mode
- geometric mean
- harmonic mean
- coefficient of kurtosis
- coefficient of skewness
- frequency distribution

The descriptive statistics fall into two categories: those above the gray line on the Describe menu and the one (Frequency Distribution) below the gray line. Any combination of the descriptive statistics above the gray line can be selected.

The selected statistics are checked on the menu. Deselect a statistic by choosing it again. When all statistics are deselected, None is checked. Selecting None will deselect all Describe statistics.

Numeric Descriptive Statistics For A Single Variable

If you select **Frequency Distribution**, any checked statistics above the gray line become unchecked. Remove **Frequency Distribution** by selecting **None** or any statistic above the line.

This example deals with the non-graphic descriptive statistics that you might wish to compute for a single variable (that is, a single StatView data column). The dataset we are examining is again **Lipid Data**, which is included on the StatView diskette.

- Open **Lipid Data**.
- Assign **X** to the column containing the variable you wish to describe. For our dataset, we are interested in describing the variable **Cholesterol**. Double click the cursor on the column heading **Cholesterol** to assign it as **X₁**.
- Select **Mean, Std. Dev., etc.**
- Select **Mode**.
- Select **Percentiles**. Note that the median is the 50th percentile; use this command when you want to investigate the median.
- Select **Geometric Mean**.
- Select **Harmonic Mean**.
- Select **Kurtosis & Skewness**.

As you become experienced with StatView you may not wish to make all these selections. Indeed, many people will find the first three selections sufficient.

The view window appears after the first statistic is selected and then is updated as each new statistic is chosen. The results of the checked statistics for the **Cholesterol X₁** variable are displayed in tabular form:

X ₁ : Cholesterol					
Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
191.232	35.674	3.66	1272.648	18.655	95
Minimum:	Maximum:	Range:	Sum:	Sum Squared:	# Missing:
115	285	170	18167	3593733	0
* < 10th %:	10th %:	25th %:	50th %:	75th %:	90th %:
8	142	168.5	191	212	233
* > 90th %:	Mode:	Geo. Mean:	Har. Mean:	Kurtosis:	Skewness:
9	190	187.925	184.582	.036	.302

The first two rows of this table represent the statistics associated with the **Mean, Std. Dev., etc.** selection. The third row and the first entry of the fourth row represent the percentiles selection. Each of the last five entries of the final row represent the other descriptive statistics selections and are labeled accordingly. By following the statistics as reported here, you can generate many descriptive statistics.

Measures of Central Tendency and Skewness

If the distribution of Cholesterol values is symmetric, but not necessarily normal, then the *50th percentile* (also referred to as the *median*), *mean*, and *mode* will be identical. For these data, the mode is 190. By definition, the mode is the most frequently occurring value in a distribution. The mode is the least valuable measure of central tendency, only telling us the most frequently occurring observed value. If there is more than one most frequently occurring observed value, there is no mode.

The mean and median are better measures of central tendency. The median is 191 for our Cholesterol variable and is approximately .232 units below the mean of 191.232, suggesting that the distribution is slightly *skewed*. More than 50 percent of the values are below the mean value of 191.232, the median is less than the mean, thereby suggesting a positively skewed distribution. Thus, the right tail of this distribution is longer than the left tail. This suggests that the high Cholesterol values tend to deviate more from the mean than the low Cholesterol values.

A similar conclusion could have been reached by referring to the index of skewness, .302. This index is the average of the cubed standard scores (or *z-values*) of the distribution. If the average of the cubed standard values is 0, the distribution is *symmetric*, suggesting that the extreme values are evenly distributed above and below the mean. If it is negative, the distribution is negatively skewed suggesting that the majority of extreme observed values are less than the mean. If it is positive, the distribution is positively skewed, suggesting that the majority of extreme values are greater than the mean.

Select the minimum value in the distribution of Cholesterol, 115, and change it to -1000. (Do not save this change when you close Lipid Data.) Notice when you view the table again that the median has not changed, while the mean has been reduced substantially. This demonstrates the stability of the median and the instability of the mean as measures of central tendency. The median is the single best descriptive measure of the central tendency of a distribution.

Variance and Standard Deviation

The variance is a measure of the dispersion or the extent to which there are individual differences in the distribution of values. For the Cholesterol data, the variance is 1272.648, indicating that, on the average, the squared difference between any value and the mean is 1272.648. StatView uses the variance formula for an unbiased sample estimate rather than the formula for a population. As a consequence, when comparing a StatView variance estimate to a variance estimate obtained by some other procedure, you might occasionally find that the StatView estimate is a small amount larger if the alternative method has used a formula for a population variance. The differences between the two formulae are small and only worth noting so you can be assured that any comparative discrepancies that you might observe are due to a subtle difference between formulae.

The standard deviation is simply the square root of the variance. For Cholesterol, it is 35.674. Depending upon discipline and preferences, you may want to report either a variance or a standard deviation, but usually not both.

Maximum, Minimum and Range

The maximum, minimum, and range are especially valuable in that they help you check the data quickly. The minimum is just the smallest value in the dataset, 115

for our Cholesterol variable, while the maximum is the largest value in the dataset, 285. You generally have an idea of what the span of values should be in a given distribution. Should either the minimum or maximum value be substantially lower or higher than expected then you have probably entered a value incorrectly into the dataset.

While it is very difficult to directly interpret a standard deviation as being either large or small it is possible to compare the range to the standard deviation. This comparison will provide some sense of whether a distribution is homogeneous (small variance), or heterogeneous (large variance). The ratio of the range to the standard deviation should typically define some value between 2 and 6. For our Cholesterol data, with a range of 170 and a standard deviation of 35.674, this ratio is approximately 4.76. Because 4.76 is near the center of the traditional range of 2 to 6, our data are neither heterogeneous nor homogeneous. Had our data defined a ratio near 2, or even less than 2, we would have concluded that our Cholesterol sample was extremely homogeneous. By similar logic had our data defined a ratio near 6, or even greater than 6, we would have concluded that our sample was extremely heterogeneous.

Kurtosis

When describing a distribution, four measures are typically provided: the mean, variance, skewness, and *kurtosis*. We have briefly discussed the first three. Kurtosis refers to both the peak, the center of the distribution, and tails of a distribution. The kurtosis of a distribution is computed as the average fourth power of the standard scores minus 3.

A distribution may be characterized as *leptokurtic*, extremely peaked with "slim" tails; *platykurtic*, extremely flat with fat tails; or *mesokurtic*, modestly peaked with modest tails. For a normal distribution the average fourth power of the standard scores is exactly 3. Since by convention 3 is subtracted from the average fourth power, the kurtosis of a normal or mesokurtic curve is 0. A leptokurtic distribution would be characterized by a positive index of kurtosis, a platykurtic distribution would be characterized by a negative index of kurtosis. The larger the absolute value of the index of kurtosis, the more extreme the kurtosis.

For our Cholesterol distribution the kurtosis is .036, indicating that the distribution is mesokurtic, neither peaked nor flat. If it were platykurtic, its fat tails would indicate that there are more extreme values in the distribution than you would find in a normal distribution, whereas the slim tails of a leptokurtic distribution would suggest that there are fewer extreme values in the distribution than you would find in a normal distribution.

How fat can the tails get? It is possible to define a concave distribution, U shaped, where virtually all of the values are in the tails of the distribution. Such a concave, extremely platykurtic distribution could have a kurtosis index that might exceed -2. By the same logic, you might wish to define a peaked distribution with very long, low tails. Such a distribution would be extremely leptokurtic with a kurtosis index that might be as large as 2.

Coefficient of Variation

One of the more intriguing topics in basic statistics deals with the scale of the variable being analyzed. An in-depth discussion of this topic is beyond the scope of this manual. However, it is necessary to briefly touch on it in order to appreciate the coefficient of variation, the harmonic mean and the geometric mean. Variables

usually assume the characteristics of one of four scales: *nominal*, *ordinal*, *interval* or *ratio*. When Karl Pearson was refining much of what we call "descriptive statistics," he was working with ratio variables. Such variables have no negative values, a zero that implies an absolute lack of whatever is being measured, and an equal distance between any two consecutive values.

An example of such a ratio variable is limb length. It is well documented that the variation of the length of a front limb is related to the average length of the front limb. For instance the average length of the front limb of a human is substantially greater than the average length of the front limb of a hamster. We find that the standard deviation of the length of human arms is also substantially greater than the standard deviation of the length of the front legs of hamsters. You could not compare the standard deviations of the front limbs of the two species and conclude that there is greater front limb variation in humans than there is in hamsters.

In situations where two populations differ appreciably in their means, assuming a ratio variable, the coefficient of variation should be used to compare variation. It is computed as the ratio of the standard deviation to the mean multiplied by 100. This coefficient is independent of the unit of measurement and will usually range from some value greater than 0 to 100. It is possible with some extremely heterogeneous data to have coefficients greater than 100.

Our Cholesterol variable is a ratio variable. An observed Cholesterol value of 0 means that there is absolutely no trace of cholesterol present in whatever is being sampled. In the human population Cholesterol values typically range from 118 to 300. The coefficient of variation for our Cholesterol is 18.655.

A pressing question that should be addressed deals with the magnitude of the coefficient of variation. How big is big? Simpson, Roe and Lewontin (1960), a reference prior to the widespread use of computers, reports that a majority of values in the biological sciences are within the range of 4 to 10, with 5 or 6 being a reasonable estimate of the average value. All values reported by Pearson (1898) in his original work were less than 5. In the social sciences informal studies of non-ratio variables have suggested that the coefficient of variation typically ranges between 15 and 30, with small samples of fewer than 30 observations tending to define larger coefficients than large samples. Clearly the biological sciences work with more homogeneous variables than the social sciences.

A relatively small coefficient of variation, indicating extreme homogeneity, would suggest that the associated variable is not sufficiently sensitive to represent variation within the sample being measured. Alternatively, a relatively large coefficient of variation might suggest that an instrument is not measuring a single dimension.

If you are using a variable for which an observed value of 0 implies something other than the absolute lack of whatever the associated variable is measuring, then you are not using a ratio variable and the coefficient of variation could be seriously distorted, perhaps corrupted. It is best to ignore the coefficient of variation unless you are working with ratio variables. Such variables are typically found in the biological and physical sciences, but seldom found in the social sciences.

Percentiles

The choice of percentile does not determine the observed values associated with the 99 percentile points that exist in any distribution. Instead, StatView identifies the observed values associated with five critical percentiles: the 25th, 50th, 75th, 10th, and 90th. The first three of these are referred to as the first, second and third

quartile points, respectively; the second quartile is also called the median. The 10th and 90th percentiles are usually the “rule of thumb” points that segregate the extreme portions of the distribution from the rest of the distribution. Note that the median is the 50th percentile; to investigate the median, show the percentiles.

To the extent that the difference between the values associated with the 25th and the 50th percentiles equals the difference between the values associated with the 50th and 75th percentiles, the middle 50 percent of the associated distribution will tend to be symmetric. In addition, if the difference between values associated with the 10th and 50th percentiles is also equivalent to the difference between the values associated with the 50th and 90th percentile points, then the middle 80 percent of the distribution will be symmetric.

Standard Error of the Mean

The Std. Error item in the Describe menu stands for the Standard Error of the Mean. If we were to assume that our data came from a population, which we do when we compute our standard deviation, we might wonder about the adequacy of the mean in terms of representing the population mean. If we were to assume that our sample of Cholesterol values was randomly selected from a population of Cholesterol values we could make certain statements about what to expect if we were to take subsequent samples of 95 Cholesterol values from the population. If we were to compute the mean for each sample of 95 Cholesterol values we would eventually have a sampling distribution of means.

The mean of this sampling distribution would be the population mean and the standard deviation of this sampling distribution would be equivalent to the population standard deviation, estimated to be 35.674 from our data, divided by the square root of the sample size, or the count, used. The estimated standard deviation of the sampling distribution of means is approximately $(35.674/9.74)$ or 3.66, which is also referred to as the standard error of the mean. We would expect that approximately 68% of all sample means would be within 3.66 Cholesterol units of the population mean. Thus, for our Cholesterol data it would appear as though there would not be too much fluctuation in the mean from sample to sample. In a sense, the standard error of the mean is an indication of just how accurately the population mean can be portrayed by the sample mean. To get a better understanding of the use of the standard error of the mean as a descriptor, you should refer to a basic statistics book that discusses confidence intervals.

Geometric Mean

The geometric mean is not a general descriptor of the values of a distribution. It is commonly used to average economic indices. It is most useful with a variable undergoing a constant rate of change. It represents a mean that would be defined by the data if they were transformed in a specific manner. Recall, from the section on the coefficient of variation, that there are situations when the means and variances associated with a ratio variable tend to be systematically related. In such situations, as the mean increases, so does the variance. Frequently the variance can be made independent of the mean by representing each value of the distribution by its common logarithm. When you compute the mean of a logarithmic distribution and then transform that mean back to the metric of the original untransformed distribution, the result is a geometric mean. Note that if you have a value of 0 in the distribution, the geometric mean can not be computed. For our Cholesterol data, the geometric mean is 187.925. The geometric mean is always less than the arithmetic mean when all values of the distribution are positive.

Graphic Description of Data For A Single Variable

Harmonic Mean

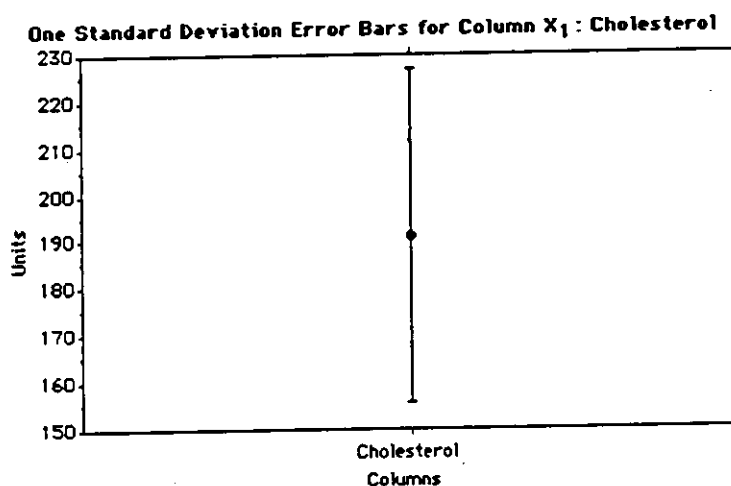
The *harmonic mean*, also derived from transformed data, is most often used to average rates and ratios. It is based on a reciprocal transformation, which replaces a value by its reciprocal (where the reciprocal is defined as 1 divided by the value). The reciprocal of the mean of a distribution of reciprocals is referred to as a harmonic mean. Like the geometric mean, it is not really descriptive of the observed distribution. Instead, it may be thought of as being descriptive of a transformed distribution. Note that if you have a value of 0 in the distribution, the harmonic mean cannot be computed. For our Cholesterol data the harmonic mean is 184.582. The harmonic mean is always less than the geometric and arithmetic means.

When you use the graphics window to view your data, there are several options available with the Describe menu.

- Open Lipid Data and assign X to Cholesterol.
- Select Mean, Std. Dev., etc. from the Describe menu.
- Select Scattergram from the View menu.

Standard Deviation Error Bars

This scattergram is a univariate rather than a bivariate scattergram. It has only an ordinate, as there are no values for the abscissa. (Throughout this chapter, ordinate is the Y axis and the abscissa is the X axis.) A standard deviation error bar appears:

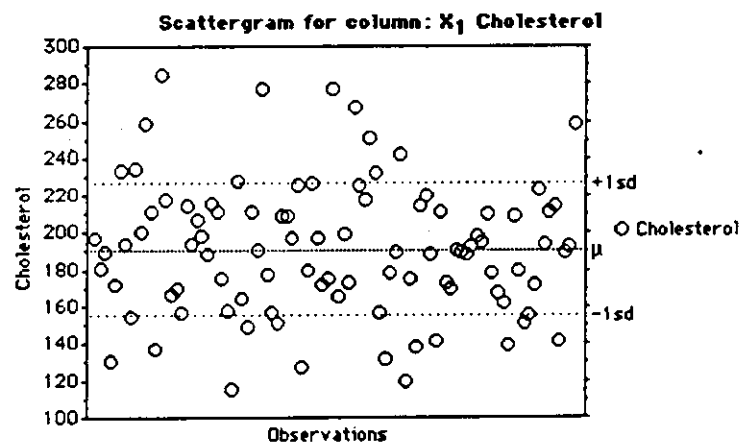


This standard deviation error bar for Cholesterol has a dot at the the mean, 191.232. A line representing one standard deviation unit, 35.674, is extended above and below the mean. This line is intended to convey some sense of the degree of dispersion about the mean. Usually a majority of the observed values fall within a band that extends from one standard deviation unit above the mean to one standard deviation unit below the mean.


Univariate Scattergrams

- Click the composite/paging tool, the top right control of the tool palette.

The graph will be redrawn with all the observed values entered. Notice in the following graph that standard deviation error band and mean are noted by lines and marked on the right axis. Also note that this graph represents each observed value with a circle. The vertical placement of the observed values is with regard to the axis, Cholesterol. The circles are placed horizontally in such a fashion that we can count the number of individuals with a particular observed value. For instance four individuals have Cholesterol counts of 191. A quick count shows that 67 of the 95 observed values, or 70%, are within plus or minus one standard deviation of the mean. If these data were normally distributed we would expect approximately 65 of the 95 observed values, or 68%, to be within plus or minus one standard deviation of the mean.



The default is to show one standard deviation above and below the mean. The tool palette for this view contains a tool unique to this display, the standard deviation

control, . This control allows you to specify the width (in standard deviations) for the band displayed by the standard deviation lines.

Confidence Intervals

Suppose you wished to estimate the population mean from the sample mean. We always assume that there has been some sampling error with any sample mean. Thus, we can never be sure that the sample mean and the population mean are exactly the same. We can, however, estimate a band of values which we might confidently predict as spanning the population mean. Such a span is a confidence interval. You can also specify the probability or degree of confidence that you wish to associate with the confidence interval. Such a value represents the probability that the band will span the population mean with repeated applications. That is, we never know if a particular confidence interval spans the population mean, but we do know that if we select a probability of .95, then 95 percent of the time that we construct such a confidence interval it will span the population mean.

- Open Lipid Data and assign X to Cholesterol.
- Select Confidence Intervals from the Describe menu, and the Confidence Intervals dialog box appears:

Select Distribution:

☒ t ☐ normal Std. Dev.:

Select Confidence Intervals:

☒ 95% ☐ 90% ☐

You can select from two distributions available for computing confidence intervals: the t-distribution and the normal distribution. Choose the normal distribution only if you are constructing confidence intervals for data in which you know the true standard deviation of the population. If you select normal distribution, enter the population's standard deviation in the text entry rectangle to the right of its radio button. When you don't know the population standard deviation, the sample standard deviation is used in conjunction with the family of t-distributions to construct a confidence interval. Confidence intervals based on sample standard deviations are always wider than confidence intervals based on population standard deviations.

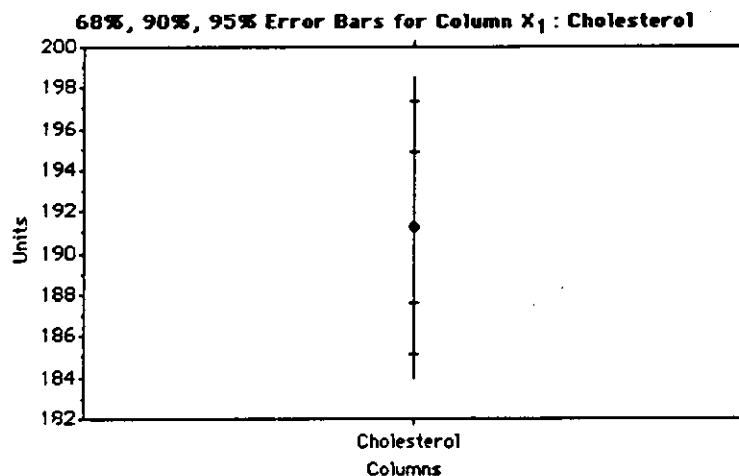
Regardless of whether you have selected a t- or normal distribution, you may have up to three probability levels associated with your confidence interval. Two of the probability levels have been preselected; they are 95% and 90%. The text entry rectangle next to the last check box allows you to enter your own choice for the probability to be used for the third interval. If you wish to display an interval based upon a single standard error bar, enter 68%.

- Click t for the distribution.
- Select the 95% box, the 90% box, and enter 68 in the confidence interval text entry rectangle. (Entering values into this rectangle automatically checks the associated box.)
- Click OK. A table appears which lists the values defining each confidence interval.

X ₁ : Cholesterol					
t 95%:	95% Lower:	95% Upper:	t 90%:	90% Lower:	90% Upper:
7.267	183.964	198.499	6.08	185.151	197.312
t 68%:	68% Lower:	68% Upper:			
3.659	187.572	194.891			

The band for the 95% confidence band extends from a low of 183.964 to 198.499; the 90% confidence band extends from a low of 185.151 to a high of 197.312; the 68% confidence band extends from a low of 187.572 to a high of 194.891. The values under the t's in table above are the t-values from a t table associated with the 95%, 90%, and 68% confidence levels respectively for 94 degrees of freedom.

- Select Scattergram from the View menu.



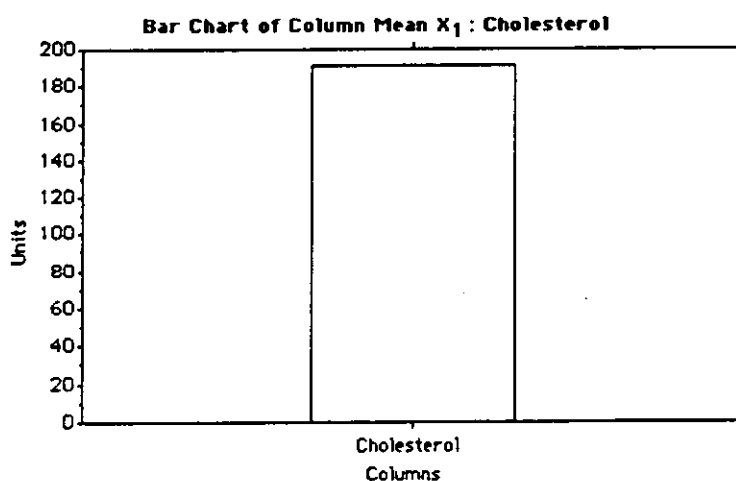
The two interior tick marks are associated with the 68% confidence interval and span from 187.572 to 194.891. The second set of exterior tick marks are associated with the 90% confidence interval and span from 185.151 to 197.312. The end of the lines represent the demarcations for the 95% confidence interval and span from 183.964 to 198.499. Thus, the probability of spanning the population mean within the interval is .95. As you become more confident that a confidence interval spans the population mean, the span associated with the confidence interval widens.

z-Score Distribution

A z-score distribution is a quick method for displaying a standard score frequency distribution. It is possible to graph the distribution of standard scores associated with any variable within StatView. A standard score is an observed value's deviation from the mean, expressed in standard deviation units.

Assuming that Cholesterol is still X₁ and that you entered the information for the confidence intervals as outlined in the previous section, it is possible to obtain a graphic representation of Cholesterol in standard score form (also referred to as z-scores). Note, that the standard score information is also available if **Mean**, **Std.dev.**, etc is selected.

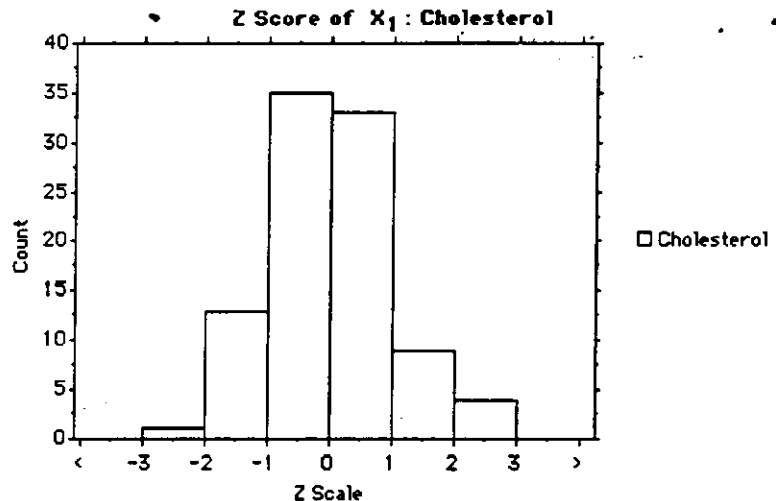
- Select **Bar Chart** from the **View** menu. You should see a chart with a single bar in the center:



This chart is, unfortunately, not too informative. It simply represents a bar whose height on the axis is comparable to the mean of variable X_1 .

- Click the composite/paging tool.

The graph changes to a histogram representing the z-score frequency distribution of Cholesterol. The abscissa, or baseline, associated with the z-scores usually ranges from +3 to -3. The bars are drawn one z-score unit wide while the ordinate represents the frequency of z-scores associated with each bar.

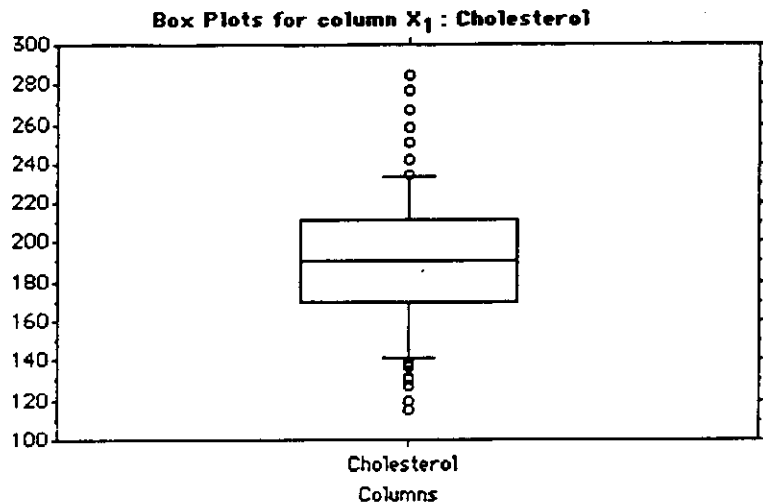


A z-score of 0 is always associated with the mean. For this illustration of our Cholesterol data, the tails of the distribution extend to -3 and +3. Frequently it is possible to see the skew of the distribution when looking at this graph. For our Cholesterol data there is only a slight positive skew, which is difficult to perceive in this graph.

Box Plot

Tukey originally developed the box plot. Note that our box plots are based upon the work of William Cleveland and differ from Tukey's box and whisker plots in the manner in which outliers are plotted. In the numerical description of Cholesterol, we computed five percentile ranks (10, 25, 50, 75, and 90). We briefly discussed these five percentile ranks and the information that they provided. The box and whisker plot is a graphic method for displaying these five percentile points.

- Open Lipid Data and assign X to Cholesterol.
- Select Percentiles from the Describe menu.
- Select Box Plot from the View menu.



This box plot is derived from the five percentiles. The top of the box represents the 75th percentile, a Cholesterol value of 212. The bottom of the box represents the 25th percentile, a Cholesterol value of 168.5. The middle 50% of the Cholesterol values are contained within the span defined by the box boundaries. The line in the middle of the box represents the median for Cholesterol, 191. If the distribution is symmetric the median will be in the exact middle of the box. The lines extending above and below the box are referred to as "whiskers". The top whisker is drawn from the Cholesterol value associated with the 75th percentile, 212, to the Cholesterol value associated with the 90th percentile, 233. The bottom whisker is drawn from the Cholesterol value associated with the 25th percentile, 168.5, to the Cholesterol value associated with the 10th percentile, 142.

The small, somewhat overlapped circles under the lower whisker and seven small circles above the upper whisker represent observed values below and above the 10th and 90th percentile values, respectively. Extreme values are clearly apparent with box plots, as are outliers. StatView always shows the highest 10% of observed values and the lowest 10% of observed values. When an observed value is an outlier, it will appear to be farther removed from the whisker than the other extreme values. For our Cholesterol data, the box plots make it clear that the high Cholesterol values are more extreme than the low Cholesterol values.

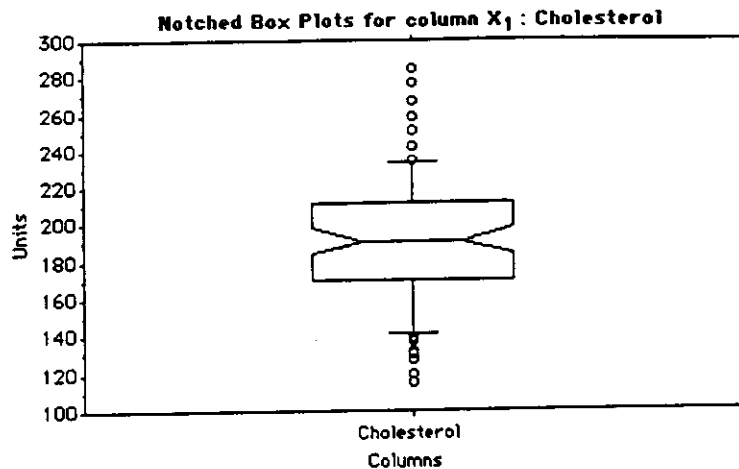
Two controls appear on this view which are unique to box plots. The notch control,




changes the box plot to a notched box plot. The notches represent 95% confidence bands about the median.

- Click the notch. A notched box plot appears:

The notches are quite narrow on this plot. The confidence band for the median ranges from approximately 190 to 200. Our previous discussion regarding confidence bands may also be applied here with the exception that we are dealing with a median rather than a mean.



The second control, , eliminates the representation of the the extreme twenty percent of the observed values, the 10% below the 10th percentile and the 10% above the 90th percentile:


The Cumulative Frequency Curve

It is sometimes informative to plot observed values against their percentiles in a percentiles plot. The percentiles plot displays a cumulative frequency curve.

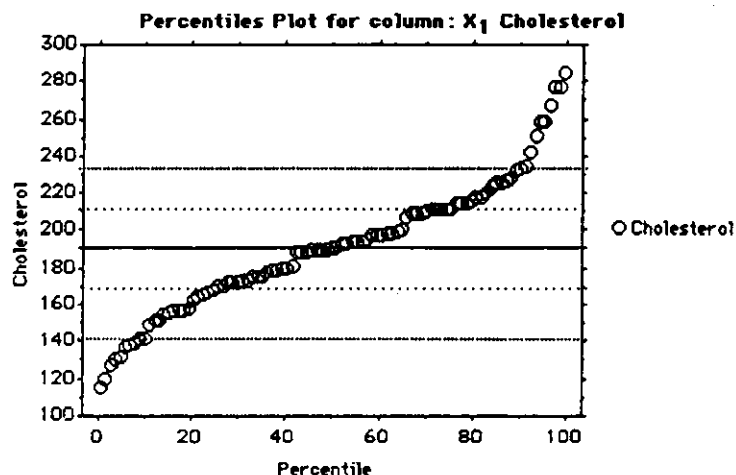
To obtain a percentile plot you must have Percentiles selected in the Describe menu.

- Open Lipid Data and assign X to Cholesterol.
- Select Percentiles from the Describe menu.
- Select Scattergram from the View menu

The cumulative frequency curve allows you to very quickly estimate the percentile associated with any observed value in a distribution. The ordinate represents observed values, cholesterol, and the abscissa represents the values from 0 to 100. These abscissa values represent the cumulative percent below the upper limit of an observed value. The percentile plot will rise from lower left to upper right, regardless of the observed values used.

- Click the bottom control of the view control panel, , which is unique to percentile plots.

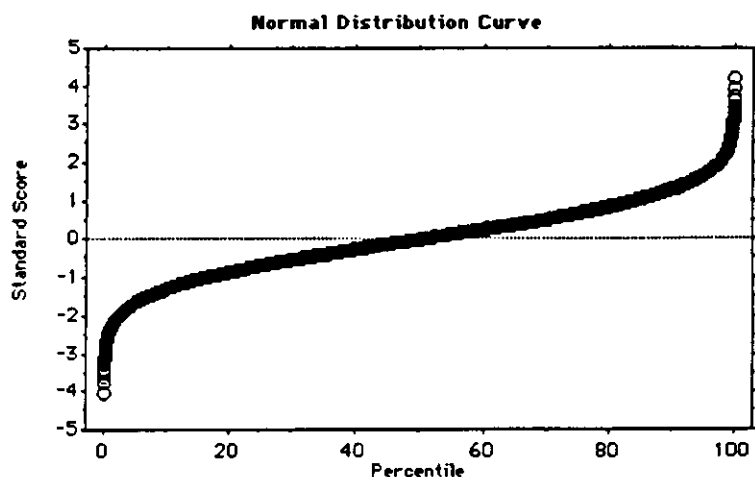
(If this tool is not visible, click the composite/paging tool.) This will cause horizontal lines to be displayed on the plot, representing the five percentile values (10th, 25th, 50th, 75th, and 90th percentiles) indicated in the table view and used with the box and whisker plot.



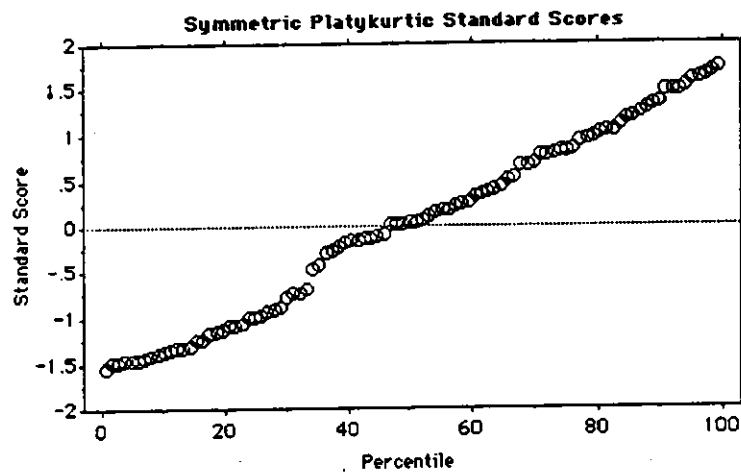
To estimate the percentile associated with an observed Cholesterol count of 200, we simply scan across to the points associated with 200, then scan down to the percentile – approximately the 58th percentile in this case. We conclude that approximately 58 percent of the sample have Cholesterol levels below 200.

Notice that this percentile plot associated with our mesokurtic Cholesterol distribution has a minor middle “hump”. The plot is a slowly rising line except at the upper right. This slow rise suggests that the associated distribution is essentially mesokurtic. Furthermore, the abrupt increase in the percentile plot slope in the upper right suggests that a very small change in percentile is associated with a rather large change in observed value. Some of the high Cholesterol values are substantially higher than the rest of the Cholesterol values. If there was an abrupt upward change at the lower end of the plot it would suggest some low outlier; a small change in percentile value at the low end of the Cholesterol continuum associated with a large change in observed Cholesterol value.

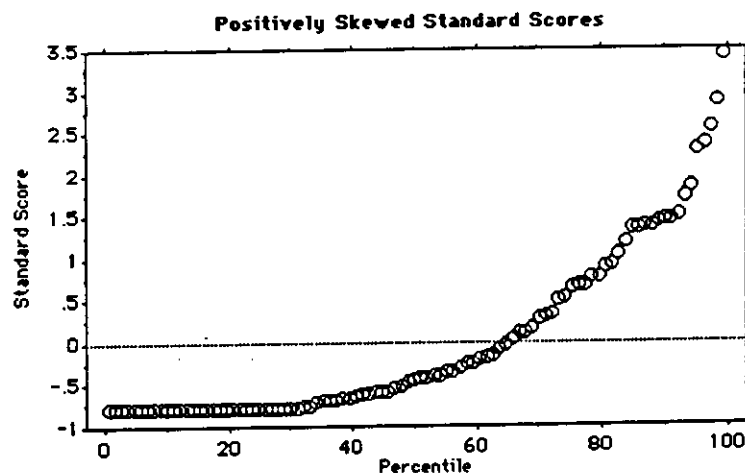
There are certain properties implicit in this view. We know that, by definition, half of the observed values will be to the right of the 50th percentile and half will be to the left of the 50th percentile. Furthermore, we know that the graphed points will go from the lower left to the upper right. We also know that if a distribution is normal, the relative frequency curve will be an S-shaped, *sigmoid*, curve. Deviation from the S-shape allows us to make certain inferences about the shape of the associated distribution. To aid in an interpretation of the percentile plots, we plot cumulative frequency curves of four distributions: normal, platykurtic symmetric, positively skewed, and negatively skewed.



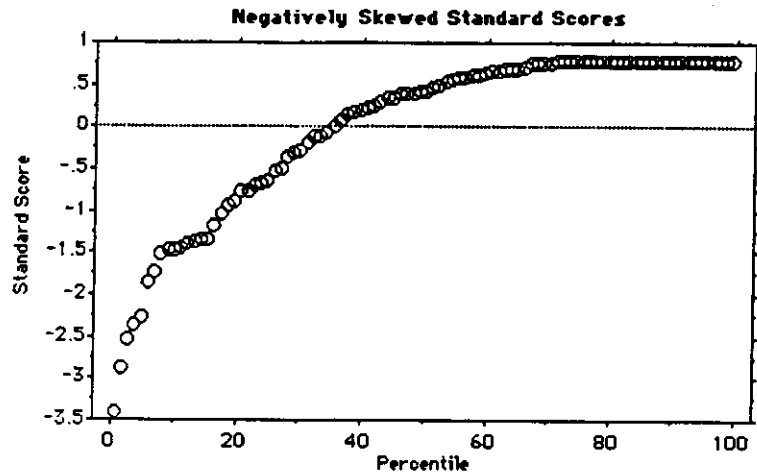
For a normal distribution, a small change in an observed value near the center of the distribution will be associated with a relatively large percentile change. A change in an observed value at either the low or high end of the observed value distribution will result in only small percentile changes. Therefore, the relative frequency curve of observed values rises most rapidly at the low and high ends of the percentile range and has a plateau in the middle part.



For a symmetric, platykurtic distribution, a change of an observed value at almost any location will cause a relatively constant change in the percentile (except at the extreme ends of the distribution, where large observed value change will be associated with small percentile change). Therefore, the relative frequency curve should rise at a constant rate.



For a positively-skewed distribution a change of one observed value unit near the lower end of the distribution will be associated with relatively large percentile changes. As you move toward the right tail of the distribution, a change of one observed value unit will have less and less percentile change associated with it. Thus, a relative frequency curve associated with a positively skewed distribution will rise slowly from lower left to the upper right with the rise becoming more apparent as the curve approaches the right side of the graph.

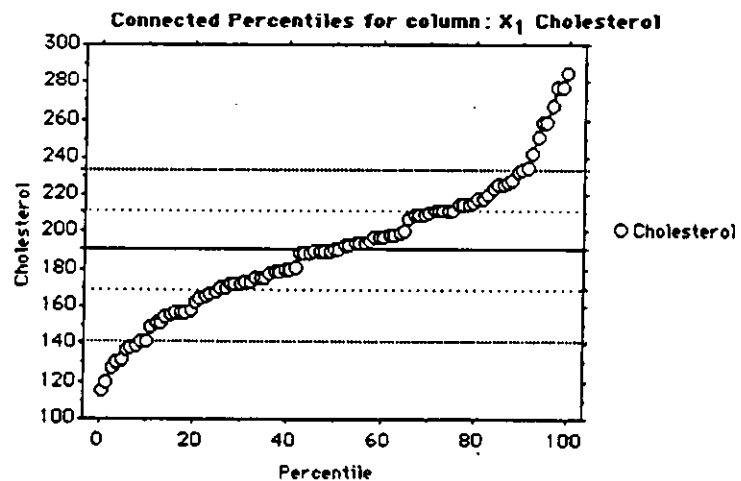


For a negatively-skewed distribution a change of one observed value unit near the lower end of the distribution will be associated with relatively small percentile changes. As you move toward the right tail of the distribution, a change of one observed value unit will have more and more percentile change associated with it. Thus, a relative frequency curve associated with a negatively skewed distribution will rise rapidly from lower left to the upper right with the rise becoming less apparent as the curve approaches the right side of the graph.

The normal distribution is a mesokurtic curve, and you should therefore think of the percentile plot associated with mesokurtic distributions as being sigmoid. The other graphs shown above are leptokurtic curves and are characterized by a cumulative frequency curve that shows a very steep slope as the plot rises from lower left to upper right.

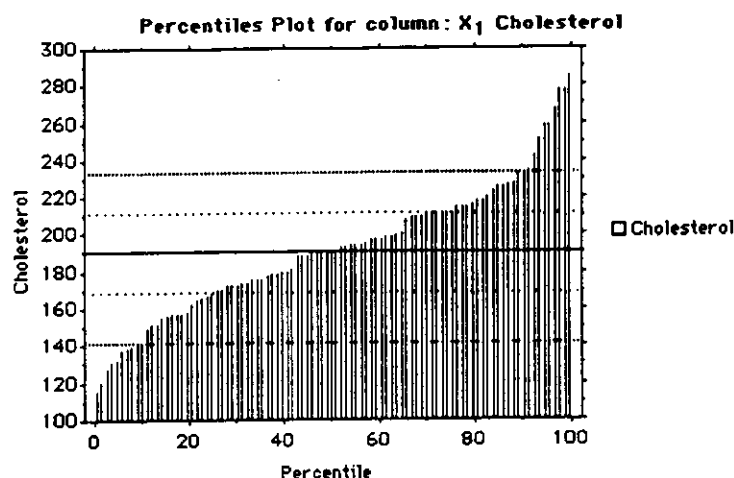
At times you may have so few observed values in a particular range that it is difficult to ascertain the slope of the percentile plot. In such a situation it is helpful to have all points connected by a line.

- Select Line Chart from the View menu.



The resulting percentile plot is identical to the previous percentile plot except that it also has a line going through all of the plotted points.

- Select Bar Chart from the View menu.



The resulting plot represents still another way of looking at the percentile plot. Rather than plotting the points, a bar is dropped from what would be the center of a point to the percentile baseline. As may be clear, these are two alternative graphic methods for viewing percentile data. The graphic method you choose is a function of your preference.

Summary

This discussion has shown how StatView allows observation of the statistical variation of a variable in several different ways. We started by computing the mean, standard deviation and other summary statistics for the Cholesterol data, observing the results first in tabular and then in graphic form.

Next, percentiles were used as a reference point for our analysis. Once again both tables and graphs were observed. Although we looked at the mean, standard deviation and other summary statistics first, and then the percentiles, you may select all or any combination of the descriptive statistics and percentiles simultaneously.

Graphic Comparisons

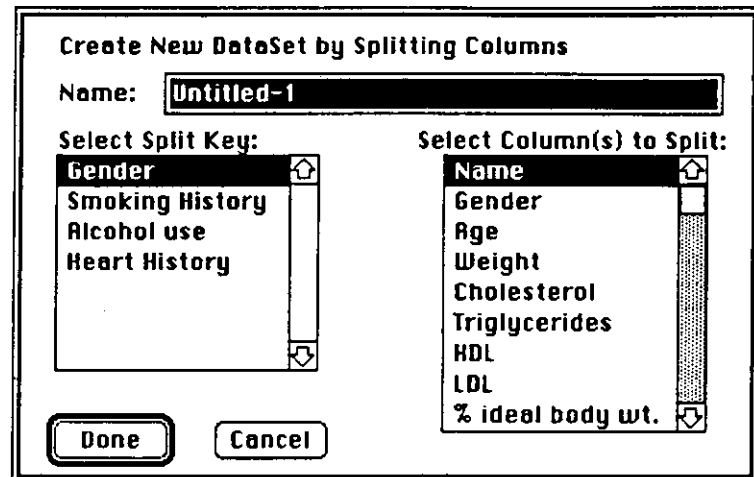
StatView computes descriptive statistics for as many X variables as you specify. More importantly, StatView graphic features let you compare the distributions of your variables through composite graphs. The graphs include error bars and box plots. For many of these statistics, it is easier to use columns that have been split with the Split Columns command in the Tools menu. See Chapter 7 for more information on this command.

Making Descriptive Comparisons Around the Mean

Virtually all of the graphic comparison options of StatView assume that variables selected for comparison have a common unit of measure. Thus, if you attempted to compare Systolic BP with Cholesterol, you would get a graphic comparison with the unit of measurement being a combination of both variables. The lower end of the continuum would be determined by Systolic BP while the upper end of the continuum would be defined by Cholesterol. The comparative graphic representations would be impossible to interpret.

In this example, we will not compare two of the existing variables in Lipid Data. Instead, we will take the Cholesterol variable and split it into two cholesterol measurements, one for females and one for males. StatView makes it especially easy to split variables on the basis of associated categorical variables.

- Open Lipid Data.
- Select **Split Columns** from the Tools menu. A dialog box appears with the split key list on the left and the column to be split list on the right.



- Select **Gender** as the split key and **Cholesterol** as the column to split.
- Click **Done** to create the dataset.

When the dialog box disappears you will note a new dataset **Untitled-1**. This dataset contains two columns, the first representing the cholesterol measurements of the male subjects, the second the cholesterol measurements of the females:

Untitled-1		
	male - Cholesterol	female - Cholesterol
1	197	181
2	190	131
3	172	285
4	233	215
5	194	178
6	155	227
7	234	197

- Assign **X** to male - Cholesterol and to female - Cholesterol, in that order.
- Select **Mean, Std. Dev., etc.** from the **Describe** menu.
- Select **Confidence Intervals** from the **Describe** menu and click **OK**.

The following table appears showing the results for the **X₁** column, male - Cholesterol:

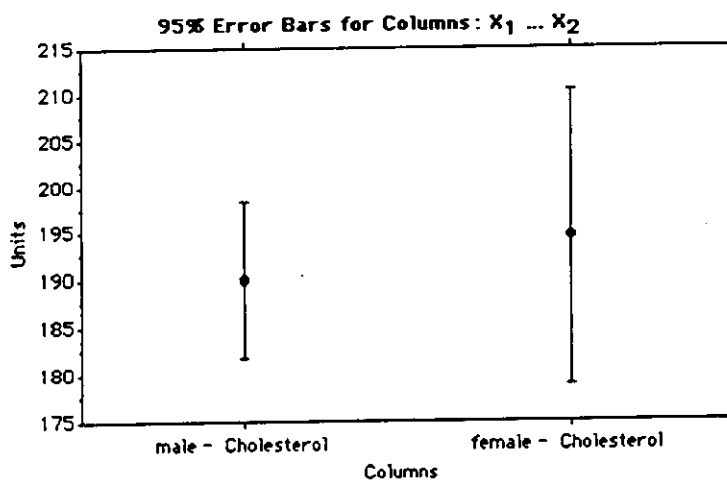
X ₁ : male - Cholesterol					
Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
190.085	35.299	4.189	1246.021	18.57	71
Minimum:	Maximum:	Range:	Sum:	Sum of Sqr.:	* Missing:
115	277	162	13496	2652602	0
t 95%:	95% Lower:	95% Upper:			
8.355	181.729	198.44			

- Click the bottom arrow in the scroll bar to view the X₂ column. You see the results for the next column, the X₂ column, female - Cholesterol.

X ₂ : female - Cholesterol					
Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
194.625	37.322	7.618	1392.94	19.176	24
Minimum:	Maximum:	Range:	Sum:	Sum of Sqr.:	* Missing:
131	285	154	4671	941131	47
t 95%:	95% Lower:	95% Upper:			
15.76	178.865	210.385			

While we could impose interpretations on these data similar to those imposed in the individual variable description section, it will turn out that the graphic views of X₁ and X₂ together will be much easier to interpret, from a comparative perspective, than the table views. Note that when StatView split the data values, it created a dataset with enough rows to handle the 71 male values. Because there are only 24 female values, the remaining rows (47) are filled with missing values.

- Select Scattergram from the View menu. Make sure that the composite/paging control is set for a composite view, showing both columns.



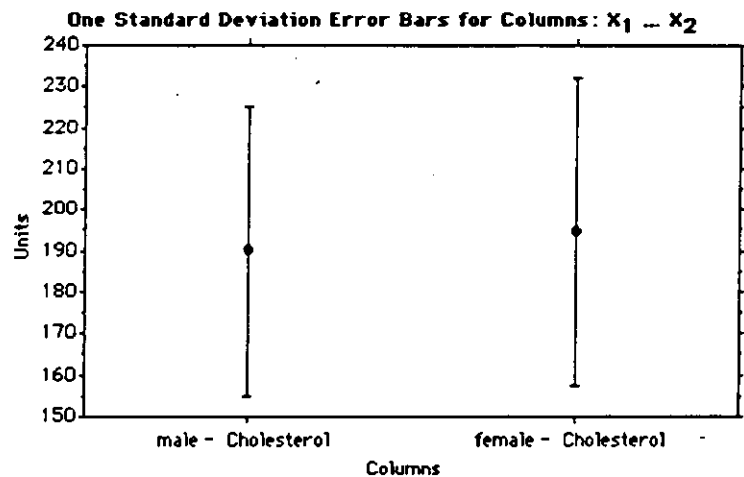
This view shows a single-tiered error bar for each variable. The ends of the bars represent the 95% confidence intervals, and are marked by the short horizontal lines. The means for these two groups, 190.085 for the males and 194.625 for the females, appear to be different, but the error bars overlap, indicating that the two cholesterol samples could be from the same population. It is reasonable to conclude that the difference of 4.54 that we see between the two means is just a consequence of chance.

More precisely, if we repeatedly sampled cholesterol from the female population, using samples of size 24, and constructed 95% confidence intervals for each sample, we would expect that 95% of the confidence intervals would span the

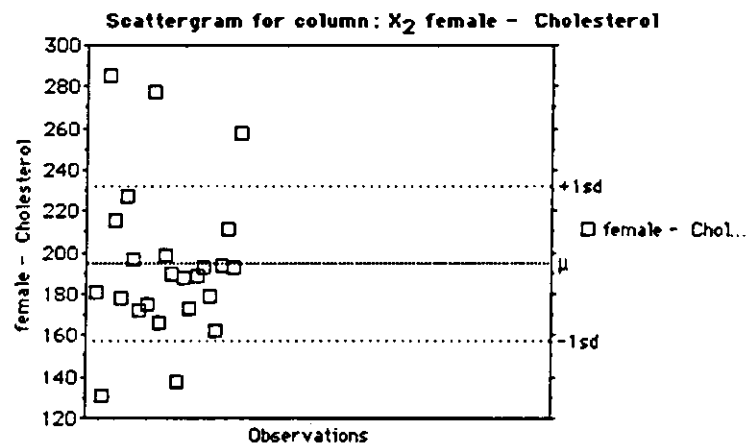
actual population mean cholesterol level, and that 5% of them would not span the actual population mean. It is apparent that the female cholesterol confidence band completely spans the male cholesterol confidence band. If we were to do a t-test (which we will do later), we might conclude that the female cholesterol mean is not significantly ($p < .05$) different from the male cholesterol mean.

- Select **Confidence Intervals** from the **Describe** menu.
- Click **Remove** to remove the confidence intervals from the selected descriptive statistics.

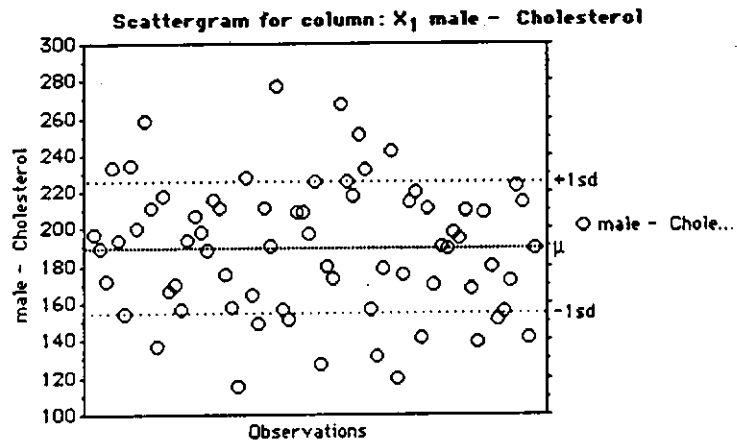
The view shows a single-tiered standard deviation bar for each variable. The center of each bar represents the mean of the associated variable. This view allows you to quickly look at the two variables to determine whether or not the two associated distributions might overlap with each other on the same continuum. It would appear from our scattergram that the male and female Cholesterol distributions overlap almost completely.



A much more informative view may be seen by clicking the composite/paging tool to view a separate univariate scattergram for each variable. We have discussed the interpretation of the univariate scattergram in the graphic section for individual variables. It is possible to page through the view windows for the two variables to compare the univariate scattergrams for outliers and the extent to which the two variables exhibit different skewness. Notice that for the females about 67 percent of the observed values, 14 observed values, are below the mean. This would imply that the female cholesterol distribution is positively skewed. This observation can be confirmed by computing the coefficient of skewness for these data, .906.



The male distribution, however, appears to be almost symmetric, with 36 observed values below the mean and 35 observed values above the mean. This can be confirmed by computing the index of skewness for the male Cholesterol data. This index is .063, while a symmetric distribution would be 0.

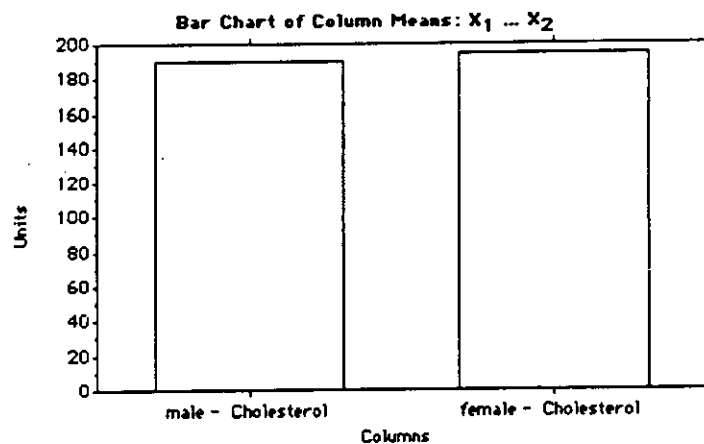


These two scattergrams are difficult to compare because they differ in sample size. The male sample being much larger than the female sample gives the erroneous visual impression that the males are much more heterogeneous than the females. If the samples were of approximately the same size such a visual impression might tend to be valid.

It is difficult to understand the difference in variance between the males and the females when comparing their scattergrams. For some datasets there will be very clear sub-groupings of the observed values either at the mean, above the mean, or below the mean. Such groupings suggest the existence of subgroups within the dataset. However, these scattergrams suggest that both the male and female subgroups tend not to have subgroups with regard to Cholesterol count.

- Select Bar Chart from the View menu. Once again make sure that the paging/composite tool is set for a composite view.

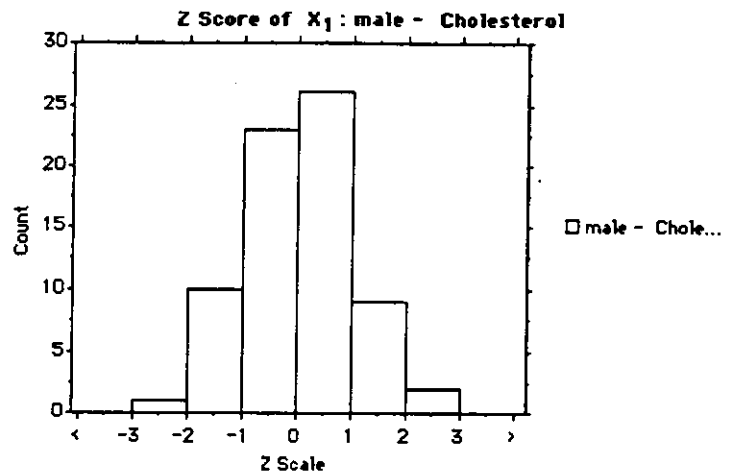
This chart compares the variable means:



- Click the composite/paging tool, and the view displays the z-score distribution for each variable grouping.

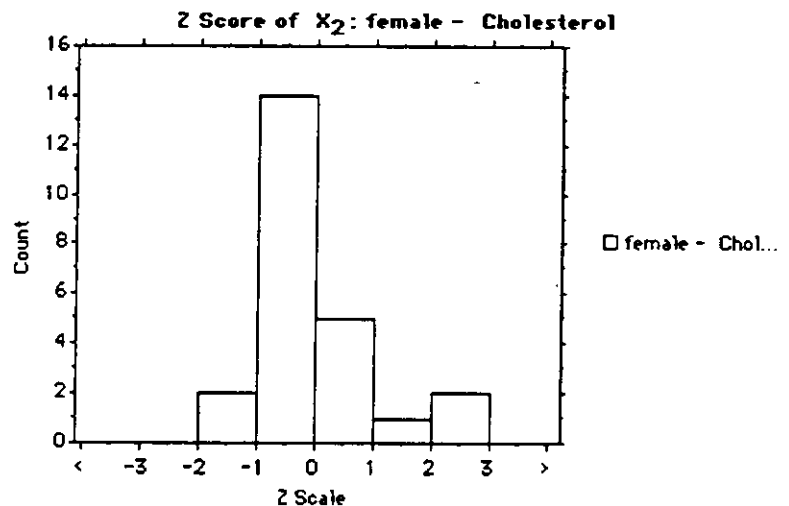
As noted in the discussion of single group graphics, the bars drawn represent a standard deviation unit. The bar between 0 and 1 represents a frequency count of the number of points between the mean and one standard deviation above the

mean, that is, with z-scores from 0 to 1. Also as previously noted this view gives us a very rough picture of the variable distribution. The male and female z-distributions can be observed by paging through the two pages.



These two views of the standard scores are somewhat redundant when considering the conclusions that we have already made. For the z-scale bar chart associated with the male Cholesterol count, there is no discernable skewness, as the distribution is almost symmetric.

For the bar chart associated with the female Cholesterol distribution it is very easy to note the positive skewness. Notice that the right side of the distribution extends from 0 to 3 while the left side of the distribution extends only from 0 to -2. Clearly the right side of the distribution is longer than the left side.



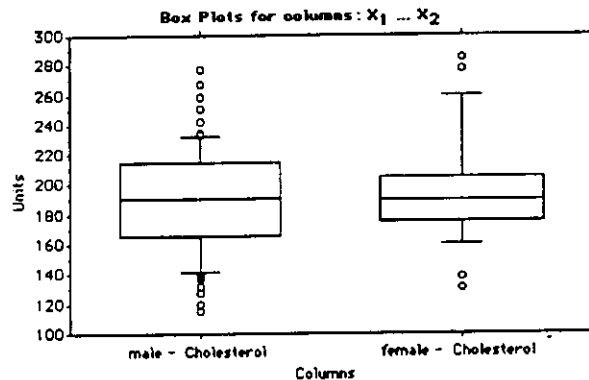
- Select **Table** from the **View** menu.
- Select **None** from the **Describe** menu, clearing previously selected statistics.
- Select **Percentiles** from the **Describe** menu.

The table view shows the results for the X₁ variable, male - Cholesterol. By turning the pages, you see the results for female - Cholesterol:

X2: female - Cholesterol					
* < 10th %:	10th %:	25th %:	50th %:	75th %:	90th %:
2	159.6	174	189.5	205	259.9
* > 90th %:					
2					

- Select **Box Plot** from the **View** menu. Make sure that the paging/composite tool is set for composite (click the tool if necessary).

Note that in a box plot the total box plot is always contained in the view; therefore, visual comparisons of the box lengths will be a valid procedure.



This composite view shows the box plots for each of the X variables selected: in our example, the male and female Cholesterol levels. With these two box plots a great deal of comparative information may be seen. You will recall that we discussed box plots in detail when we discussed the graphic description of a single variable.

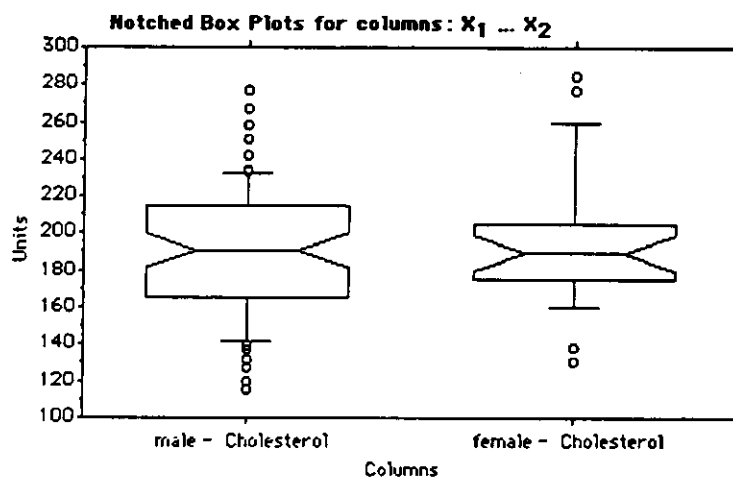
Because the box associated with the males is "thicker", we can conclude that the middle 50 percent of the Cholesterol distribution from the male sample is more heterogeneous than the middle 50 percent associated with the female sample. Notice the difference in whisker lengths for the females and the similarity in whisker lengths for the males. The fact that the whisker associated with the higher female Cholesterol counts is clearly longer than the whisker associated with the lower female Cholesterol counts suggests a positively skewed distribution. There will most likely be a longer tail at the upper end of the female Cholesterol distribution than at the lower end.

Notice also that the extreme Cholesterol counts for the men are close to the whisker ends, whereas the extreme Cholesterol counts for the females are somewhat removed from the ends of the whiskers. These female outliers, when considered within the context of the homogeneity of the middle 50 percent of the female distribution, provide an explanation for the greater female variability. As a group, females tend to have a homogeneous Cholesterol count, but those females who have extreme Cholesterol counts, either high or low, tend to be very extreme.

This explains why the standard deviation for female Cholesterol is larger than the standard deviation for male Cholesterol. It is possible that the standard deviations for these datasets are misleading. For a majority of the male and female samples, the females tend to be more homogeneous than the males, as is demonstrated by the females' "slimmer" box in the box plot.

- Click the composite/paging tool, and the view window displays the box plot for each variable on a separate page.

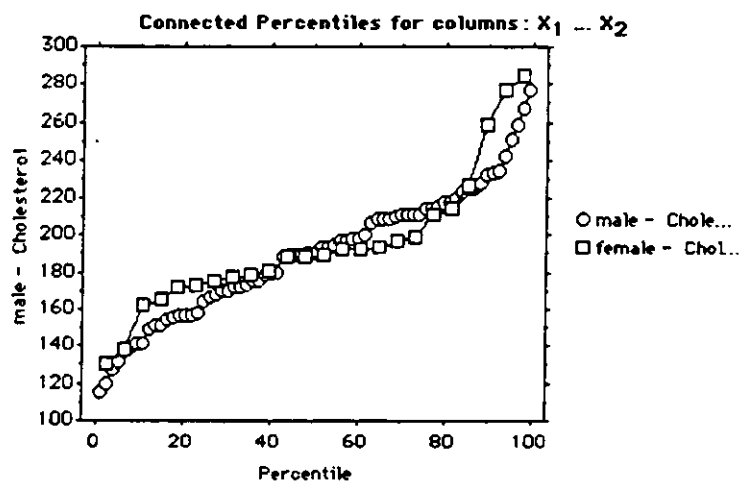
- Click the composite/paging tool again and you are returned to the composite view.
- Click the notch control. The resulting notched box plots allow you to make a quick visual comparison of the male Cholesterol median confidence band with the female Cholesterol median confidence band:



The width of the notches suggests that the two median confidence bands are not only approximately the same width but also overlapping almost completely. The medians may be assumed to be the same.

Three other views of percentiles are available: scattergram, bar, and line chart. These percentiles plots may be a composite or may have only one variable per page. You may use the composite/paging tool to switch between the composite and single variable presentations.

- Select Line Chart from the View menu.



This view is an effective descriptive comparison between male and female Cholesterol. Note that the ordinate represents observed values, Cholesterol level, and that the abscissa represents percentiles. This type of plot was discussed in detail in the univariate description section.

Notice that the two percentile plots are reasonably comparable. The female plot rises more rapidly than the male plot at both the low and high ends of the Cholesterol continuum. This would suggest that the female distribution is more mesokurtic than the male distribution. Furthermore the gradual, almost

unchanging slope associated with the male plot also suggests that the male plot may tend to be a bit platykurtic. The very abrupt increase in slope at the end of the female plot may be attributed to the female outliers.

Ideally, if two distributions are similar they would have been identical in appearance, but not necessarily superimposed.

Summary

These examples have shown how the graphic features of StatView can compare either several variables or several samples on a single variable. The first example used the mean as a starting point for observing Cholesterol variation among males and females. Next, percentiles were used as a reference point for our comparisons. In each case, the variation between the variables was made clear through different graphic views of error bars and box plots.

Once again, it is important to understand that while we looked at the mean statistics first and then the percentiles, you may select to compute one or all of the descriptive statistics.

- Select **Table** from the **View** menu.
- Select all of the descriptive statistics except **Frequency Distribution**.

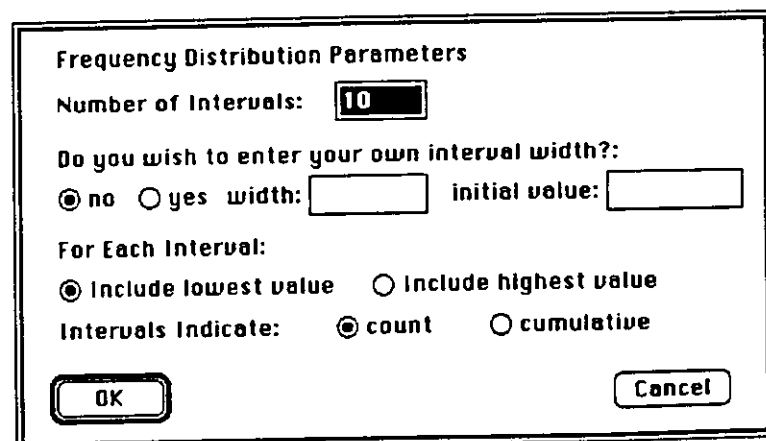
The table includes the results for all selected statistics. Each page of the view window contains the results for a different X variable. All pages of the results may be printed.

StatView computes frequency distributions for X variables. Frequency distributions treat category and continuous (integer, long integer, and real) data differently. The first example is of the Cholesterol distribution which is defined by data that are continuous.

Frequency Distribution

Frequency Distribution of Continuous Data Variables

- Open **Lipid Data** and assign X to **Cholesterol**.
- Select **Frequency Distribution** from the **Describe** menu, and the following dialog box appears:



The dialog box is titled "Frequency Distribution Parameters". It contains the following elements:

- Number of Intervals:** A text box containing the value "10".
- Do you wish to enter your own interval width?:** Two radio buttons, "no" (selected) and "yes".
- width:** A text box next to the "yes" radio button.
- initial value:** A text box next to the "yes" radio button.
- For Each Interval:** Two radio buttons, "Include lowest value" (selected) and "Include highest value".
- Intervals Indicate:** Two radio buttons, "count" (selected) and "cumulative".
- OK** and **Cancel** buttons at the bottom.

You can specify the number of intervals in the distribution to plot using the box labelled **Number of Intervals**. The maximum number of intervals is 1000. The next parameter determines the interval width. If you do not enter your own width, StatView creates a width equal to the data range plus one divided by the number of intervals. You can enter your own interval width and the initial value in the respective text entry rectangles.

The next parameter in the dialog box determines whether an interval includes the lowest value in the interval or the highest value. If you include the lowest value, each interval contains the count of data values that are greater than or equal to the lower limit of the interval and less than the upper limit of the interval. If you include the highest value, each interval contains the count of data values that are greater than the lower value and less than or equal to the upper value.

The final parameter determines whether the intervals indicate the count of values or the cumulative count of values. You specify whether you want to have the height of a bar determined by the frequency of observed scores in the interval or by the cumulative frequency of scores up to the upper limit of the interval.

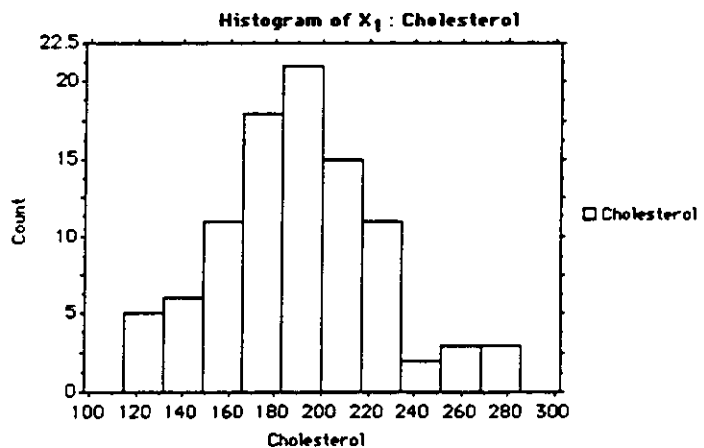
- Click OK. The following table appears:

X ₁ : Cholesterol				
Bar	From: (≥)	To: (<)	Count	Percent
1	115	132.1	5	5.263%
2	132.1	149.2	6	6.316%
3	149.2	166.3	11	11.579%
4	166.3	183.4	18	18.947%
5	183.4	200.5	21	22.105%
6	200.5	217.6	15	15.789%
7	217.6	234.7	11	11.579%
8	234.7	251.8	2	2.105%
9	251.8	268.9	3	3.158%
10	268.9	286	3	3.158%

—Mode

This table summary of the total Cholesterol data as it will be used to construct a histogram. The range is 171; the highest observed value plus one (286) minus lowest observed value (115). The default interval length is (171/10) or 17.1 units. The intervals are then constructed from the lowest to the highest score. The table summarizes the histogram bars in terms of the frequency within the bar and the percentage associated with the bar. Either summary may be used to determine the heights of the bars in the histogram. If there is a modal interval, it is labeled.

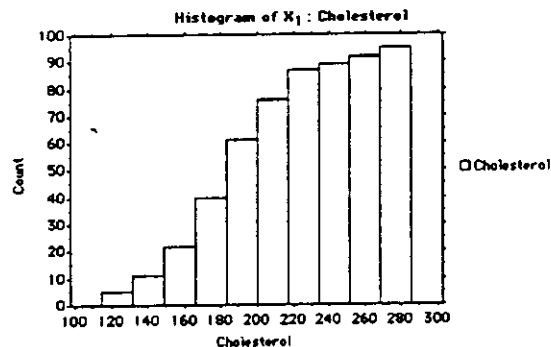
- Select Bar Chart from the View menu.



You now have a graphic representation of the table. If you have selected more than one X variable, scrolling the pages will show you the histograms for successive X variables.

It is also possible to represent the data cumulatively. Cumulative data are defined by intervals whose counts represent the frequency of observed values below the upper limit of the associated bar. Such a distribution is very similar to a percentile plot, but the axes differ. The bars increase in height from lower left to upper right. The ordinate represents frequencies.

- Select Frequency Distribution from the Describe menu.
- Click Cumulative.
- Click OK.



The interpretation of this cumulative histogram is quite similar to the interpretation of the percentile plot.

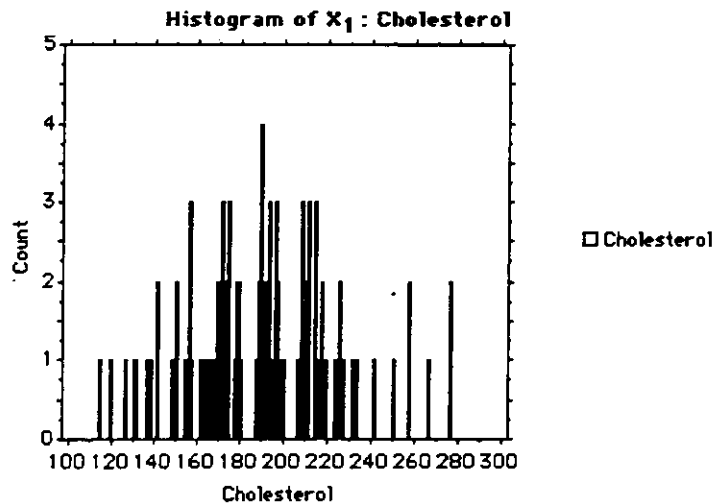
Now suppose that you want the most accurate graphic portrayal possible for a given set of data. Ideally you would want a histogram with a bar corresponding to every observed value in the dataset.

- Select None from the Describe menu to clear all statistics.
- Select Mean, Std. Dev., etc. from the Describe menu.
- Select Table from the View menu.

The table view provides you with the basic descriptive data for Cholesterol. The range, 170 for Cholesterol, provides some sense of how many bars will be required to represent all possible observed values. For our Cholesterol data, each observation is measured to the nearest unit, and therefore we will need 170 bars. If our data were measured to the nearest .5, we would require 340 bars. Each bar should span from the real lower limit to the real upper limit of the observed value associated with it. The table view also shows us that the lowest value in the Cholesterol distribution is 115. The real lower limit of 115 is 114.5 and the real upper limit of 115 is 115.5. If we select an interval width of 1, for 1 Cholesterol unit, and use 114.5 as our initial value, then we will need 170 intervals to cover the range of values from 115 to 285, the maximum observed Cholesterol value.

- Select Frequency Distribution from the Describe menu.
- Enter 170 for the Number of Intervals.
- Enter 1 for the interval width.

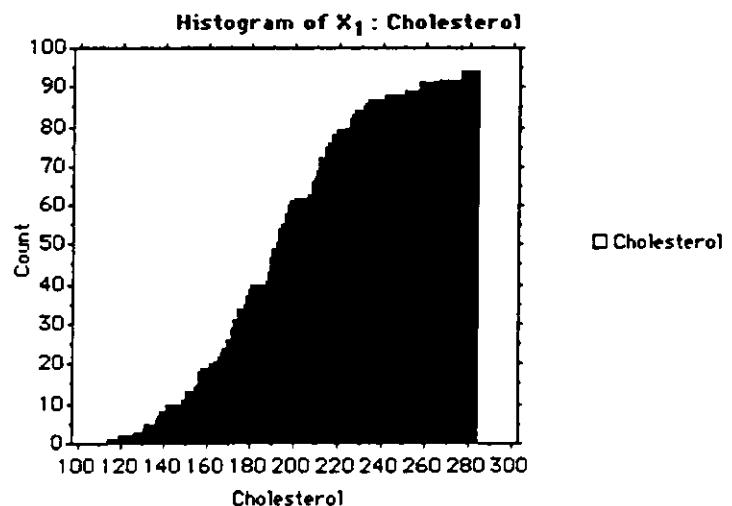
- Enter 114.5 as the initial value.
- Click Include lowest value.
- Click Count.
- Click OK.
- Select Bar Chart from the View menu. The resulting histogram has a bar to represent each unique observed value and represents every peak and trough in the data:



Because there are gaps in the data (that is, possible observed values that do not occur in the dataset), some of the histogram bars appear to be a bit isolated and have wide troughs.

The histogram based on cumulative frequencies looks much better than the histogram based on simple counts.

- Select Frequency Distribution from the Describe menu.
- Click Cumulative.
- Click OK.



The data gaps are now represented by the rough surface of the distribution.

Frequency Distribution of Category Data Variables

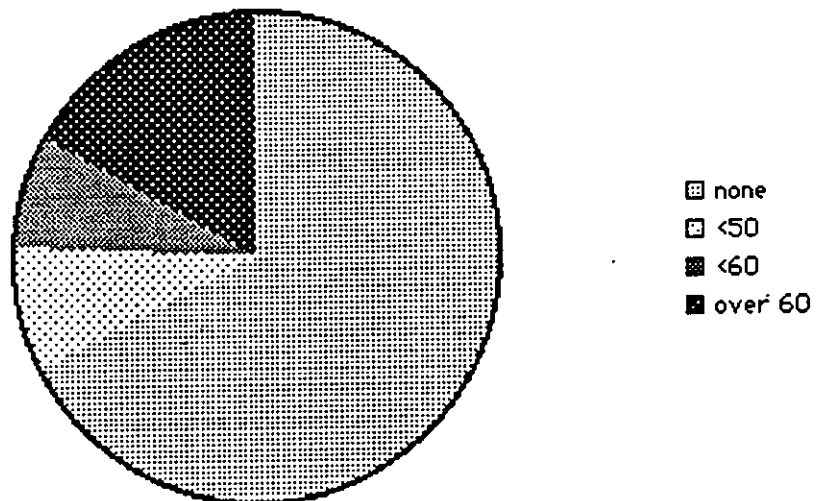
Suppose we wished to look at the heart history data in Lipid Data: whether there has been a history of heart attack and the general age at which it occurred. This is a categorical variable and has four categories: none for no history of attack, <50 indicating at least one heart attack in the family prior to the age of 50, <60 at least one heart attack in the family prior to age 60 but after age 50, >60 at least one heart attack in the family after age 60.

- Open Lipid Data.
- Assign X to Heart History.
- Select **Frequency Distribution** from the **Describe** menu.

The value for the number of intervals defaults to the highest number of elements contained in the category column selected. If the number of intervals selected is greater than the number of elements, StatView only displays the information on the category elements. If the number of intervals chosen is less than the number of elements in the category set, StatView displays information on only the selected number of intervals, and indicates with a message that the category set contains more elements than are being displayed. Interval width and lowest/highest value do not apply for category variables and their settings are ignored. You may choose to display count or cumulative information.

- Select **Count** and click **OK**.
- Select **Pie Chart** from the **View** menu.

Pie Chart of X₁ : Heart History

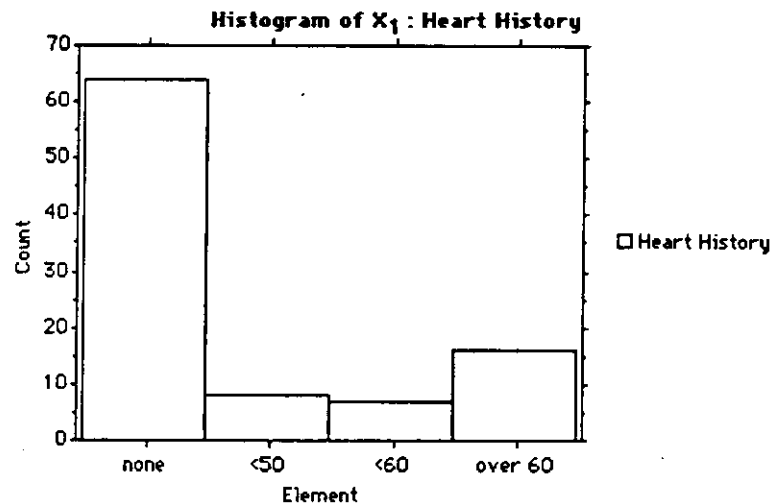


Notice that the order of the categories in the pie chart corresponds to their order in the legend. The order of the categories in the legend is a function of their order at data entry. The category of "none" corresponds to "element 1" at the time that the categories were defined. In a somewhat analogous fashion the "over 60" category corresponded to "element 4" at the time that the categories were defined.

You can very quickly estimate that approximately 60-75% of the sample has no history of heart attack. It is also possible to conclude that the proportion of respondents with a history of heart attacks prior to age 50 is approximately the same as the proportion of respondents with a history of heart attacks before 60 but

after 50. You may also conclude from the pie chart that the proportion of respondents with a history of heart attacks after 60 is about equal to the proportion of respondents with a history of heart attacks before 60.

- Select **Bar Chart** from the **View** menu.



You now have a histogram view of the heart history data. It would be more difficult to make comparative statements with the histogram than with the pie chart. It is a matter of personal preference whether you prefer bar charts or pie charts. StatView allows you to use either or both.

- Select **Table** from the **View** menu.

The table displays the category element name, the frequency with which that element appears in the current X column, and the percentage of the data it represents.

X₁ : Heart History

Bar:	Element:	Count:	Percent:	
1	none	64	67.368%	-Mode
2	<50	8	8.421%	
3	<60	7	7.368%	
4	over 60	16	16.842%	

Although pie charts are available for all frequency distribution data, they make the most sense for category data. Such data have names associated with regions rather than bar numbers. It is very difficult to recall what a bar represents without having a copy of either the histogram table or the histogram chart.

The logic of a pie chart is to represent the bars in such a fashion that the proportion of the whole associated with any one category is readily seen. It also facilitates visual comparisons among the bars. The visual complexity of a pie chart increases with the number of bars associated with it.

Summary

StatView allows you to create frequency distributions of your data with up to 1000 bars. The interval widths, initial values, whether to include observed values equal

to the lower limit or the upper limit of the interval in the count, and whether the histogram is a frequency count or cumulative count can all be specified.

The frequency information can be viewed in both tabular and graphic forms. The graphic forms include bar charts, also referred to as histograms, and pie charts.

Finally, StatView recognizes the difference between category and continuous data in determining and displaying frequency distributions.

Chapter 6 — The Compare Menu

The statistics in the **Compare** menu expect different variable assignments and display results differently depending on those assignments. The table below lists the required variables for each type of statistic and the associated result format.

Type	Variables Required	Results
OneX*	One X variable	A result is calculated for each X variable.
ManyX	Several X variables	One result is calculated using all X variables.
OneXOneY	One X variables and one Y variable	If there is a single X variable and multiple Y variables, a result is calculated for each Y variable. If there is a single Y variable and multiple X variables, a result is calculated for each X variable. If there are multiple X and Y variables, a result is calculated for each X_iY_i pair (matching subscripts).
ManyXOneY	Several X variables and one Y variable	A result is calculated for each Y variable.
OneXManyY	One X variable and several Y variables	One result is calculated using all Y variables.

*All statistical tests in the **Describe** menu are OneX statistics.

The following discussions use several data sets found in the Sample Datasets folder. For each statistic example, we state which datasets are used; make sure the correct dataset is open. The statistics are presented here in the order they appear on the menus.

Compare Percentiles

The **Compare Percentiles** command compares 19 corresponding percentiles of two variables, an X and a Y. It is a OneXOneY statistic. (See the table at the beginning of this chapter.) The percentiles compared are 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 96, 97, 98, 99.

In particular, this procedure allows the user to compare two data distributions. The n th percentile of a dataset is the value such that $n\%$ of the data are less than or equal to the value. The percentiles of one distribution are plotted against the corresponding percentiles of a second distribution. Usually such comparisons are made between distributions that represent measures on the same variable. The graph is always drawn as a box with the abscissa defined by the X variable and the ordinate defined by the Y variable.

If the two distributions are comparable, the 19 plotted points will fall on a straight line passing from the lower left corner to the upper right corner of the graph. If a point is above the line then the sample associated with the ordinate has a higher score associated with the given percentile point than the sample associated with the abscissa. Alternatively, if a point is below the line, the sample associated with the abscissa has a higher score associated with the point than the sample associated with the ordinate. In general, this percentile comparison graph is most informative when comparing the distribution forms of a variable in two samples. They may be either the same group or two different groups.

For example, compare the Cholesterol Distribution for males to the Cholesterol Distribution for females.

- Select **Split Columns** from the **Tools** menu.

Create a new dataset using Gender as the split key and Cholesterol as the column to split. (See Chapter 7 for more information on splitting columns.)

- Assign X to male - Cholesterol.
- Assign Y to female - Cholesterol.
- Select **Compare Percentiles** from the **Compare** menu. The following table will appear.

Table View

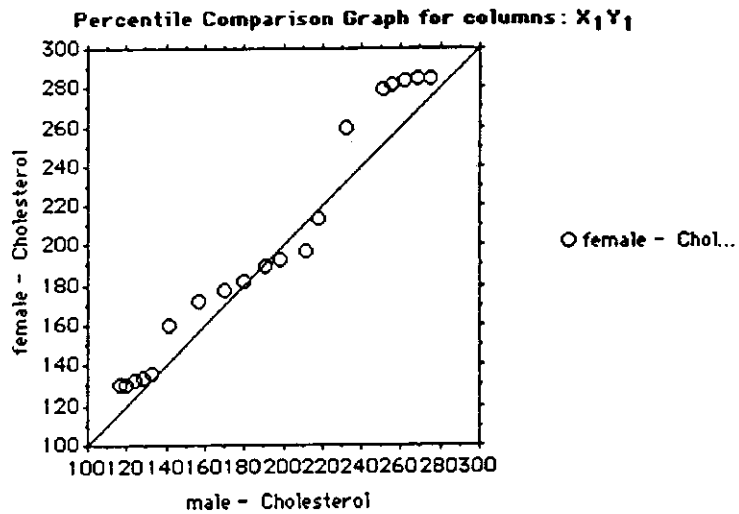
Percentile Comparison for X ₁ : male - Cholesterol			Y ₁ : female - Chole...		
% male - C...female - ...			% male - C...female - ...		
1	116.05	131	20	157	172.3
2	119.6	131	30	170	177.1
3	124.41	132.54	40	179.9	181.7
4	128.7	134.22	50	191	189.5
5	132.25	135.9	60	198.3	193
10	142	159.6	70	211.2	197.6
			80	218	213.8
			90	232.4	259.9
			95	250.55	279.4
			96	255.62	281.32
			97	261.33	283.24
			98	267.8	285
			99	274.9	285

This table view provides all the information that will be used to plot the percentile comparison graph. The names of the X and Y variables analyzed are shown in the view title. The percentile compared is indicated in the % column. The 19 variable percentile values are displayed in their respective columns.

Notice that below the median, the 50th percentile, the female Cholesterol count is higher than the male Cholesterol count at every percentile. However, between the 50th and 80th percentiles the male Cholesterol count is higher than the female Cholesterol count. At the upper extreme, the 90th to the 99th percentile the females once again exceed the males.

Graphic Views

- Select **Scattergram** from the **View** menu.



The graph drawn is a square graph; both axes are equal in length and the line $x=y$ is drawn. The points plotted are the 19 comparison percentiles listed above. At both extremes of the line, the points are above it, suggesting that the females have higher Cholesterol counts at the extremes than do the males. You can see the data converging on the line near the median. This convergence suggests that the middles of the two distributions are similar.

The scattergram view contains a tool specific to the statistic, the equal axis tool,



Clicking on this control rescales the display to use the maximum display. The graph will no longer be square and the $x=y$ reference line is removed. Both of these graphs, the square and non-square form, may be viewed as Line Charts.

t-Test

Frequently, you will want to compare two means to see if they are comparable. Furthermore, when making such a comparison, you may also be questioning whether or not it is reasonable to assume that the samples from which the means were computed could have come from the same population. StatView provides three different types of t-Tests: one group t-Test, paired two group tests, and unpaired two group tests. t-Tests must be performed on raw data, not calculated means.

- Select t-Test from the Compare menu, and the t-test dialog box is displayed.

Select t-Test:

☒ One Group t-Test

Two Group Tests:

☐ Paired ☐ Unpaired

t Values: ☐ one tail ☒ two tail

OK
Cancel

You can specify which t-Test to calculate. You can also specify whether the probability values correspond to one- or two-tail tests of significance.

One Group t-Test

The one group t-Test compares a sample mean to a hypothesized population mean. It is a OneX statistic. (See the table at the beginning of this chapter.)

The hypothesis tested in this example is that the observed sample mean is consistent with the mean specified in the dialog box. Our dataset, Lipid Data, is based on 95 subjects. These subjects were educated with regard to the relationship between Cholesterol count and diet. 43 of the subjects had Cholesterol samples taken three years after their initial measures. We have ordered these subjects in our dataset so that they are the first 43 subjects. The last column of the dataset is Cholesterol Loss. It represents the difference between the initial Cholesterol measure and the subsequent Cholesterol measure some three years later (see Formula in Chapter 7 to learn how to create a new column by formula from variables in your dataset).

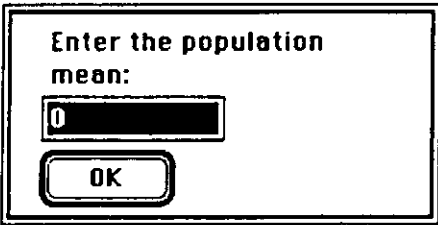
The average Cholesterol loss is 9.767 units. Unfortunately, not all subjects showed a reduction in Cholesterol count. Indeed, some subjects elevated their Cholesterol by as much as 73 units while others reduced their Cholesterol by as much as 62 units. A researcher might question whether the average Cholesterol Loss of 9.767 units is just chance variation. That is, can we assume that 9.767 is significantly greater than 0?

This is a one-tail test. Had we asked the more fundamental, and somewhat uninformed question, of whether or not 9.767 is different from 0 we would be implying a two-tail test. We do not know either the population mean or the population standard deviation with regard to Cholesterol Loss. The one group t-Test will use our sample standard deviation for Cholesterol Loss, 27.627, as an estimate of the population standard deviation. On the basis of a sample of size 43 with a standard deviation of 27.627, is it possible that the sample mean of 9.727 units does not represent a real cholesterol loss and is just a chance fluctuation from 0?

Using Lipid Data, let us assign X to the Cholesterol Loss column and perform the correct t-test.

- Assign X to Cholesterol Loss.
- Select t-Test from the Compare menu to display the t-Test dialog box.
- Click one tail.
- Click OK, performing the default One Group t-Test.

This dialog box appears prompting you to enter the population mean.



The dialog box has a title bar and a main area. The title bar contains the text "Enter the population mean:". Below the title bar is a text input field with the number "0" entered. Below the input field is an "OK" button.

- Click OK. You see:

One Sample t-Test X1 : Cholesterol loss

DF:	Sample Mean:	Pop. Mean:	t Value:	Prob. (1-tail):
42	9.767	0	2.318	.0127

Note: 52 cases deleted with missing values.

This one group test essentially addresses the question of whether or not a sample of size 43 with a mean of 9.767 could have been selected from a population that actually had a mean of 0. The table tells you that it is quite improbable that this sample was drawn from a population with a mean of 0. Specifically, the probability of obtaining a mean of 9.767 or larger just by chance alone if the population mean is 0 is .0127. Assume that you are operating at the .05 significance level, then you will assume that the mean of 9.767 represents a real positive difference because the reported probability level, .0127, is a probability that is less than .05. The dieting is probably effective in reducing cholesterol counts for the sample.

No graphic views are available.

Paired Two Group t-Test

The paired two group t-Test computes a paired t value between an X and Y column, where each row entry for both columns is assumed to be a measure on the same subject. Sometimes this test is also referred to as a dependent sample or repeated measures t-test. It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

The paired two group t-Test compares an X mean and a Y mean determined from either the same sample of respondents or two samples that are known to be dependent. The null hypothesis assumes that the two means have been defined by samples from the same population, and therefore we would expect the difference between the means to be 0. The analysis addresses the question of whether or not the observed difference between the two means is a chance difference. This difference might either be larger than we might expect by chance, which is a one-tail test, or be more extreme than we might expect by chance, which is a two-tail test.

We have many variable measures on the 95 subjects in Lipid Data, but it would be unreasonable to make comparisons between variables that represent very different measures. For instance, if we were to compare the Systolic BP with Cholesterol variables, we would most certainly find a statistically significant difference, but it wouldn't have any substantive meaning. In our discussion of the one group test, we noted that 43 subjects receiving dietary education reduced their Cholesterol count over a three-year period. You might question how their diet modification affected their weight. Did the diet modification also bring about a weight loss? Is their weight at the end of three years significantly less than it was when they began their diet modification? This is a one-tail hypothesis that specifically questions whether the mean weight associated with the third year measure is statistically less than the mean weight associated with the initial measure. We are using the same group of subjects for the computation of both means.

- Open Lipid Data. Assign X to Weight.
- Assign Y to Weight-3yr.
- Select t-Test from the Compare menu to display the t-Test dialog box.

- Click one tail.
- Click OK. The following table appears:

Paired t-Test X₁: Weight Y₁: Weight-3yr

DF:	Mean X - Y:	Paired t value:	Prob. (1-tail):
42	-1.907	-1.558	.0634

Note: 52 cases deleted with missing values.

The mean weight associated with the initial measures was approximately 164.5 whereas the mean weight associated with the measures taken some three years later was 166.5. You do not even need a statistical test to tell you that the sample did not lose weight. You can compare the two means visually and immediately see that the second mean is certainly not less than the first mean. The associated probability is .0634. Assume that you are operating at the .05 significance level, then you will assume that the mean of 166.5 is not less than the mean of 164.5 and that any difference between the third year mean weight and the initial mean weight is due solely to chance fluctuation. Evidently, the dieting, while effective in reducing cholesterol counts for the sample, does not tend to reduce weight for the sample.

No graphic views are available.

Unpaired Two Group t-Test

The unpaired two group t-Test computes an unpaired t value comparing the means between two groups in a single Y column. The groups in the Y column are specified by an X column which must be a Category or Integer column containing exactly two groups. For a category column the number of groups is the number of distinct values in the column. (Note that this is not necessarily equivalent to the number of elements defined for the category.) For an integer column, StatView calculates the number of groups as the column's Maximum Value - Minimum Value + 1. For example, you get an error message if your integer column contains the values 1 and 3, because StatView sees three groups in this column rather than two.

If you have more than two groups you should consider a one-way analysis of variance. Sometimes this test is also referred to as an independent sample or non-repeated measures t-test. It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

The unpaired two group t-Test compares two sample means determined from two independent samples. Statistically, we assume that the two means have been defined by samples from the same population, and therefore we would expect the difference between the means to be 0. The statistical analysis addresses the question of whether or not the observed difference between the two means is a chance difference, either larger than we might expect by chance, a one-tail test, or more extreme than we might expect by chance, a two-tail test.

Unlike the paired t-test example, we don't have to worry quite as much about illogical analyses. When we have independent samples StatView essentially splits a single Y column on the basis of the groupings associated with the X column. Consider again the Cholesterol counts for males and females in Lipid Data. All of our preliminary graphic work suggested that there simply is no difference between mean Cholesterol levels of males and females. If you were going to generate a

hypothesis on the basis of our preliminary analyses you would hypothesize that the male Cholesterol mean, 190.085, is not significantly different from the female Cholesterol mean, 194.625. This is a two-tail hypothesis.

- Open Lipid Data. Assign X to Gender.
- Assign Y to Cholesterol.
- Select t-Test from the Compare menu to display the t-Test dialog box.
- Click two tail.
- Click OK, performing the default Unpaired t-Test. The following table appears:

Unpaired t-Test X1: Gender Y1: Cholesterol				
DF:		Unpaired t Value:		Prob. (2-tail):
93		- .537		.5926
Group:	Count:	Mean:	Std. Dev.:	Std. Error:
male	71	190.085	35.299	4.189
female	24	194.625	37.322	7.618

Assuming a significance level of .05, the table, not surprisingly, suggests that the difference between the male and female Cholesterol levels, 4.54, is most likely due to chance and may assumed to be 0. The Cholesterol counts for the two samples are not significantly different from each other.

No graphic views are available.

Correlation Coefficient

A correlation coefficient (sometimes called Pearson's) indicates the degree of linear relationship between two variables. Generally, it is assumed that you have a single sample with two sets of observed values, an X and a Y observed value on each subject or sampling unit.

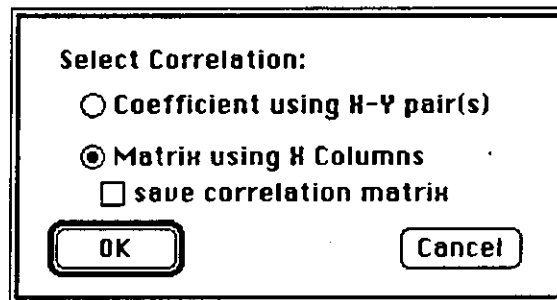
A positive correlation suggests that, as the observed values on one variable increase or decrease, so do the observed values on the other variable increase or decrease proportionately. A negative correlation coefficient suggests that, as the observed values on one variable increase, the observed values on the other variable decrease proportionately. A correlation of 0 suggests that there is not a linear relationship between the two variables.

The previous analyses of Cholesterol counts taken before and three years after exposure to dietary effects on cholesterol suggested that the Cholesterol counts were reduced over the three year period. You might wonder whether there is a linear relationship between Cholesterol counts taken during the two time periods. A positive correlation would suggest that there was most likely a general decrement in Cholesterol regardless of initial cholesterol level. A negative correlation would suggest that those with large initial Cholesterol counts tended to show extremely large cholesterol decrements while those with low initial Cholesterol counts showed an actual increase in cholesterol count to the extent that the subject's rank order on Cholesterol was actually reversed. A 0 correlation would suggest that there is a random pattern of increase and decrement with

regard to initial cholesterol. Of the three possible outcomes, that which is associated with a positive correlation seems to be the most reasonable expectation.

There are two ways that StatView can be used in calculating a correlation. Either a correlation coefficient can be computed between an X and Y variable or a correlation matrix can be computed relating a set of X variables. The correlation coefficient option provides both a brief summary table and a scattergram of the X-Y pair. The matrix option, typically associated with a number of X variables, does not provide a scattergram and does not provide a table summary of the relationship. Rather, the matrix approach defines a StatView dataset that is a correlation matrix. Such a matrix may be saved and may also be modified for future use with factor analysis.

- Select **Correlation** from the **Compare** menu and the Correlation dialog box is displayed:



Options lets you specify computing either a correlation coefficient between an X and Y variable or a correlation matrix a set of X variables.

If you choose to compute a correlation matrix, you have the option of saving the matrix as a StatView dataset.

Correlation Coefficient

Let us assume that we are going to determine the correlation between initial Cholesterol and Cholesterol measured three years later, using Lipid Data. Our expectation, as rationalized above, is that we will find a positive correlation. You must assign an X and a Y variable to compute a correlation coefficient. This choice for a correlation calculates the correlation coefficient between individual X and Y variables. It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

- Open Lipid Data. Assign X to Cholesterol.
- Assign Y to Chol-3yrs.
- Select **Correlation** from the **Compare** menu. Note that the default with an X and a Y assigned is **Coefficient using X-Y pair(s)**.
- Click **OK**. The following table appears:

Corr. Coeff. X_1 : Cholesterol Y_1 : Chol-3yrs

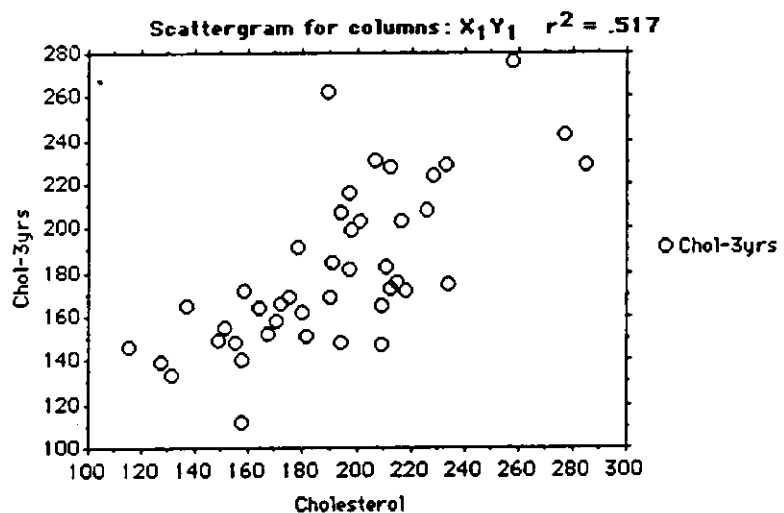
Count:	Covariance:	Correlation:	R-squared:
43	976.878	.719	.517

Note: 52 cases deleted with missing values.

As expected, the correlation, .719, is a reasonably large positive coefficient. This suggests that there is a substantive positive linear correlation between the two measures of Cholesterol.

The square of this coefficient, .517, suggests that approximately 52% of the variation in third year Cholesterol may be predicted given initial Cholesterol measurement. It would probably be worth while to investigate this relationship further with a regression analysis.

- Select Scattergram from the View menu.



It is apparent that the points of the scattergram rise from the lower left to the upper right, thereby suggesting a positive relationship. The scattergram is also informative from the perspective of calling our attention to the fact that the points become heteroscedastic as the associated Cholesterol observed values become larger. There are several outliers that might have a distorting effect on the magnitude of the correlation coefficient. In particular there is an individual who had an observed Cholesterol value of approximately 200 on the initial measure and an observed Cholesterol value of approximately 260 three years later and another with a level of 260 on the initial measurement and a level of 280 three years later, both the reverse of the general trend. If these two sets of scores are eliminated from the dataset the correlation increases in magnitude to .75. However, you are well advised to understand the nature of outliers before eliminating them from analyses. We will return to these outliers when we address the question of regression.

Correlation Matrix

This choice calculates a correlation matrix for all assigned X variables. The matrix uses only cases that are complete (non-missing or row-wise deletion for missing

values) across all X variables. It is a ManyX statistic. (See the table at the beginning of this chapter.)

- Open Lipid Data.
- Assign X to Weight, Cholesterol, Triglycerides, HDL, and LDL, in that order.
- Select **Correlation** from the **Compare** menu to display the Correlation dialog box. Note that the default with many X assigned is **Matrix using X columns**.
- Click **save correlation matrix**. Click **OK**. The following table appears:

Correlation Matrix for Variables: X1 ... X5

	Weight	Choleste...	Triglyce...	HDL	LDL
Weight	1				
Cholesterol	-.022	1			
Triglycerides	.108	.401	1		
HDL	-.276	.352	-.278	1	
LDL	.057	.962	.489	.083	1

This view gives the number of X columns used. The correlation matrix provides the names of all variables compared and the correlation coefficient values. If there are more than eight variables, the matrix occupies additional pages. This view represents an extension of the correlation coefficient. It computes all pairwise correlations between the selected X variables.

Some correlations reported in the correlation matrix may not appear to agree with the value reported as a single correlation when correlation coefficient has been selected. This discrepancy may arise from row-wise deletion of incomplete data.

When a correlation matrix is selected, the correlations are based only on the observed values associated with cases that are complete across all X variables selected. When a correlation coefficient has been selected the correlation is based only on the observed values associated with cases that are complete across the X and Y variable selected. Thus, it is possible that, when there are missing data, a correlation computed by a correlation matrix may be based on fewer cases than it would be if it were computed with a correlation coefficient.

There are no graphic views available.

Saving a Correlation Matrix as a Datafile

An option in the correlation coefficient dialog box allows the correlation matrix to be saved as a StatView dataset. Bring the new window to the front of the screen by selecting it from the **Windows** (or Σ) menu.

	Name	Weight	Cholesterol	Triglycerides	HDL	LDL
1	Weight	1.000	-.022	.108	-.276	.057
2	Chole...	-.022	1.000	.401	.352	.962
3	Trigly...	.108	.401	1.000	-.278	.489
4	HDL	-.276	.352	-.278	1.000	.083
5	LDL	.057	.962	.489	.083	1.000

Regression

The complete square correlation matrix is saved. The first column is a string column containing the variable names. The data values in the new dataset are displayed to the number of decimal places specified for analysis results, but the variables are saved to 18 decimal places. You can display more places using **Format** from the **Tools** menu.

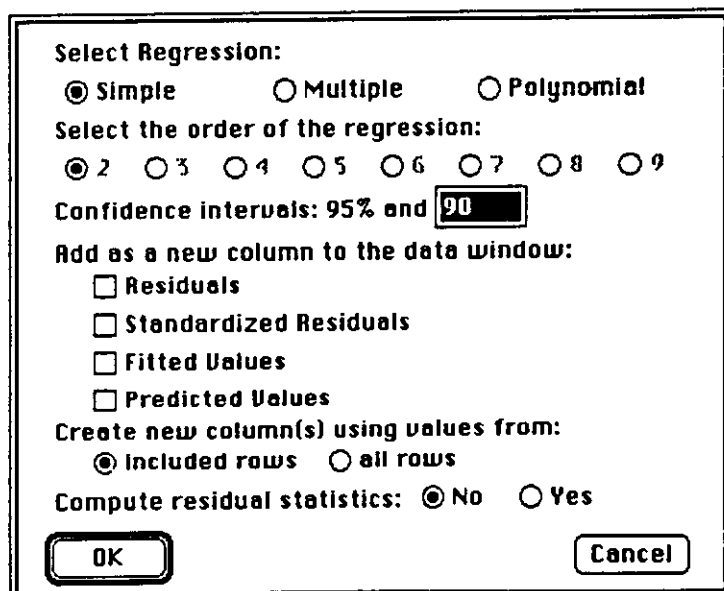
The **Regression** choice in the **Compare** menu computes three models: simple linear regression, multiple regression and polynomial regression.

Each analysis computes:

- R (Pearson's correlation coefficient), R^2 , adjusted R^2 , root mean square residual
- ANOVA table
- residual statistics
- beta coefficient table
- confidence interval tables

The following values can be computed and added in columns to the data window:

- residuals
- standardized residuals
- fitted values;
- predicted values
- Select **Regression** from the **Compare** menu, and the **Regression** dialog box is displayed.



The dialog box is titled "Select Regression:". It contains several sections: "Select the order of the regression:" with radio buttons for 2 through 9; "Confidence intervals: 95% and" followed by a text box containing "90"; "Add as a new column to the data window:" with checkboxes for Residuals, Standardized Residuals, Fitted Values, and Predicted Values; "Create new column(s) using values from:" with radio buttons for Included rows and all rows; and "Compute residual statistics:" with radio buttons for No and Yes. At the bottom are "OK" and "Cancel" buttons.

Select Regression:

☒ Simple ☐ Multiple ☐ Polynomial

Select the order of the regression:

☒ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

Confidence intervals: 95% and

Add as a new column to the data window:

☐ Residuals

☐ Standardized Residuals

☐ Fitted Values

☐ Predicted Values

Create new column(s) using values from:

☒ Included rows ☐ all rows

Compute residual statistics: ☒ No ☐ Yes

Select the regression model you wish to analyze using radio buttons at the top of the box. If you select a **Polynomial Regression**, select the order of the highest

power in the regression model using radio buttons directly below the regression type buttons.

You cannot leave the intercept out of the model, that is, you cannot force the intercept through the origin. If you have tests where this is required, you may be interested in Abacus Concept's companion product, SuperANOVA.

Two confidence intervals can be calculated. The 95% confidence interval is automatically calculated. Enter another interval in the text entry rectangle following the 95% label, if you wish.

Selecting the appropriate options saves the following information as new columns to the data window:

Option	Description
Residuals	Computed as the difference between the observed Y_i and the fitted Y_i .
Standardized Residual	Computed as the conversion of the residuals to unit normal deviate form, mean of 0 and standard deviation of 1.
Fitted Values	Computed by the regression model as the fitted Y_i .
Predicted Values	If you wish to apply the regression equation to additional data not present in the dataset, the resulting values are referred to as predicted values. To compute predicted values, enter the independent values after the last complete case for the model in the X column. That is after the last case where both X and Y values are present. These new entries would be the last rows of the dataset. The Y values must be left missing.

Selections made using these check boxes are added as new columns to the end of the dataset. If a case has been deleted from the model, the new column contains a missing value in that row. The rows containing the X-Y pairs used in the regression analysis contain missing values in the Predicted Values column.

You may specify whether the above values are determined for just the included rows of the dataset or for all rows of the dataset.

If you select **Included Rows**, the values are calculated for just the included rows of the dataset; excluded rows contain missing values.

If you select **All Rows**, the values are calculated for all rows of the dataset regardless of their included or excluded state. With this option, you can fit a model for one group of data and use this model to predict for all groups. The included rows are used to estimate model coefficients. The excluded rows are used to compute predicted values

You may choose to compute residual statistics. StatView automatically selects this if you choose to add any new columns to the data window in the dialog box.

Simple Regression

This choice calculates a simple regression between a dependent Y variable and an independent X variable. It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

Previously, with correlation and Lipid Data, we looked at the correlation between initial Cholesterol count and Cholesterol count three years after the subjects received instruction on reducing cholesterol through dieting. The resulting correlation coefficient was quite large, $r=.719$. If we wished to predict possible Cholesterol reduction over a three year period through dieting could we predict the future Cholesterol value for an individual given the individual's current Cholesterol count? Note that we are not predicting Cholesterol reduction. This is a regression question.

The regression model assumes two sets of individuals from the same population: those from whom the regression statistics are derived and those to whom the regression results are generalized. A regression equation is derived to predict the dependent variable from the independent variable and then summary statistics are computed to determine how well the regression model will work.

- Open Lipid Data. Assign Y to Chol-3yrs.
- Assign X to Cholesterol.

In order to compute predicted values, we must enter new data at the end of the X variable, Cholesterol, column.

- Enter the data points 160 and 260 at the bottom of the Cholesterol column. This will create two new rows in the dataset. These will be the Cholesterol scores for which predicted Chol-3yrs scores will be obtained.
- Select Regression from the Compare menu to display the Regression dialog box.
- Click Predicted Values.under the heading Add as a new Column... The Compute residual statistics box will activate.
- Click OK. The following table appears:

Simple Regression X1 : Cholesterol Y1 : Chol-3yrs				
Count:	R:	R-squared:	Adj. R-squared:	RMS Residual:
43	.719	.517	.506	25.589

Analysis of Variance Table				
Source	DF:	Sum Squares:	Mean Square:	F-test:
REGRESSION	1	28782.079	28782.079	43.956
RESIDUAL	41	26846.666	654.797	p = .0001
TOTAL	42	55628.744		

Residual Information Table			
SS[e(i)-e(i-1)]: e ≥ 0:	e < 0:	DW test:	
42372.135	21	22	1.578

Note: 54 cases deleted with missing values.

The count represents the number of paired observations used in computing the correlation. R is the correlation between Cholesterol and Chol-3yrs. The value for R^2 is the square of the correlation coefficient, and is interpreted as the proportion of variance of the dependent variable that is predictable from the independent

variable. Thus, you can also estimate R^2 from the ANOVA table as the ratio of the regression sum squares to the total sum squares. The adjusted R^2 is the square of the correlation adjusted for sample size and the number of coefficients in the regression equation. It is the unbiased estimate of the population squared correlation coefficient. The root mean square residual is just the square root of the mean square for residual of the ANOVA table. The root mean square residual represents the standard deviation of the residuals which are the errors of prediction.

The ANOVA table represents a partition of the total sum of squares into predictable, regression, and unpredictable, residual sum of squares. The regression mean square is the variance of the fitted values while the residual mean square is the variance of the residual values. The F-ratio, listed under F-test, is formed as the ratio of the regression Mean Square to the residual mean square. The p value under F-test represents the probability of an F-ratio of 43.956, with 1 and 42 degrees of freedom, would occur by chance sampling fluctuation. If you are operating at the .01 level of significance, you would conclude that these are significant results. That the amount of predictable variance, regression mean square, is significantly greater than 0.

The regression model assumes that the residuals are uncorrelated. In the situation where the data are from an ordered sequence, it is assumed that the residuals are not dependent upon their neighbors.

It is very important to note that our Cholesterol data are not appropriate for the Durbin-Watson test. However, to facilitate an understanding of this test, let us assume that the order of the paired observations is fixed and meaningful. To test the assumption that the residuals occur independently of the order of the paired observations the Durbin-Watson is used. The Durbin-Watson statistic, 1.578, is the sum of squares of the difference between successive residuals, divided by the residual sums of squares. To interpret the Durbin-Watson statistic it is necessary to consult a Durbin-Watson table. If the statistic is less than the range in the table, it is assumed that the residuals are independent. If the statistic is greater than the range in the table, it is assumed that the residuals are not independent and that they are correlated. If the statistic falls within the range in the table, the test is inconclusive.

Using the table provided by Neter and Wasserman, we find that for a sample of 43 the range in the table for the statistic is 1.48 to 1.57 for a .05 level of significance. Thus, if we consider our value of 1.578 as exceeding 1.57 then we must conclude that the residuals are not independent. However, remember that this conclusion is meaningless within the context of our Cholesterol data since the order of the paired observations is arbitrary.

Also in the residual information table, we have a count of the frequency of positive residuals and the frequency of negative and 0 residuals. These two frequencies should be approximately equal if the residuals are independent. If we find widely discrepant frequencies, we can conclude that a linear regression model may not be the most appropriate regression model for the data.

- Click the scroll bar. You see:

Simple Regression X_1 : Cholesterol Y_1 : Chol-3yrs

Beta Coefficient Table

Variable:	Coefficient:	Std. Err.:	Std. Coeff.:	t-Value:	Probability:
INTERCEPT	47.328				
SLOPE	.702	.106	.719	6.63	.0001

Confidence Intervals Table

Variable:	95% Lower:	95% Upper:	90% Lower:	90% Upper:
MEAN (X,Y)	173.631	189.392	174.945	188.079
SLOPE	.488	.915	.523	.88

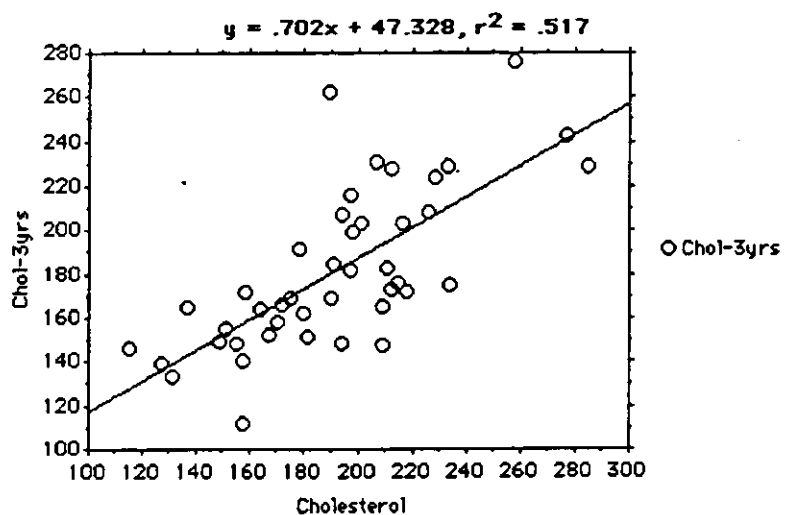
Predicted : Column 26

The beta coefficient table provides the information required for the linear regression equation. Assume Y_i to be the i th fitted value for Chol-3yrs. Assume X_i to be initial Cholesterol count for the i th person. Then the general regression equation is: $Y_i = .702 X_i + 47.328$. These coefficients are taken directly from the beta table. The beta table also has a two-tailed test to determine if the slope is significantly different from 0. For simple regression, this test provides information that has already been determined in the ANOVA table. Notice that when the F-ratio has one degree of freedom, it is the square of the corresponding t-ratio.

The confidence intervals table provides the 95% and 90% confidence bands for the slope and for the mean of the dependent variable, Y_1 . The second confidence interval value, in this case 90%, is user-entered through the Regression dialog box. The interpretations for these confidence intervals is the same as the interpretation discussed in the Confidence Interval section of Chapter 5.

If you bring the dataset window to the front of the screen and scroll to the right you will see that Column 26 has been added to the data window. Its last two rows contain the predicted values, 159.569 and 229.720, that correspond to the values 160 and 260 entered at the bottom of the Cholesterol Column.

- Select Scattergram from the View menu.



A scattergram of the points along with the fitted regression line for predicting Y_1 given X_1 is displayed. Notice in this plot that the regression line appears to pass beneath most of the points associated with initial Cholesterol counts greater than approximately 230, and Chol-3yrs. This suggests that the residuals are not independent of the observed values and that there will be positive residuals associated with these points.

This scattergram contains a tool specific to the **Simple Regression**, the confidence

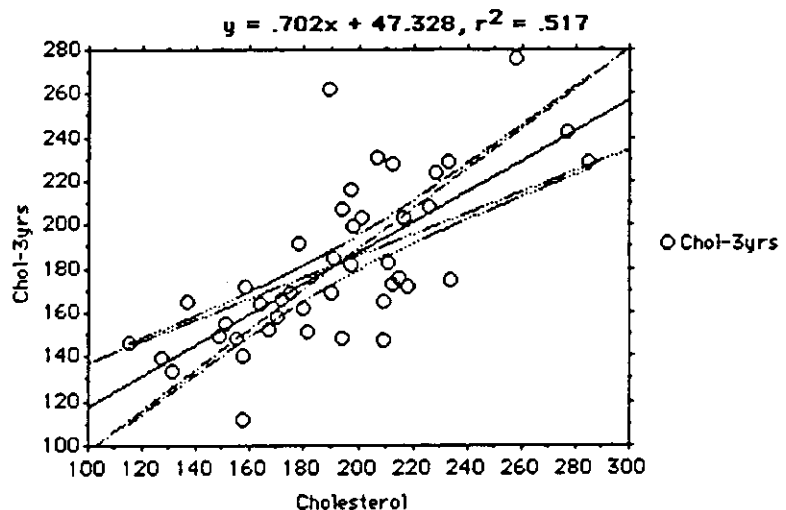
bands control:  Clicking this control presents the following dialog box.

Select confidence information for a Simple Regression
☐ 95% confidence limits for slope of regression line
☐ 95% confidence bands for the true mean of Y
☐ 90% confidence limits for slope of regression line
☐ 90% confidence bands for the true mean of Y

OKCancel

For both confidence intervals, 90% and 95%, you can plot:

- the confidence bands for the slope of the regression line
- the confidence band for the mean of the dependent variable Y₁
- Click the both choices for the 95% confidence bands, and click OK.



The confidence bands for the slope of the regression line appear to pass through the regression line, regardless of the level of confidence. Actually, the confidence bands touch the regression line at the point of intersection of the X and Y means on the regression line. These confidence bands provides some sense of the stability of the slope of the regression line. These bands provide a range within which the true regression line may fall.

The confidence bands for the mean of the dependent variable, Y given X, follow a similar pattern. The further removed the regression line is from the point of intersection of the X and Y means, the less stable is the predicted Y point of the regression line for the associated X. It is clear for our Cholesterol data, as well as for most other data, that the consequences of sample instability are most severe at the extreme ends of the X distribution.

Of course, if you had selected the 90% confidence band, the bands would be wider than for the 95% confidence band.

- Click the confidence bands tool once more and remove the confidence bands.

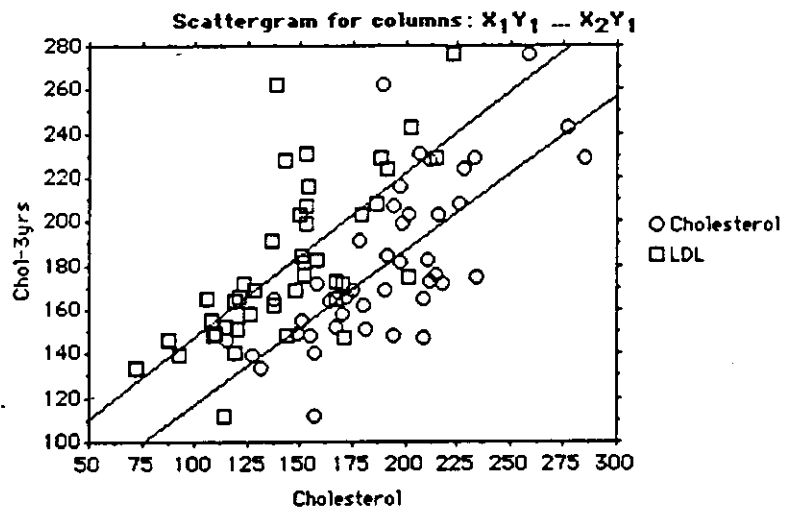
You can plot more than one regression model on a single graph. If you assign a second Y variable, Y_2 , or a second X variable, X_2 , the scattergram will display the results for this new variable. Recall from the demonstration of the correlation matrix that LDL, Low Density Cholesterol, had a high correlation with the initial measure of Cholesterol. It should also have a high correlation with Cholesterol measured three years later, especially if the initial Cholesterol is highly related to Cholesterol measured three years later, Chol-3yrs. However, assuming that the degree of linear relationship is essentially the same, you may wonder whether the slopes of the two regression lines are the same or whether the nature of the relationship has changed.

- Assign X to LDL.

The second regression model, using LDL as an independent variable, is calculated. The scattergram now displays the results for the dependent Y, Chol-3yrs, against the second X variable, LDL.

- Click the paging/composite tool.

The scattergram now displays the results of the second analysis within the context of the first regression scattergram. Notice that two sets of points are plotted and represented by different shapes. Clearly, the slopes of the two regression lines are for all practical purposes parallel thereby implying that the slopes for the two lines are not different and that the nature of the relationship has not changed.



Care must be taken when using this option. If you select two variables for X that have a discrepant metric associated with them, such as HDL and Cholesterol, the resulting scattergram can take on a rather deceptive form that does not represent either scattergram adequately.

- When you close Lipid Data do not save the changes made to the dataset in this example.

Multiple Regression

The Multiple Regression selection from the regression dialog box calculates a multiple regression between a dependent Y variable and one or more independent X variables. Variables are entered into the equation in the order of their subscripts,

X_1 to X_n . It is a ManyXOneY statistic. (See the table at the beginning of this chapter.)

Let's use the other blood variables in Lipid Data (Triglycerides, HDL, and LDL) to predict Chol-3yrs. These are reasonable substitutes for initial Cholesterol level. However, rather than using a single variable with a single regression equation we will use three variables with a single regression equation. The rationale for using several independent variables is that each may have a unique effect upon the dependent variable. Such a model assumes that the variables predict more collectively than each could predict independently of the other two.

- Open Lipid Data. Assign Y to Chol-3yrs.
- Assign X to Triglycerides, HDL and LDL, in that order.
- Select Regression from the Compare menu. You see the Regression dialog box.
- Click Multiple.
- Save the following three values to the data window by clicking Residuals, Standardized Residuals, and Fitted Values.
- Click OK. The following table appears:

Multiple Regression Y1 :Chol-3yrs 3 X variables				
Count:	R:	R-squared:	Adj. R-squared:	RMS Residual:
43	.721	.52	.483	26.171

Analysis of Variance Table				
Source	Df	Sum Squares	Mean Square	F-test
REGRESSION	3	28916.585	9638.862	14.073
RESIDUAL	39	26712.139	684.927	p = .0001
TOTAL	42	55628.744		

Residual Information Table			
SS[e(i)-e(i-1)]: e ≥ 0:	e < 0:	DW test:	
42643.36	21	22	1.596

Note: 52 cases deleted with missing values.

This information is interpreted in a fashion identical to the interpretation discussed above in the simple regression section with the exception that the ANOVA table F-ratio is no longer the square of a t-value from the beta coefficient table.

This multiple regression model accounts for approximately 52% of the variance associated with Chol-3yrs. The multiple regression model accounts for a statistically significant portion of the variance, ($F=14.073$; $DF=3,39$; $p<.0001$). That is, the proportion of variance accounted for, 52%, is more than you would expect just by chance alone.

- Click the down arrow on the scroll bar.

Multiple Regression Y1 :Chol-3yrs 3 X variables					
Beta Coefficient Table					
Variable:	Coefficient:	Std. Err.:	Std. Coeff.:	t-Value:	Probability:
INTERCEPT	50.252				
Triglycerides	.028	.065	.062	.427	.6715
HDL	.618	.42	.184	1.473	.1487
LDL	.694	.146	.654	4.751	.0001

Residual : Column 26 Std. Residual : Column 27 Fitted : Column 28

If there are more than 6 coefficients, 6 independent variables, the table is continued on successive pages. Assume Y_i to be the i th fitted value for Cholesterol-3yrs from the 3 variables. Assume X_{1i} to be the Triglyceride value for the i th person. Furthermore assume X_{2i} to be the HDL value for the i th person and X_{3i} to be the LDL value for the i th person. Then the general regression equation to determine Y_i is:

$$Y = .028X_{1i} + .618X_{2i} + .694X_{3i} + 50.252$$

The entries of the beta coefficient table may be interpreted in a fashion similar to the interpretation applied to the beta coefficient table associated with simple regression. The t-value and associated probability is a two-tailed test to determine whether or not the associated regression coefficient is significantly different from 0. It would appear from this table that only the LDL coefficient is significant, $p < .0001$. The standard coefficient is the standardized regression coefficient for use with standard scores, as opposed to the coefficients which are used with the observed scores.

The note on the bottom of the page names the columns which contain the computed values selected from the regression dialog box. If any values could not be added to the dataset because of memory constraints, a message would also be found here.

- Click the down arrow on the scroll bar.

Multiple Regression Y1:Chol-3yrs 3 X variables

Confidence Intervals and Partial F Table					
Variable:	95% Lower:	95% Upper:	90% Lower:	90% Upper:	Partial F:
INTERCEPT					
Triglycerides	-.103	.159	-.081	.137	.183
HDL	-.231	1.467	-.089	1.325	2.171
LDL	.398	.989	.448	.94	22.574

- When you close Lipid Data do not save the changes made to the dataset in this example.

There is an additional statistic associated with the multiple regression model that is not associated with the simple regression model: the partial F-test. Not surprisingly, the partial F value is the square of the t-value associated with the coefficient in the beta coefficient table. It has (1, residual DF) degrees of freedom. If significant, this partial F-value identifies a variable that is a statistically significant given that all other independent variables have been included in the model. The partial F addresses the question of whether the addition of this specific independent variable, given the other variables in the regression model, significantly contributes to the predictable variance. When using StatView, it is convenient to check the significance associated with the t-value of the beta coefficient table to determine whether or not the partial F is significant. Only the LDL variable makes a significant contribution to the multiple regression equation.

Polynomial Regression

This section in the regression dialog box calculates a polynomial regression between a dependent Y and a polynomial independent X variable. Specify an order (degree) of the polynomial between 2 and 9. It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

Simple regression dealt with fitting straight lines. There are datasets for which straight lines regressions may be “supplemented” by considering more complex lines. One such model for fitting complex lines to a single independent and dependent dataset is the polynomial regression model sometimes referred to as *curvilinear regression*. A polynomial regression line is a curved regression line.

The first-order regression line predicts Y from X as a straight line. A second-degree polynomial predicts Y from X and X^2 and is a curve with a single point of inflection or bend. A third-degree polynomial predicts Y from X, X^2 , and X^3 and is a curve with two points of inflection or two bends. StatView allows you to specify models up to ninth-order polynomials.

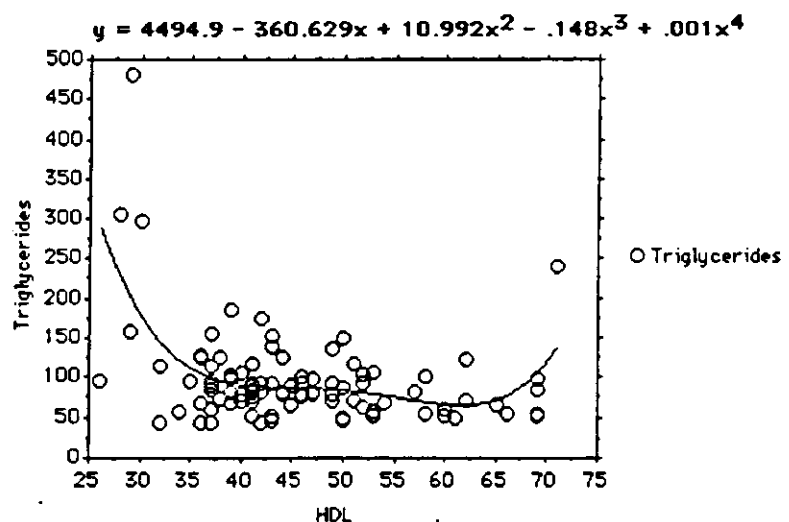
A major problem for polynomial regression is determining what order to use. Theoretically, the highest order regression for any set of data would be $k-1$, where k is the number of distinct values for the independent variable. However, parsimony compels us to find the lowest possible order required to describe the data. As the order of a polynomial model increases, the proportion of predictable variance always increases, but you must determine whether the increase is useful.

- Open Lipid Data.
- Assign Y to Triglycerides and X to HDL.
- Select Regression from the Compare menu. You see the Regression dialog box.
- Click Polynomial Regression, an order of 4 and click OK.

Note that we did not request that residual statistics, the Durbin-Watson test, be computed. Because the observations that are being used as an illustrative example are not in a sequential order, this residual test is inappropriate.

The ideal strategy to employ when using polynomial regression is to start with a high order polynomial equation and look at the scattergram. If there are no outliers, then look at the tables. Assuming that all outliers have been eliminated from the dataset, you systematically reduce the polynomial order until the regression sum of squares becomes significant and the probability for the highest order coefficient becomes significant. Some researchers investigate even lower order models. How high is high? A good rule of thumb, but none-the-less quite arbitrary, is order four unless you believe a higher order is more appropriate.

- Select Scattergram from the View menu.



This view of the scattergram clearly shows several outliers. In particular, one HDL value is 480 while the typical value is between 75 and 100. Most likely those values beyond 220 just don't fit with the other cluster of points and should not be included in the analysis. Polynomial regression is extremely sensitive to outliers.

- Activate the dataset window and click on the column heading for Triglycerides.
- Choose **Select Range** from the Tools menu. This will allow you to specify the range of values for Triglycerides to be included in the analysis. Enter 220 for the upper bound.
- Click **Done**. Now only Triglyceride values less than 220 will be included in the analysis. (Remember to clear this range when you are done with this example if you continue to use this sample dataset.)
- Select **Table** from the View menu.

Polynomial Regression X₁ : HDL Y₁ : Triglycerides

Count:	R:	R-squared:	Adj. R-squared:	RMS Residual:
91	.253	.064	.02	30.94

Analysis of Variance Table

Source	DF:	Sum Squares:	Mean Square:	F-test:
REGRESSION	4	5629.796	1407.449	1.47
RESIDUAL	86	82327.105	957.292	p = .2183
TOTAL	90	87956.901		

No Residual Statistics Computed

If there are missing values a note will be placed on the bottom of the page indicating the number of cases deleted because of missing values.

The initial summary table is identical to the summary table used with multiple regression. It is interpreted in an identical fashion to the multiple regression summary table. The fourth order polynomial regression equation does not predict a statistically significant proportion of the variance of the dependent variable since $F(4,86) = 1.47$ and $p = .2183$. There is no need to continue looking at the other tables. We should reduce the polynomial order to 3 and recompute the tables.

- Select **Regression** from the Compare menu.
- Click an order of 3 and click **OK**.

StatView automatically recomputes the polynomial regression with the new order.

Polynomial Regression X₁ : HDL Y₁ : Triglycerides

Count:	R:	R-squared:	Adj. R-squared:	RMS Residual:
91	.243	.059	.026	30.846

Analysis of Variance Table

Source	DF:	Sum Squares:	Mean Square:	F-test:
REGRESSION	3	5179.033	1726.344	1.814
RESIDUAL	87	82777.868	951.47	p = .1505
TOTAL	90	87956.901		

No Residual Statistics Computed

The third order polynomial regression equation does not predict a statistically significant proportion of the variance of the dependent variable, since $F(3,87)=1.814$ and $p=.1505$. There is no need to continue looking at the other tables. We should reduce the polynomial order to 2 and recompute the tables.

- Select Regression from the Compare menu.
- Click an order of 2.
- Click OK.

Once again, StatView automatically recomputes the polynomial regression with the new order.

Polynomial Regression X ₁ : HDL Y ₁ : Triglycerides				
Count:	R:	R-squared:	Adj. R-squared:	RMS Residual:
91	.243	.059	.037	30.671

Analysis of Variance Table				
Source	DF:	Sum Squares:	Mean Square:	F-test:
REGRESSION	2	5174.252	2587.126	2.75
RESIDUAL	88	82782.649	940.712	p = .0694
TOTAL	90	87956.901		

No Residual Statistics Computed

Again we see a lack of fit. The second order polynomial regression equation does not predict a statistically significant proportion of the variance of the dependent variable since $F(2,88)=2.75$ and $p=.0694$. There is no need to continue looking at the other tables. The curvilinear model, polynomial regression, does not fit the data. The only model left is the simple regression model, order one.

- Select Regression from the Compare menu.
- Click Simple Regression.
- Click OK.

The table view contains:

Simple Regression X ₁ : HDL Y ₁ : Triglycerides				
Count:	R:	R-squared:	Adj. R-squared:	RMS Residual:
91	.243	.059	.048	30.498

Analysis of Variance Table				
Source	DF:	Sum Squares:	Mean Square:	F-test:
REGRESSION	1	5173.846	5173.846	5.562
RESIDUAL	89	82783.056	930.147	p = .0205
TOTAL	90	87956.901		

No Residual Statistics Computed

The first order, linear, regression equation does predict a statistically significant proportion of the variance of the dependent variable, assuming a critical probability level of .05, since $F(1,89)=5.562$ and $p=.0205$. If the additional table is reviewed it will be apparent that the regression coefficient is significant.

- Click the down arrow on the scroll bar.

Simple Regression X₁ : HDL Y₁ : Triglycerides

Beta Coefficient Table

Variable:	Coefficient:	Std. Err.:	Std. Coeff.:	t-Value:	Probability:
INTERCEPT	123.44				
SLOPE	-.799	.339	-.243	2.358	.0205

Confidence Intervals Table

Variable:	95% Lower:	95% Upper:	90% Lower:	90% Upper:
MEAN (X,Y)	80.614	93.32	81.653	92.281
SLOPE	-1.472	-.126	-1.362	-.236

As expected, the regression coefficient is significant. When describing the Triglycerides given HDL, the linear regression model is superior to the polynomial regression model. The linear regression equation for the i th fitted value is: $Y_i = -1.472X_i + 80.614$.

Unfortunately, we do not have a single set of variables in our illustrative datasets for which polynomial regression provides an adequate fit. However, for convenience of discussion let us assume that the third order model fit our data.

- Select Regression from the Compare menu.
- Click Polynomial Regression.
- Click an order of 3.
- Click OK.

The first table has already been reported and discussed. Let us assume that the F-ratio is significant.

- Click the down arrow on the scroll bar.

Polynomial Regression X₁ : HDL Y₁ : Triglycerides

Beta Coefficient Table

Variable:	Coefficient:	Std. Err.:	Std. Coeff.:	t-Value:	Probability:
INTERCEPT	142.506				
x	-2.022	16.666	-.614	.121	.9037
x ²	.025	.35	.765	.072	.9425
x ³	-1.691E-4	.002	-.398	.071	.9436

If the order of the polynomial is higher than 7, the table is continued on the next page. If you chose to predict values, there will be a note on the bottom of the page that gives the name of the column which contains the predicted values.

Following the prescribed strategy, we would expect all three regression coefficients (X , X^2 , X^3) to be significantly different from 0, $p < .05$. This information would be determined from the probability column. Like the multiple regression table, the square of the t-ratio is the partial F-ratio. Thus, we know that a statistically significant regression weight will be associated with a statistically significant partial F-ratio. From the above table the third order polynomial regression equation for fitting the i th Triglyceride value from the i th HDL value is:

$$Y_i = 142.506 - 2.022X + .025X^2 - .0001691X^3$$

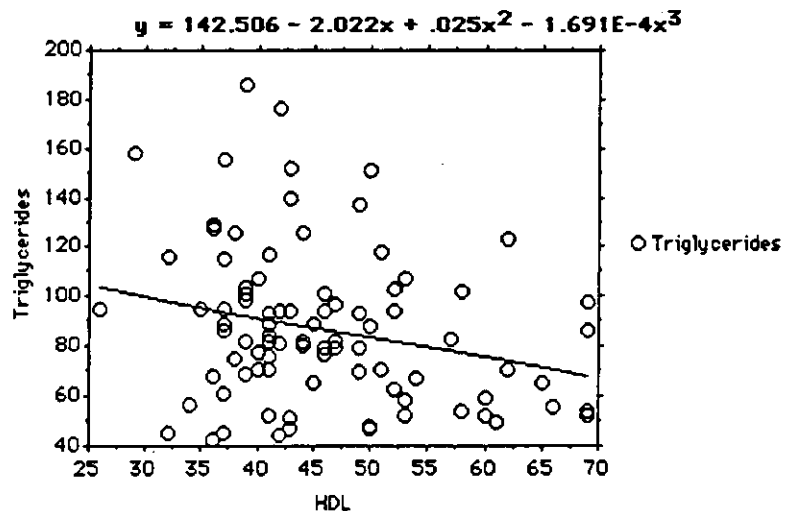
- Click the down arrow on the scroll bar.

Polynomial Regression X_1 : HDL Y_1 : Triglycerides

Confidence Intervals and Partial F Table					
Variable:	95% Lower:	95% Upper:	90% Lower:	90% Upper:	Partial F:
INTERCEPT					
x	-35.148	31.105	-29.731	25.687	.015
x ²	-.67	.721	-.556	.607	.005
x ³	-.005	.005	-.004	.004	.005

This table provides the confidence intervals for the regression weights reported in the previous table. It also reports the partial F-ratios. Those polynomials associated with a significant partial F-ratio are making a statistically significant contribution to the predictable variance. Most of the information in this table is derived from the previous table.

- Select Scattergram from the View menu.



This view of the scattergram should clearly show the curvilinear nature of the third order polynomial regression line. Because of the actual poor fit of the polynomial regression model to the data and because the linear model is most appropriate, the third order polynomial regression line appears to be almost linear.

We have provided a strategy for interpreting a polynomial regression model. We feel that this is the most effective strategy for interpretation. There are other strategies, but we have found them to be dangerously misleading with our illustrative datasets. We have also discovered that we have no data that adequately fit a polynomial model. Such data are difficult to find. Data that most usually conform to the polynomial model are time series data.

Stepwise Regression

When you consider the multiple regression model, it is important to keep in mind that a frugal solution is frequently desirable. We want the most efficient regression equation with the smallest number of variables. We want to be sure that every variable in the multiple regression equation makes a statistically significant contribution to the predictable variance. Most importantly, we want to predict as much of the variance of the dependent variable as is possible from the composite of independent variables. This whole process is complicated by the fact that the independent variables may be correlated with each other and consequently each predict a "same" part of the variation in the dependent variable.

StatView computes a multiple linear regression using the forward stepwise regression with elimination of unnecessary variables. The forward selection procedure selects as the next variable for the regression model that independent variable with the highest partial correlation with the dependent variable. Essentially, the partial F-ratio associated with each remaining variable is computed based upon the inclusion of a remaining variable into the existing equation. Of those variables not included in the regression equation, that variable with the largest partial F-ratio is selected for inclusion and then new partial F-ratios are computed.

The stepwise procedure used by StatView improves this forward selection by including an additional evaluative step in the procedure. With the inclusion of each new variable in the model, all variables previously entered are reevaluated. This reevaluation will remove a variable from the model if the variable's variance contributions are accounted for by variables subsequently entered into the model, that is, if a variable's partial F-ratio becomes less than some predetermined value it is removed from the model.

You cannot leave the intercept out of the model, that is, you cannot force the intercept through the origin. If you have tests where this is required, you may be interested in Abacus Concept's companion product, SuperANOVA.

The StatView stepwise procedure continues until no variables currently in the equation can be removed and the variable with the highest partial correlation not in the equation fails the F-to-Enter test.

At each step, the results you have seen in the regression examples above are displayed. The following values can be computed and added to the data window:

residuals

standardized residuals

fitted values

predicted values

- Select Stepwise Regression from the Compare menu. The following dialog box is displayed:

Stepwise Regression Parameters:

F-to-Enter F-to-Remove

Force variables into the regression?:
☒ No ☐ Yes, force **K** variables to

Add as a new column to the data window:
☐ Residuals
☐ Standardized Residuals
☐ Fitted Values
☐ Predicted Values

Create new column(s) using values from:
☒ included rows ☐ all rows

Compute residual statistics: ☒ No ☐ Yes

Text entry rectangles at the top of the dialog box allow you to specify the F-to-Enter and the F-to-Remove values which control the entry and removal of variables into the equation. The value 4 is the default for F-to-Enter. The value 3.996 is the default for F-to-Remove. The F-to-Remove value must be less than or equal to the F-to-Enter value.

Below these, radio buttons allow you to force variables into the equation. If you select Yes, then you must specify the sequence of variables to be forced. These X variables enter the equation first and stay regardless of their F-ratios. Variables are specified in the following manner: to force variables X₃ through X₆, click Yes, **force variables** and enter 3 to 6. These numbers refer to the X variables' subscripts. First variables are entered in the order of their subscripts, with X₃ entering first.

You can add the following information as new columns to the data window:

Option	Description
Residuals	The difference between the observed Y_i and the fitted \hat{Y}_i .
Standardized residuals	The residuals in unit normal deviate form.
Fitted values	The fitted \hat{Y}_i values determined by the regression model.
Predicted Values	\hat{Y}_i values predicted for additional independent values. To compute predicted values, enter the independent values after the last complete case for the model, leaving missing values in the dependent column. That is after the last case where both X and Y values are present. The program calculates predicted values using the determined regression equation.

Selections made using these check boxes are added as new columns to the end of the dataset. If a case has been deleted from the model, the new column contains a

missing value in that row. The rows containing the X-Y pairs used in the regression analysis contain missing values in the Predicted Values column.

Radio buttons allow you to specify whether the above values are determined for just the included rows of the dataset or for all rows. If you select **Include Rows**, the values are calculated for just the included row of the dataset; excluded rows contain missing values. If you select **All Rows**, the values are calculated for all rows in the dataset regardless of their included or excluded state. In this manner, you may fit a model for one group of data and use this model to predict for all groups.

You may compute residual statistics without adding them as a new column. StatView automatically selects this if you add any new columns to the dataset.

Assign a dependent Y variable, Y₁, and one or more independent X variables. It is a ManyXOneY statistic. Only one stepwise regression can be calculated at a time. You can not assign two Y variables.

The illustrative dataset, Lipid Data, has a collection of variables that deal with blood and blood flow. Is there some combination of these variables that predict Cholesterol count? It was demonstrated in the correlation section that LDL is highly correlated with Cholesterol count, $r=.96$. If we were to use LDL with the other variables in a multiple regression analysis the analysis would not tell us too much about the variables and Cholesterol because LDL would predict most of the variation of Cholesterol. However, if we were to establish a multiple regression equation from the variables, excluding LDL, would there be some combination of these remaining variables that predict a statistically significant ($p<.05$) portion of the Cholesterol variation? Stepwise regression will address this question.

- Open Lipid Data. Assign Y to Cholesterol.
- Assign X to Triglycerides, HDL, Weight, Systolic BP and Diastolic BP, in that order.
- Select Stepwise Regression from the Compare menu.
- Enter a value of 3.953 for F-to-Enter. This is the critical F-ratio, for 1 and 93 degrees of freedom, that must be exceeded for significance if alpha is set equal to .05. Any variable with a partial F-ratio of 3.953 or greater will be entered into the regression equation.
- Enter a value of 3.953 for F-to-Remove. Any variable with a partial F-ratio less than 3.953 will either be removed from the regression equation if it has been previously entered with a larger F-ratio or it will not be entered at all.
- Click OK. The following summary appears:

Stepwise Regression Y₁:Cholesterol 5 X variables

Summary Information

F to Enter	3.953
F to Remove	3.953
Number of Steps	2
Variables Entered	2
Variables Forced	0...0

No Residual Statistics Computed

When the stepwise regression has been completed, this summary page appears. This summary information notes the dependent variable name and the number of independent variables from the dataset that were used in the analysis. The F-to-enter and F-to-remove values used in the procedure are displayed as well as the number of steps and the number of variables entered during the procedure. If any variables were forced to enter the regression, their variable subscripts would be shown. A message states that no residual statistics were calculated. If residual statistics had been specified, the Durbin-Watson test results would appear there.

- Click the down arrow.

Stepwise Regression Y₁:Cholesterol 5 X variables

STEP NO. 1 VARIABLE ENTERED: X₁: Triglycerides

R:	R-squared:	Adj. R-squared:	RMS Residual:
.401	.161	.152	32.859

Analysis of Variance Table				
Source	DF:	Sum Squares:	Mean Square:	F-test:
REGRESSION	1	19218.259	19218.259	17.8
RESIDUAL	93	100410.646	1079.684	
TOTAL	94	119628.905		

The title gives the current step and the name of the variable entered/removed at this step. These tables provide the same information as the tables shown earlier in this chapter under the discussion of Regression.

The first step identified Triglycerides as the best single predictor of Cholesterol. In the second table, the value of R is the multiple correlation and represents the correlation between the observed dependent variable and the fitted scores predicted using Triglycerides as the only variable in the multiple regression equation. For this example the correlation between the fitted Cholesterol, predicted from Triglycerides, and the observed Cholesterol is .401. The proportion of Cholesterol variance that can be predicted by this single variable regression equation is .161, with the unbiased estimate of this value being .152. The root mean square residual, standard deviation of the residuals, is 32.859.

The ANOVA table is interpreted in a fashion that is identical to that discussed with simple and multiple regression. The F-test, also the F-value, is substantially larger than the critical F-to-Enter thereby implying that the proportion of predictable variance, regression mean square, is greater than you would expect by chance.

- Click the down arrow.

STEP NO. 1 Stepwise Regression Y₁:Cholesterol 5 X variables

Variables in Equation				
Variable:	Coefficient:	Std. Err.:	Std. Coeff.:	F to Remove:
INTERCEPT	168.413			
Triglycerides	.235	.056	.401	17.8

Variables Not in Equation		
Variable:	Par. Corr:	F to Enter:
HDL	.527	35.441
Weight	-.071	.471
Systolic BP	-.11	1.118
Diastolic BP	.114	1.209

This page contains summary information regarding the regression equation based upon the selected variable and the partial F-ratios for those variables not selected.

If any of these variables were forced into the equation a bullet (•) appears by the parameter name in the table.

The above table is interpreted in a fashion similar to the interpretation that was used for the multiple regression example. From the above table the regression equation for predicting an initial Cholesterol count for the i th individual, Y_i , using the Triglycerides value, X_2 , may be defined as:

$$Y_i = 168.413 + .235 X_2$$

Of course the partial F-ratio of 17.8 is significant, $p < .05$, since it exceeds 3.953. Note that this F-ratio is labeled F-to-Remove since it determines whether the variable remains in the equation. The partial F-ratios have been computed for the other independent variables not included in the equation. Their corresponding partial correlations are also reported. The partial correlation for an independent variable represents the correlation between the independent variable and the dependent variable, removing the effects of the independent variable(s) included in the regression equation. Notice that the partial correlations are always proportional to the partial F-ratios.

The next variable to be selected for inclusion in the regression equation will be the variable with the largest significant, exceeding 3.953, partial F-ratio. This will be the same variable that has the largest significant partial correlation.

Tables like the above two tables are determined for each variable selected for inclusion, or removal from, the regression equation. For our example there are only two variables included and none removed from the regression equation. The tables that follow are the tables associated with the last variable included in the equation. They are structurally similar to the previous table.

- Click the down arrow.

Stepwise Regression Y_1 :Cholesterol 5 X variables

(Last Step) STEP NO. 2 VARIABLE ENTERED: X_2 : HDL

R:	R-squared:	Adj. R-squared:	RMS Residual:
.628	.394	.381	28.07

Analysis of Variance Table				
Source	DF:	Sum Squares:	Mean Square:	F-test:
REGRESSION	2	47141.995	23570.997	29.916
RESIDUAL	92	72486.911	787.901	
TOTAL	94	119628.905		

This second step identified HDL as the best predictor to be used with Cholesterol to define a two variable multiple regression equation. In the above table, the multiple correlation between the observed dependent variable and the fitted scores is .628. The fitted scores were determined predicted by using HDL and Triglycerides as the independent variables in the multiple regression equation. The proportion of Cholesterol variance that can be predicted by this new two variable multiple regression equation is .394, with the unbiased estimate of this value being .381. The root mean square residual, standard deviation of the residuals, is 28.07.

Notice that the regression degrees of freedom are 2 in the ANOVA table. We can no longer make a simple comparison between the F-ratio and the F-to-Remove. The

regression sum of squares is now determined as a function of two independent variables. The regression equation is accounting for more variance because it has more variables in it. The predictable variance is significantly greater than 0 ($p < .05$). Furthermore, the ratio of the regression sum of squares to the total sum of squares is still R^2 . So you might say that the proportion of predictable variance is greater than 0.

- Click the down arrow.

STEP NO. 2 Stepwise Regression Y_1 :Cholesterol 5 X variables

Variables in Equation				
Variable:	Coefficient:	Std. Err.:	Std. Coeff.:	F to Remove:
INTERCEPT	79.735			
Triglycerides	.317	.049	.541	40.97
HDL	1.778	.299	.503	35.441

Variables Not in Equation		
Variable:	Par. Corr.:	F to Enter:
Weight	.078	.561
Systolic BP	-.116	1.246
Diastolic BP	.093	.793

From the above table the regression equation for predicting an initial Cholesterol count for the i th individual, Y_i , using the Triglycerides value, X_2 , and HDL, X_3 , may be defined as:

$$Y_i = 79.735 + .317 X_2 + 1.778 X_3$$

Notice that the regression weight associated with the Triglycerides has changed with the inclusion of the HDL variable, as has the intercept.

The new partial F-ratios have been computed for the other independent variables not included in the equation, as have their partial correlations. This partial correlation represents the correlation between the independent variable and the dependent variable, removing the effects of the two independent variables included in the regression equation.

None of the remaining partial F-ratios exceed 3.953. Therefore, none of the remaining variables will be included in the regression equation and the computations are complete.

Note that comparisons of R^2 values shows that Triglycerides and HDL, together do not predict Cholesterol as well as the single variable LDL.

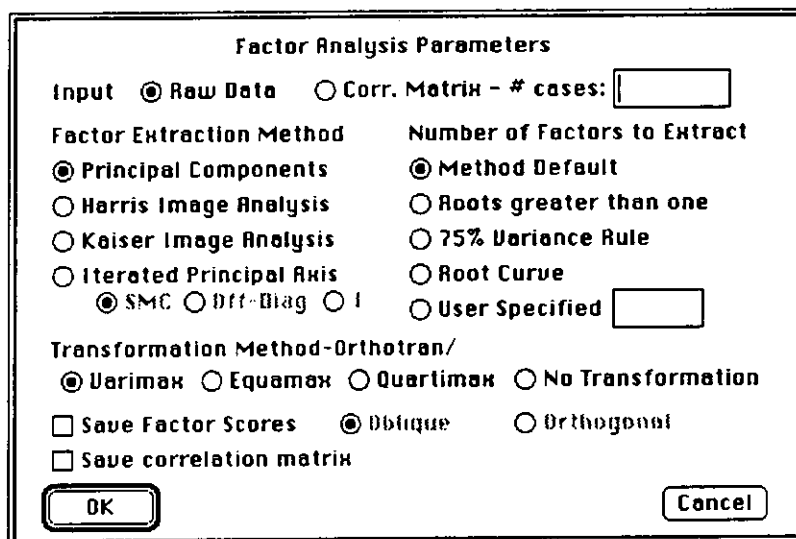
Factor Analysis

This menu choice performs a factor analysis of a correlation matrix. The input can be raw data or a correlation matrix. Select the initial factor extraction method from the dialog box. The choices are principal components, Kaiser image, Harris image, and iterated principal axis analysis. Specify either no transformation for the solution or one of: equamax, varimax, and quartimax. You also can specify the criteria that are used for determining the number of factors. Factor scores can be computed and saved. Plots of the unrotated solution, the orthogonal solution, and the oblique primary are provided.

Parameters

The discussion below is a brief clarification of the factor analysis tables produced. With minor exceptions, most of the discussion is contained in general form in one of the following three reference books: *Factor Analysis* by R. Gorsuch (1984), *Multivariate Analysis With Applications in Education and Psychology* by N. Timm (1975), and *The Foundations of Factor Analysis* by S. Mulaik (1972).

- Select **Factor Analysis** from the **Compare** menu. The following dialog box is displayed:



The dialog box is titled "Factor Analysis Parameters". It contains several sections of controls:

- Input:** Two radio buttons: ☒ Raw Data and ☐ Corr. Matrix - # cases: [text box].
- Factor Extraction Method:** Four radio buttons: ☒ Principal Components, ☐ Harris Image Analysis, ☐ Kaiser Image Analysis, and ☐ Iterated Principal Axis. Below the last one are three radio buttons: ☒ SMC, ☐ Off-Diag, and ☐ I.
- Number of Factors to Extract:** Four radio buttons: ☒ Method Default, ☐ Roots greater than one, ☐ 75% Variance Rule, and ☐ Root Curve. Below these is ☐ User Specified [text box].
- Transformation Method-Orthotran/:** Four radio buttons: ☒ Varimax, ☐ Equamax, ☐ Quartimax, and ☐ No Transformation.
- Save Factor Scores:** ☐ Save Factor Scores, ☒ Oblique, and ☐ Orthogonal.
- Save correlation matrix:** ☐ Save correlation matrix.
- Buttons:** "OK" and "Cancel" at the bottom.

Input Data

There are two types of data that can be analyzed: raw data and correlation matrix data.

- Raw data is subject by case data (subjects being X variable columns and cases being rows). For such data, it is desirable that there be more observations than subjects. If you wish to compute factor scores, the raw data input must be selected.
- Correlation matrix data requires the input data (the dataset) be a Pearson correlation matrix. The correlations must not be determined from different samples of subjects; they must be determined from the same total pool of subjects.

StatView expects the correlation values to be located in the lower left corner of the correlation matrix. Thus, you may use either a square correlation matrix or a lower left correlation matrix as input. If **Correlation Matrix** is selected in the dialog box, you must enter the **Number of Cases** used to determine the correlation matrix. This number is used for any multivariate significance tests performed on the data.

Factor Extraction Method

The factor extraction method, also referred to as the *initial factoring procedure*, determines the magnitudes of the communality estimates of the variables. These influence the magnitudes of the eigenvalues which ultimately influence decisions regarding the number of factors to extract. Eigenvalues are also referred to as *characteristic roots*. There are four factor extraction methods available: principal

components analysis; Harris image analysis; Kaiser image analysis; and iterated principal axis.

The principal components analysis is probably the oldest factoring procedure. It performs a simple eigenvalue-eigenvector analysis of the correlation matrix in its original form. This well established factor method is utilized in a variety of disciplines and discussed in the applied research texts of these disciplines.

Harris image analysis, a more recent factoring procedure, requires a nonsingular correlation matrix. It has a tendency to extract more factors than either of the other two non-image analysis factor extraction methods. It is properly referred to as a psychometric factor procedure, meaning that the model is based on a theory of variable sampling as opposed to the more traditional theory of statistical or subject sampling. This method factors a modification of the original correlation matrix, the image variance covariance matrix. Because of the large number of factors that define an image factor solution, the final rotated solution usually has a large number of zero loadings. However, the non-zero loadings are not always as large in magnitude as those large loadings observed in the more traditional factor analytic model. Furthermore, the Harris image analysis will determine factors that are defined by a single variable, pseudo-specific factors.

Kaiser image analysis is a factor model that was defined by Kaiser in a manuscript devoted to a clarification of the Harris image analysis. It is distinct from the Harris image analysis in that it rescales the factor solution so it represents a factor analysis of the original correlation matrix and allows the user to impose the interpretive model of traditional factor analysis. Both image analyses define the same number of factors. While this is a trivial algebraic modification, the final, transformed Kaiser image analysis solution will have loadings that are quite different from those loadings of the Harris image analysis solution. (Factor loadings are reported by StatView in the primary pattern matrix described later in this chapter.)

The iterated principal axis is an iterative factor extraction method. This method requires that some estimate of the communalities be placed in the diagonal of the correlation matrix prior to the initial factoring. The communality estimates are modified with each iteration. The factoring continues until the communality estimates stabilize. This factoring method requires significantly more computation time than the other factoring methods. If you select this method of factor extraction, you are also required to select a method for determining the initial communality estimates. In the dialog box, selecting SMC causes the squared multiple correlations to be the initial estimates. Selecting Off-Diag causes the largest off-diagonal entries of the correlation matrix to be the initial estimates. Selecting 1 causes 1 to be the initial estimate. The distinction between these initial communality estimates is quite easy to follow and is discussed in a number of factor analytic texts, including Gorsuch. If you selected SMC and your correlation matrix is singular, the largest off-diagonal entries of the correlation matrix is used as the initial communality estimates.

Number of Factors to Extract

The initial factoring method determines the properties of the initial factors. The number of factors retained for interpretation is almost always a function of the eigenvalues (sometimes referred to as *latent roots* or just *roots*). There are three general criteria that are used for determining the number of factors: roots greater than 1, root curve analysis, and extraction of 75% of the variance.

The criterion of Roots greater than 1 specifies that as many factors will be retained as there are eigenvalues greater than or equal to 1.

The criterion of 75% variance rule is determined by the sum of *all* eigenvalues, which is also the matrix variance. Keeping in mind that the eigenvalues are determined in order of descending magnitude, it becomes clear that each eigenvalue accounts for successively less variance than the eigenvalue preceding it. As soon as the sum of the proportionate contributions of the eigenvalues exceeds .75, it is assumed that all relevant matrix variance has been accounted for. The rank order of the eigenvalue that pushes the sum of the proportionate contributions over .75 is assumed to be an index of the total number of factors to be retained for further analysis.

The root curve criterion is based upon the work of Cattell (1966) and Cattell and Jaspers (1967). Essentially, this criterion determines the eigenvalue associated with the point of inflection of a plot of the eigenvalues from largest to smallest. Inasmuch as the eigenvalues are always determined in order from largest to smallest, the rank order of the eigenvalue associated with the point of inflection is considered to be an estimate of the number of factors. It is frequently employed with interactive factor analyses.

Method default represents a criterion that is unique to the factoring method employed, and may or may not determine the same number of factors as one of the first three criteria. The default method for principal components is a combination of two criteria. It is the larger of the numbers determined by either the 75% variance rule or by the root curve analysis.

The default method for the two image analysis models follows Harris (1962), and defines the number of factors by a mathematically precise method that works only for the image analysis model: Harris eigenvalues greater than 1. Harris eigenvalues are the eigenvalues of the image variance-covariance matrix. If you decide to apply one of the three criteria discussed above in place of the method default, the criterion selected is applied to a modification of the Harris eigenvalues. If you attempt to enter a specified number of factors, and this specified number of factors is greater than the number that might be determined by the image analysis method default, then the specified number is over-ridden by the number determined by the method default.

The iterated principal axis method default is simply the number of eigenvalues greater than 1.

You can specify the number of factors to extract. If you have decided to specify the number of factors prior to the analysis, you should be aware of many salient points. Traditionally, the maximum number of factors that you might expect in a factor analysis is half the number of variables being analyzed. Under no circumstances should the estimated number be larger than the number of variables. If, for some reason, you over-estimate the number of factors, the estimate is adjusted to the maximum possible number for the given matrix. This may be less than the number of variables, especially if the number of subjects defining the correlations is less than the number of variables.

Transformation Method

The factor method and the number of factors represents the first part of factor analysis. The transformation method represents the second part of factor analysis. If you decide that you want no transformation, then the initial factor solution is considered to be the final solution matrix.

If you want to define the final solution by some transformation solution, you have the option of three types of orthogonal solutions: varimax, equamax, and quartimax. Regardless of your choice of orthogonal solution, an oblique solution,

the orthotran solution, is computed from your orthogonal solution. All three orthogonal solutions attempt to define a simple structure solution with the constraint that the factors remain orthogonal or uncorrelated. The orthotran solution relaxes the constraint of orthogonality, tolerating correlated factors, and determines a clearer simple structure than the orthogonal solutions.

All three orthogonal solutions are computed by maximization of the orthomax criterion. StatView computes normalized solutions. If a quartimax solution is selected, a majority of the variance is allocated to the first factor with the consequence that the other factors do not have a good simple structure. When an equamax solution is selected, the variance is allocated equally to all factors. When a varimax solution is selected, a solution between the equamax solution and quartimax solution is obtained. Typically, the varimax criterion is maximized unless you have a compelling reason for maximizing either of the other two criteria.

The orthotran solution is a general transformation solution that refines the simple structure of an orthogonal solution. It refines the simple structure by allowing the factors to be correlated, an oblique solution. If correlated factors do not refine the simple structure, then the solution matrix remains an orthogonal solution. If a better simple structure is obtained from an oblique solution, then the solution matrices are the primary pattern and the primary intercorrelation matrix or the reference structure solution and the primary intercorrelations. Whenever an orthogonal solution is selected, StatView determines the associated orthotran solution as well.

Save Factor Scores

The third part of factor analysis deals with the factor scores. If you have a non-singular correlation matrix, it is possible to compute regression estimate factor score weights. Checking **Save Factor Scores** causes StatView to compute and save the factor score weights. The factor scores are added as new columns to the end of the dataset. This option is available only if raw data has been input.

If you did not determine a transformation solution, you obtain *unrotated* factor scores. If you computed a transformation solution, select whether you wish to have orthogonal or oblique factor scores. If you select orthogonal, the factor scores show 0 intercorrelations. If you select oblique, the factor scores are correlated. More precisely, the intercorrelation of the primary factors represents the intercorrelations you would obtain if you were to actually compute the intercorrelations of the factor scores.

Save Correlation Matrix

If you check **Save correlation matrix**, the computed correlation matrix is saved to a new dataset window. It is saved as a square correlation matrix.

Tabular Views

This is a ManyX statistic. (See the table at the beginning of this chapter.) If you are inputting raw data, assign X to each variable to be used in determining the correlation matrix. If you are inputting a correlation matrix, assign X to each column of the correlation matrix.

- Open the **Eight Physical Variables** dataset. (Notice that this dataset is a correlation matrix.)

- Assign X to each column of the dataset.
- Select **Factor Analysis** from the **Compare** menu. We are inputting a correlation matrix.
- Click **Correlation matrix**, and enter 305 as the number of cases.
- Click **OK**, using the default settings; **Principal Components**, **Method Default**, and **Varimax** for the analysis.
- Page through the view as you follow the explanation below.

The Factor Analysis displays the progress of the calculation. This is useful to track the progress of longer calculations.

Factor Analysis for physical variables: X1 ... X8

Summary Information

Factor Procedure	Principal Component Analysis
Extraction Rule	Method Default
Transformation Method	Orthotran/Varimax
Number of Factors	2

When the factor analysis has been completed, this summary page appears. This page notes the dataset name and the number of variables from the dataset that were used in the analysis. The factor procedure is noted along with the procedure used to determine the number of factors, the transformation procedure, and the actual number of factors defined. If factor scores were computed and saved, the columns they were saved in are noted beneath the summary table. Also, if the correlation matrix is either singular or ill-conditioned, such is noted underneath the table.

Correlation matrix

	height	arm span	forearm ...	lower le...	weight	bitrocha...	chest girth	chest wi...
height	1							
arm span	.846	1						
forearm l...	.805	.881	1					
lower leg...	.859	.826	.801	1				
weight	.473	.376	.38	.436	1			
bitrochan...	.398	.326	.319	.329	.762	1		
chest girth	.301	.277	.237	.327	.73	.583	1	
chest wid...	.382	.415	.345	.365	.629	.577	.539	1

The correlation matrix table is fundamental to factor analysis. The correlation matrix is the variance-covariance matrix of the variables in a standard score format. The ij th entry of the correlation matrix (row i and column j) is the correlation between variable i and variable j . The correlation matrix is printed in triangular form because half of the correlation coefficients of a correlation matrix duplicate the other half of the correlation coefficients. The correlation between variable i and variable j is exactly the same as the correlation between variable j and variable i . Virtually all books on factor analysis and multivariate analysis discuss the correlation matrix.

Partials in off-diagonals and Squared Multiple R in diagonal

	height	arm span	forearm ...	lower le...	weight	bitrocha...	chest girth	chest wi...
height	.816							
arm span	.346	.849						
forearm L...	.072	.584	.801					
lower leg...	.479	.179	.188	.788				
weight	.183	-.196	.1	.056	.749			
bitrochan...	.103	-.005	.027	-.122	.492	.604		
chest girth	-.146	.091	-.116	.131	.491	.054	.562	
chest wid...	-.086	.248	-.087	-.025	.238	.177	.12	.478

Guttman (1954) provided ample algebraic evidence that for a composite of variables that are going to be factor analyzed, the partial correlations between the variables should approach 0. Furthermore, he argued, the multiple correlations for the variables should be reasonably high. The partial correlation between any two variables, say height and lower leg in our example, is the correlation (.479) that exists between the two variables, removing the effects of the other variables in the matrix. That is, in the eight physical variables, the partial correlation between height and lower leg is an estimate of the correlation between the two variables based upon variation that is common to the two variables, but not common with any other variables in the matrix.

The square of this partial correlation (.229) represents the proportion of variance of either variable that could be predicted in a linear regression sense only by the other variable and not by any other variable in the matrix. Alternatively, the squared multiple correlation for a variable represents the proportion of variance for that variable that is common with all other variables in the matrix. The squared multiple correlation for height suggests that approximately 82 percent of the variation in height may be predicted in a linear regression sense from the other seven variables.

Measures of Variable Sampling Adequacy

Total matrix sampling adequacy : .845

height	.864
arm span	.816
forearm length	.858
lower leg len...	.887
weight	.78
bitrochanteri...	.851
chest girth	.824
chest width	.898

Bartlett Test of Sphericity- DF: 35 Chi Square: 2116.975 P: .0001

Addressing Guttman's (1954) expectation of 0 partial correlations and large multiple correlations, Kaiser (1970) developed an index, called *variable sampling adequacy*, of the extent to which a matrix of partial and multiple correlations conforms to 0 partials and large multiple correlations. The notion of variable sampling adequacy follows logically from Guttman's discussion of partial and multiple correlations. Specifically, to the extent that a composite of variables is logically homogeneous (measuring the same universe of content) they are especially appropriate for factor analysis. The measure of variable sampling adequacy reported in StatView is derived from Kaiser's (1970) equations. This index quantifies the extent to which a composite of variables, and the variables within the composite, conform to the desired expectation of the partial correlations tending toward 0.

Kaiser argues that the sampling efficiency for the total composite of variables, total matrix sampling adequacy (or MSA), should be greater than .500 in order to assume that Guttman's assumptions have been minimally met. For the eight physical variables, the index is .845, which suggests that these data do indeed represent a homogeneous collection of variables and are suitable for factor analysis. As the index approaches 1, you may assume that the data are conforming almost perfectly to the assumption of 0 partial correlations.

It is quite possible that one or more variables are different from the other variables in the composite and from each other. In such a case, the total MSA is depressed, perhaps even appearing to be less than .50. Such variables each have a low index of variable MSA because they may not logically belong to the same psychometric universe of content as the other variables in the composite of variables. They will have an unpredictable influence on any factor analysis done on the composite. Eliminating those variables with low indices of sampling efficiency will result in an improved index of total MSA, and a composite of variables that are perhaps more appropriate for factor analysis. For our example, the lowest index of variable MSA is .78 for weight. Even this lowest value is substantially larger than the minimum value of .50, thereby reaffirming that these variables are appropriate for a factor analysis.

The measure of sampling adequacy is one of two evaluations that should be performed prior to any attempts to interpret the results of a factor analysis. The second evaluation is associated with the statistical significance of the correlations. It has been demonstrated that interpretable factors may emerge from data that are totally random. Bartlett's test of sphericity is the multivariate analog of the statistical test that is frequently applied to a single correlation coefficient to see if it is significantly different from 0. The test of sphericity is used to determine if, in general, the collection of correlations in the correlation matrix are different from 0. Ideally, a significant chi-square value is determined, thereby suggesting that the collection of correlations are different from 0 and most likely do not occur as a function of chance. For our illustrative data, the chi-square value is 2116, which is significant at the .0001 level. These values are reported at the bottom of the summary of the MSA values. Thus, our correlations are, in general, significantly different from 0 correlations.

Eigenvalues and Proportion of Original Variance

	Magnitude	Variance Prop.
Value 1	4.673	.584
Value 2	1.771	.221
Value 3	.481	.06
Value 4	.421	.053

StatView uses a powerful algorithm to determine eigenvalues: the HOW method. This method is based on the works of Householder, Ortega and Wilkinson. The method is discussed in great detail by Wilkinson (1965). Many properties of eigenvalues are beyond the scope of this discussion. Virtually all subjective criteria used to determine the number of factors can be applied to the table above. When looking at this table of eigenvalues, it may be noted that the eigenvalues are presented in an order that corresponds to their size. Typically, there are as many eigenvalues as there are variables, and the sum of the eigenvalues is equal to the sum of the diagonal elements of the matrix from which they are determined. The variance proportion is an estimate of the proportion of variance that the eigenvalue and its associated eigenvector account for when they are used to define a factor.

Usually, StatView divides the number of variables by two to determine an initial estimate of the number of eigenvalues (which is also an initial estimate of the number of factors). The many rules for determining the number of final factors are then applied to the eigenvalues that have been determined (see dialog box discussion on number of factors above). You may override the number of eigenvalues determined initially by equating the number of factors in the dialog box to the number of desired eigenvalues. The eigenvalues as displayed by StatView are of no great value in the final interpretation of the factor solution. They are displayed for purposes of completeness and for those who wish to address subjectively the number-of-factors question.

Eigenvectors

	Vector 1	Vector 2	Vector 3	Vector 4
height	-.398	.28	-.101	-.107
arm span	-.389	.331	.113	.068
forearm leng...	-.376	.345	.015	-.047
lower leg len...	-.388	.297	-.145	.124
weight	-.351	-.394	-.213	-.114
bitrochanter...	-.312	-.401	-.073	-.713
chest girth	-.286	-.436	-.421	.63
chest width	-.31	-.314	.853	.221

Like the eigenvalues, the *eigenvectors* are included in the display for purposes of completeness. The eigenvectors are computed with the eigenvalues. For every eigenvalue, there is an associated eigenvector. The eigenvectors are not used when interpreting the final factor solution.

Unrotated Factor Matrix

	Factor 1	Factor 2
height	.859	-.372
arm span	.842	-.441
forearm leng...	.813	-.459
lower leg len...	.84	-.395
weight	.758	.525
bitrochanter...	.674	.533
chest girth	.617	.58
chest width	.671	.418

Once the number of factors have been determined, it is necessary to determine the correlation of each variable with each factor, a structure value typically referred to as a loading. Most modern day factor analysts view this initial, *unrotated factor matrix* as the initial step in determining a desirable factor solution matrix. The square of a structure value represents the proportion of variance of the variable associated with the row that can be predicted by the factor associated with the column.

Computing the sum of the squared structure values by row results in a proportion, the *final communality estimate*, that represents the total proportion of variance of the variable that can be predicted by the factors.

Communality Summary

	SMC	Final Estimate
height	.816	.877
arm span	.849	.903
forearm leng...	.801	.872
lower leg len...	.788	.861
weight	.749	.85
bitrochanter...	.604	.739
chest girth	.562	.717
chest width	.478	.625

Prior to a factor analysis, the total proportion of variance of a variable that is predictable is estimated by the squared multiple correlation of the variable. Both the communality estimates and the squared multiple correlations (SMC) are reported in the communality summary table. Some analysts prefer to think of the squared multiple correlations as the initial communality estimates, while others prefer to think of the largest off-diagonal entry associated with the variable as the initial communality estimate. In the situation where a singular (determinant equal to 0) correlation matrix is analyzed, the initial communality estimate is assumed to be 0.

For the eight physical variables, it can be seen from the communality summary table that approximately 82 percent of the variation in height is predictable in a linear regression equation using the other seven variables. This conclusion is derived from the squared multiple correlation of height. When two factors are used to predict height, approximately 88% of the variation is predictable, an improvement of approximately 6%.

On occasion, StatView reports a final communality estimate that is slightly larger than 1. When this occurs, the associated variable is referred to as a *Heywood case*. Generally, the associated factor analysis will not be flawed. For further information on the Heywood case, the interested user is referred to the Gorsuch book.

Orthogonal Transformation Solution-Varimax

	Factor 1	Factor 2
height	.9	.26
arm span	.93	.195
forearm leng...	.919	.164
lower leg len...	.899	.229
weight	.251	.887
bitrochanter...	.181	.84
chest girth	.107	.84
chest width	.251	.75

When interpreting a factor solution (attempting to name the factors), it is substantially easier to deal with solutions where the variables have high loadings on just one factor or 0 loadings on most factors (a simple structure). Simple structure is best achieved by allowing the final factors to be correlated with each other: an oblique solution. Some analysts prefer solutions in which the factors are uncorrelated: orthogonal solutions. StatView uses a general orthogonal transformation procedure to compute an orthogonal solution. The general solution algorithm provides a choice of either the varimax orthogonal solution, the equamax orthogonal solution, or the quartimax orthogonal solution. With only a few exceptions, the contributions of the factors to a factor solution:

- are evenly distributed across all factors with an equamax solution

- tend to be concentrated on just a few factors with the quartimax solution
- are neither of the two extremes just noted when defined as a varimax solution

For the eight physical variables, the StatView default, a varimax transformation solution, was determined as the orthogonal solution. Notice that for this solution, the first four variables show large positive loadings on correlations with the first factor, while the last four variables show large positive loadings on the second factor. The correlations of the first four variables with the second factor are quite low. Similarly, the correlations of the second four variables with the first factor are quite low. However, many of these low correlations are none-the-less statistically significant. An oblique solution, correlated factors, will reduce the magnitude of these low loadings. Small, but statistically significant, loadings in an orthogonal factor solution are suggestive that the solution might be better described as an oblique solution.

Oblique Solution Primary Pattern Matrix-Orthotran/Varimax

	Factor 1	Factor 2
height	.919	.033
arm span	.973	-.047
forearm leng...	.971	-.08
lower leg len...	.928	-4.82E-4
weight	-.001	.922
bitrochanter...	-.064	.89
chest girth	-.146	.911
chest width	.043	.768

When determining an oblique solution, StatView uses an algorithm that simply takes a given orthogonal solution and releases the restriction of orthogonality. The algorithm, the orthotran solution, always defines a simple structure solution that is good as or better than the associated orthogonal simple structure solution.

There are two types of oblique solution in factor analysis a *primary pattern solution* and a *reference structure solution*. StatView determines both types of solution. These solutions are quite similar; indeed, one is a column rescaling of the other. The pattern solution defines loadings that are regression coefficients for predicting the standard score from of a variable in terms of the defined factors. The reference structure solution defines loadings that are correlations. Both solutions have good simple structure in the sense that the high loadings are high, and the low loadings are near 0.

Oblique Solution Reference Structure-Orthotran/Varimax

	Factor 1	Factor 2
height	.795	.029
arm span	.841	-.041
forearm leng...	.839	-.069
lower leg len...	.802	-4.17E-4
weight	-.001	.797
bitrochanter...	-.056	.77
chest girth	-.127	.788
chest width	.037	.664

When comparing a primary pattern solution to a reference structure solution, it is immediately apparent that the large loadings are larger in the primary pattern solution. Sometimes these primary pattern values become larger than 1, simply

because they are regression weights. Regardless of whether you use a primary pattern or reference structure solution, the conclusions should be the same. For the eight physical variables, it is clear that the first four variables are associated with the first factor and not at all associated with the second factor. Using similar logic, it is apparent that the second four variables are associated with the second factor. If we were to attempt to name the factors, we would attempt to name the first factor so that it represented the essence of the variables loading on it. The first factor could be named *bone structure*. The last four variables all deal with flesh and the second factor could be named *flesh factor*.

Clearly, for these data we would have arrived at the same factor name if we had attempted to interpret the factors from an orthogonal solution. Is it reasonable to assume that body weight or flesh is independent of bone structure? If your response is "yes," then you might be satisfied with an orthogonal solution. If, however, you assumed that taller people are in general heavier and fleshier, than shorter people, then you would be satisfied with an oblique solution.

Primary Intercorrelations-Orthotran/Varimax

	Factor 1	Factor 2
Factor 1	1	
Factor 2	.503	1

When utilizing an oblique solution, you are obligated to define the intercorrelations of the factors. Regardless of whether you are using a primary pattern or reference structure, it is the primary factor intercorrelations that are reported. For the eight physical variables, there is only one correlation, the correlation between the flesh factor and the bone factor. According to the StatView solution, the correlation between these two factors for this particular dataset is .503. Typically, you do not define oblique solutions with factor intercorrelations much larger than .50. If you were to have large factor intercorrelations, say .70 or higher, this would suggest that the factor solution may have been under-factored; that more factors should have been extracted in the initial solution.

Variable Complexity-Orthotran/Varimax

	Orthogonal	Oblique
height	1.166	1.003
arm span	1.088	1.005
forearm leng...	1.063	1.013
lower leg len...	1.13	1
weight	1.159	1
bitrochanter...	1.092	1.01
chest girth	1.032	1.052
chest width	1.221	1.006
Average	1.119	1.011

Variable complexity refers to the factor density of a variable. Usually for ideal simple structure, which you never have with an unrotated initial solution, each variable is accounted for by no more than one factor. When you have this ideal simple structure, the average *variable complexity* will be 1. That is, on the average, each variable is defined by no more than one factor.

To the extent that simple structure is not achieved, each variable is defined by more than one factor, and the average variable complexity will be greater than unity. The average variable complexity of an oblique solution will always be less than the average variable complexity of an orthogonal solution.

For the eight physical variables it is apparent that the average variable complexity for both orthogonal and oblique solutions is low. The variables height, lower leg length, weight and chest width are a bit more factorially dense than the other variables in the orthogonal solution. Notice how the oblique solution has reduced the complexity of these variables. This reduction in complexity may be verified by comparing the orthogonal solution with the oblique reference structure solution. The oblique solution is a near perfect simple structure solution.

Proportionate Variance Contributions

	Orthogonal	Oblique		
	Direct	Direct	Joint	Total
Factor 1	.543	.313	.431	.743
Factor 2	.457	.265	-.009	.257

Factors make different proportionate contributions to the common or explained variance. The greater the proportionate variance contribution of a factor, the more important the factor is in terms of explaining the intercorrelations of the variables. The *direct proportionate contribution* of a factor, regardless of whether you are looking at an orthogonal or an oblique solution, represents the proportion of the common variance that the factor accounts for independent of the other factors. The *joint proportionate contribution* of a factor is only defined for an oblique solution since it deals with shared variance or variance that is common to more than one factor. At times, you may note trivial negative joint contributions. Such negative contributions tend to be negligible. It is the positive joint proportionate contribution (such as the .431 in the eight physical variables example) that is of interest.

This suggests that 43% of the common variance may be attributable to the covariation of the two factors. When looking at the proportionate variance contributions of the oblique factors relative to the orthogonal factors, we see that for the eight physical variables data, the bone factor accounts for a greater proportion of the common variance than the flesh factor, but it accounts for substantially more variance (74%) as an oblique factor. Of that 74%, 26% is accounted for independently of the other factor.

Factor Score Weights for Oblique Transformation Solution-Orthotran/Varimax

	Factor 1	Factor 2
height	.321	-.129
arm span	.355	-.172
forearm leng...	.361	-.185
lower leg len...	.331	-.145
weight	-.175	.39
bitrochanter...	-.192	.386
chest girth	-.225	.408
chest width	-.131	.318

The final display provided by the StatView factor analysis is a display of factor score weights. If you wish to convert the original variable observations to standardized factor scores (for the eight physical variables data, a flesh score and a bone score), the columns of the factor score weight matrix may be thought of as the standardized regression weights for converting the eight physical variables in standard score format to two standard scores, one for flesh and one for bone. If the oblique factor score weights are used, then the intercorrelations of the factor scores is precisely and exactly .503, as defined by the factor intercorrelation matrix.

Factor Score Weights for Orthogonal Transformation Solution-Varimax

	Factor 1	Factor 2
height	.276	-.045
arm span	.297	-.077
forearm leng...	.299	-.089
lower leg len...	.281	-.058
weight	-.063	.332
bitrochanter...	-.08	.324
chest girth	-.107	.337
chest width	-.04	.274

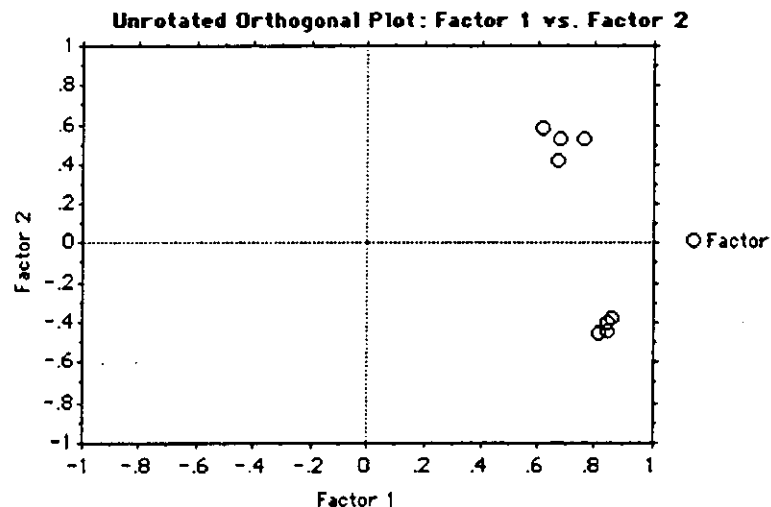
If the orthogonal factor score weights are used, then the intercorrelation of the factor scores is exactly .00. Whether you use orthogonal or oblique factor scores is really a matter of personal preference. If you want to interpret an oblique solution, then the oblique factor scores should be used.

Graphic Views

StatView provides three plots: those associated with the unrotated factor solutions, those associated with the orthogonal solution and those associated with the oblique solution. Within any particular set of plots, all pairwise factor plots are presented.

Unrotated Solution

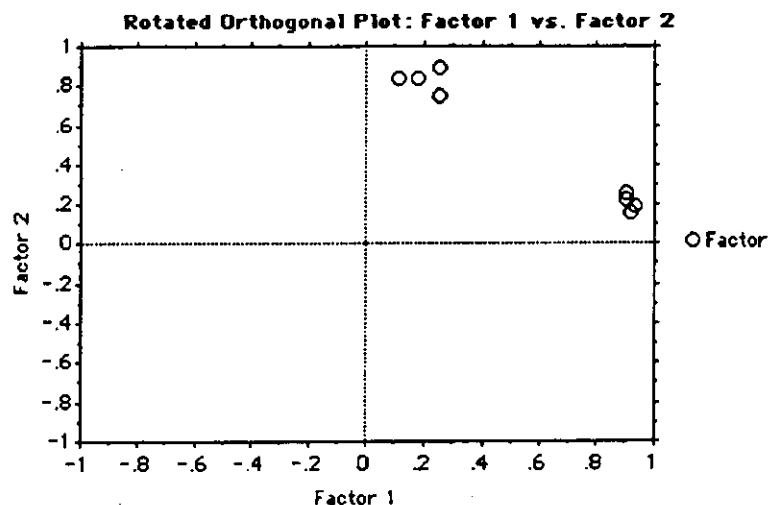
- Select Scattergram from the View menu.



The plot of the unrotated solution allows you to make a quick judgment regarding the potential simple structure of the factor solution. For the eight physical variables problem, two distinct clusters of points are apparent in the unrotated plot. An ideal factor solution for the variables, from a simple structure perspective, would have one axis passing through the cluster of variables 1 through 4 in the upper right hand quadrant, and the other axis passing through the other cluster. If the data were under-factored (which is not possible with the eight physical variables), then you might see points scattered throughout all four quadrants with no definitive clusters of points. If the data were over-factored, you would see many points near the point of intersection of the two axes, and perhaps one or two points defining a cluster.

Orthogonally Rotated Solution

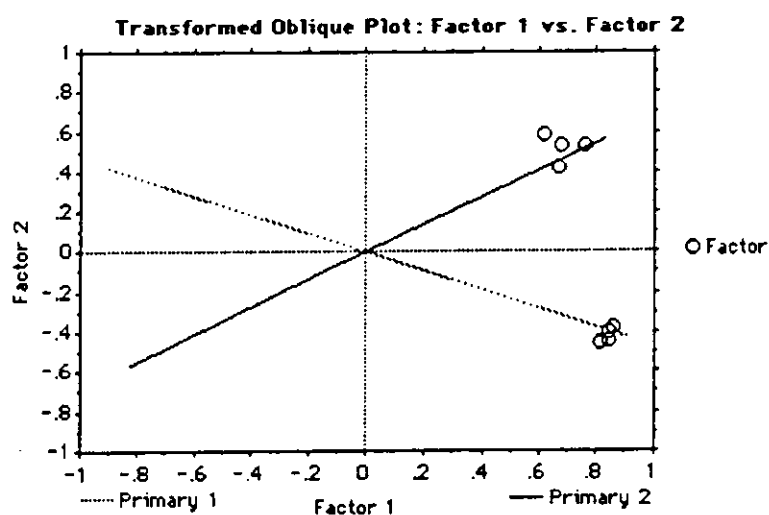
- Click the down arrow in the scroll bar.



The plot of the orthogonally rotated solution should show the axes either near the clusters or passing through them. If the axes pass through the clusters, then an orthogonal solution, uncorrelated factors, is appropriate for the data. If the axes are near, but not through, the clusters (as they are for the eight physical variables), then an oblique solution is most appropriate. A dataset that has been over-factored is apparent in the plot of the orthogonal solution since most of the plotted variables will be right at the origin and small clusters of a few variables will be near an axis. A plot of an under-factored orthogonal solution will be similar to the plot of the unrotated solution.

Oblique Solution

- Click the down arrow in the scroll bar.



The plot of the oblique solution should show the oblique axes, primary axes, passing through the clusters of points as they do for the eight physical variables. Notice that the plotted primary axes are not a right angles. This is because the axes are correlated. In this example, the simple structure of the oblique solution is quite good. This is because the primary axes pass directly through the clusters. In a case where the orthogonal solution passes axes through the clusters it may be noted

ANOVA

that the oblique solution and the orthogonal solution are identical and the factor intercorrelations are 0.

StatView can calculate a full interaction model for factorial or repeated measures experiments.

Factorial experiments may have up to 16 grouping factors (also called between factors) and need not be balanced (that is, cell frequency does not have to be the same throughout the model). In fact, StatView can handle factorial models with missing cells if the model is connected.

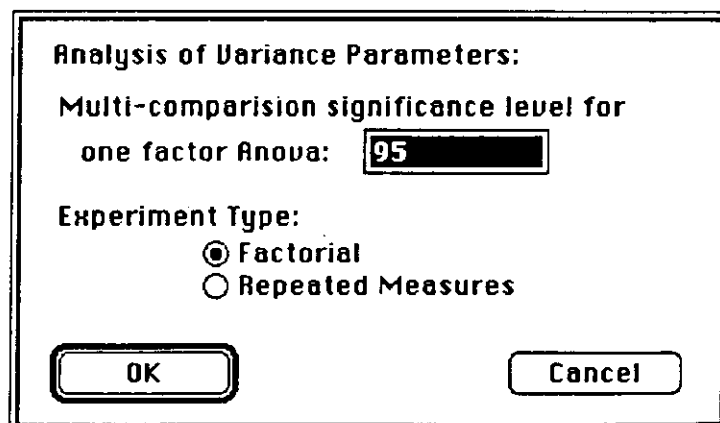
Repeated measures models can have up to 15 grouping factors and a single within factor. If you specify a repeated measures design, StatView automatically builds the correct ANOVA table for this type of model. The following restrictions hold for StatView's ANOVA feature.

- StatView only solves full interaction models. A full interaction model contains each factor as a main effect and every possible combination of the factors as interaction effects.
- For Repeated measure experiments there can be no more than one repeated measures factor (also referred to as a within factor).
- Repeated measures models with 2 or more grouping factors must be balanced.

For different model designs, Abacus Concepts produces SuperANOVA, an advanced general linear modeling application that can solve virtually any ANOVA model. Abacus Concepts offers a free demonstration version of SuperANOVA.

Multiple comparison tests (Scheffé F-test, Fisher's PLSD, Dunnett t-test) are performed for all single factor models. The significance level for the Dunnett is not reported by StatView. In order to interpret the Dunnett you must check a Dunnett table. Such tables are found in most statistics books that treat the topic of multiple comparisons (see Winer p. 874). Please note that StatView computes significance levels for Fisher's PLSD even when the overall significant level for the factor does not meet your significance level. Do not report significant mean differences unless your overall F test is significant.

- Select ANOVA from the Compare menu. The following dialog box is displayed:



The dialog box is titled "Analysis of Variance Parameters:". It contains a label "Multi-comparison significance level for one factor Anova:" followed by a text input field containing the value "95". Below this is a label "Experiment Type:" followed by two radio button options: "Factorial" (which is selected) and "Repeated Measures". At the bottom of the dialog are two buttons: "OK" and "Cancel".

If you are analyzing a single factor model, enter the significance level for the multiple comparison tests in the text entry rectangle. StatView will default to a significance level of 95%.

Choose whether the experiment type is factorial or repeated measure.

Assigning Variables

Specify the experiment design through X and Y variable assignments. The number of grouping factors is determined by the number of X variables. The dependent variable is specified via the Y variable. For repeated measure experiments, the within-factor is specified with multiple Y variables.

WARNING: Columns assigned as grouping (X) variables for ANOVAs, unpaired t-tests and some nonparametrics must be Category or Integer columns. The number of groups in the column must match the number of groups expected by the statistic you are computing.

For a Category column the number of groups is the number of distinct values in the column. (Note that this is not necessarily equivalent to the number of elements defined for the category.)

For an Integer column StatView calculates the number of groups as the column's Maximum Value - Minimum Value + 1. Each group must be represented in the column. For example, if you use an integer column containing the values 1 and 3 for an ANOVA, you will see an error message warning that a group contains no values. This is because StatView sees three groups and the group associated with 2 has no values. (This message also appears if all the records within any group have missing values.) If you use this column for a two group statistic you will see an error message warning you that the X column has the incorrect number of groups.

We recommend that you use category columns as grouping variables because category columns allow you to clearly label the groups you are analyzing.

To facilitate your understanding of StatView ANOVA, the examples below illustrate different structural models. This discussion uses several datasets:

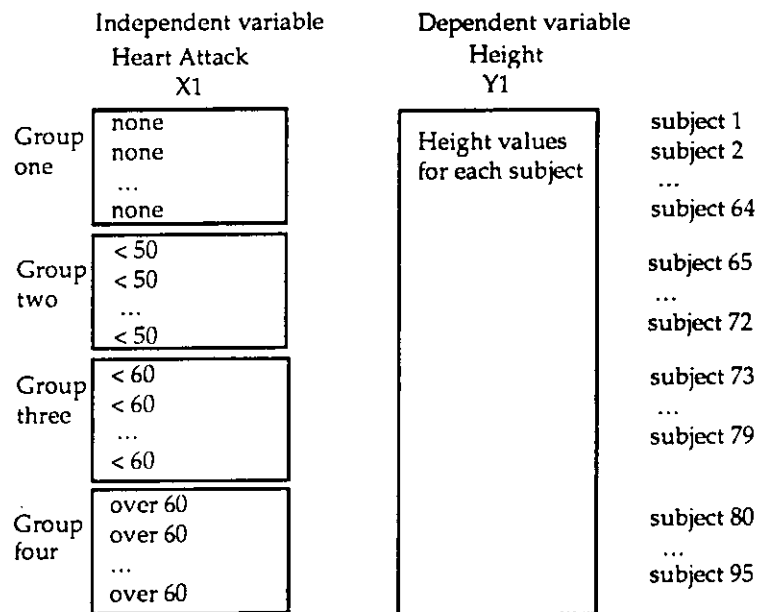
- Lipid Data
- Winer 2 Factor Balanced (Winer, p. 437)
- Afifi & Azen 2 Factor Unbalanced (Afifi and Azen, p. 166)
- Winer 1 Factor Repeated Measure (Winer, p. 268)
- Winer 3 Factor Repeated Measure (Winer, p. 525)

Single Factor Factorial, Non-Repeated Measure

Since this example is computing a single factor ANOVA, assign one X variable to the single grouping factor. The column assigned as the X variable must be a Category or Integer column and must be suitable as a grouping variable. (Please read the warning on the preceding page.) The Y variable specifies the dependent data column. A single factor factorial design ANOVA is a OneXOneY statistic. (See the table at the beginning of this chapter.)

- Open Lipid Data.

The figure below illustrates the dataset layout for this Single Factor Factorial, Non-Repeated Measure model.



It displays an analysis of four groups on a single dependent variable. The dependent variable, Height, is represented as a single data column and should be assigned as Y₁. A second column should be the grouping column, in this example Heart History. This grouping column will, by row, note the group to which the associated dependent variable belongs. The grouping column should be assigned as X₁.

- Assign Y to Height.
- Assign X to Heart History.
- Select ANOVA from the Compare menu.
- Use the default, 95, as the significance level for post hoc means comparisons. Note: This corresponds to an alpha level of .05.
- Click Factorial.
- Click OK, using the default, Factorial, as the experiment type. The following table appears:

One Factor ANOVA X₁ : Heart History Y₁ : Height

Analysis of Variance Table				
Source:	DF:	Sum Squares:	Mean Square:	F-test:
Between groups	3	85.067	28.356	1.674
Within groups	91	1541.626	16.941	p = .1782
Total	94	1626.693		

Model II estimate of between component variance = .713

The view title specifies the experiment design and the X and Y variables. The ANOVA table includes:

- the between groups degrees of freedom, sum of squares, mean square
- the within-groups degrees of freedom, sum of squares, mean square
- the total degrees of freedom, sum of squares
- F statistic and probability

For a single factor, the Model estimate of between component variance is computed. This is an unbiased estimate for the variance of the population of differential effects due to the single factor. (Afifi & Azen, p.215).

Note that the ANOVA is not significant at the .05 level. If we had chosen the .20 level, the ANOVA would have been significant.

- Click the down arrow in the scroll bar to display the table of cell means.

One Factor ANOVA X₁: Heart History Y₁: Height

Group:	Count:	Mean:	Std. Dev.:	Std. Error:
none	64	69.693	4.087	.511
<50	8	66.344	4.904	1.734
<60	7	70.071	3.306	1.25
over 60	16	69.031	4.125	1.031

For each group the following summary statistics are provided:

- name
- count of cases
- mean
- standard deviation
- standard error of the mean

If there are more than five groups, the table is continued on successive pages.

This table should always be examined to determine the number of data values in each cell of the design as well as cell means and standard deviations.

- Click the down arrow in the scroll bar to display the table of post hoc means tests.

One Factor ANOVA X₁: Heart History Y₁: Height

Comparison:	Mean Diff.:	Fisher PLSD:	Scheffe F-test:	Dunnett t:
none vs. <50	3.35	3.066 *	1.57	2.17
none vs. <60	-.378	3.255	.018	.231
none vs. over 60	.662	2.285	.11	.576
<50 vs. <60	-3.728	4.231	1.021	1.75
<50 vs. over 60	-2.688	3.54	.758	1.508

* Significant at 95%

Comparisons between treatment means are provided in the last table. For each treatment comparison the following information is provided:

- difference between the means
- Fisher's PLSD test
- Scheffé F-test
- Dunnett t-test

If the Fisher and Scheffé test are significant at the level entered in the dialog box, an asterisk (*) appears by the comparison value. No significance is computed for the Dunnett test.

Notice that the none vs. < 50 difference is marked significant in the Fisher PSLD column. You must not interpret this as a significant difference because the F test from page one was not significant. The P in PSLD stands for *protected*. The protection results from not interpreting the PSLD column unless the F test is significant. If you go back to the ANOVA choice from the **Compare** menu and select 80 for the Multi-comparison significant level, the PSLD values will correspond to an alpha of .20. While almost no one would ever use an alpha of .20, in this example the Fisher's PSLD values would now be appropriate to interpret.

There are no graphic views available.

Two Factor Factorial, Non-Repeated Measures — Balanced Model

When analyzing a two factor ANOVA, assign X variables to each of the two grouping factors. The columns assigned as X variables must be Category or Integer columns and must be suitable as grouping variables. (Please read the warning at the beginning of the section on ANOVA in this chapter.) The Y variable specifies the dependent data column. A Two Factorial design is a ManyXOneY statistic. (See the table at the beginning of this chapter.)

- Open the Winer 2 Factor Balanced dataset.

This data is from Winer (1971, p. 436). Winer describes the experiment as follows: The experiment evaluates the effectiveness of three drugs in bringing about behavioral changes in two categories of patients. A random sample of nine patients belonging to the first category (schizophrenics) is divided at random into three groups, with three patients in each subgroup. Each subgroup is assigned to one of the drug conditions. The same procedure is followed for the nine patients belonging to the second category (depressives). The rating column contains criteria ratings made on the patient before and after the administration of the drugs. The figure below illustrates the layout of the data.

Independent variables		Dependent variable	
Category of Patient	Drug	Rating	
X1	X2	Y1	
Schizophrenic	Drug a	Rating values for subjects one subject per row.	Subject 1
	Drug b		Subject 2
	Drug c		Subject 3
Depressive	Drug a		Subject 4
	Drug b		Subject 5
	Drug c		Subject 6
			Subject 7
			Subject 8
			Subject 9
			Subject 10
			Subject 11
			Subject 12
			Subject 13
			Subject 14
			Subject 15
			Subject 16
			Subject 17
			Subject 18

In this example we have two independent variables. The first independent variable, **Category of patient**, has two levels: schizophrenic and depressive. The second independent variable, **drug**, has three levels: drug a, drug b, drug c. This model will require three columns of data: two grouping columns and a dependent variable column. There is a unique group for each crossing of the levels of the independent variables. This is a balanced design because all groups have equal sample sizes. If any pair of groups had unequal sample sizes, it would be an unbalanced design.

The first independent variable column, X₁, is comprised of four unique entries, one for each unique level. The second independent variable column, X₂, is comprised of two unique values, one for each of the two levels. The third column, Y₁, is the dependent variable column. Reading across a row, the first entry denotes the individual's group on independent variable 1; the second entry denotes the individual's group on independent variable 2. The third entry denotes the individual's observed value on the dependent variable.

- Assign Y to Rating.
- Assign X to Category of Patient and to Drug.
- Select ANOVA from the Compare menu.
- Click OK.

Two factor and higher ANOVA calculations present a dialog box which specifies the number of factors in the model and indicates the progress of the computation. Note that multi-factor ANOVAs often take a long time to compute. The following table will appear:

Anova table for a 2-factor Analysis of Variance on Y₁: Rating

Source:	df:	Sum of Squares:	Mean Square:	F-test:	P value:
Category of Patient (...)	1	18	18	2.038	.1789
Drug (B)	2	48	24	2.717	.1063
AB	2	144	72	8.151	.0058
Error	12	106	8.833		

There were no missing cells found.

The view title names the experiment design and the Y variable analyzed. The note on the bottom of the page indicates whether any missing cells (cells with no observed values) were found. The ANOVA table includes:

- for the main effect of Factor A: the degrees of freedom, sum of squares, mean square, F-ratio and probability value
- for the main effect of Factor B: the degrees of freedom, sum of squares, mean square, F-ratio and probability value
- for AxB interaction: the degrees of freedom, sum of squares, mean square, F-ratio and probability value
- for Model Error: the degrees of freedom, sum of squares, mean square
- Click the down arrow in the scroll bar.

The AB incidence table on Y1: Rating

Drug:		drug a	drug b	drug c	Totals:
category:	schizophre...	3	3	3	9
		4	8	6	6
depressives		3	3	3	9
		10	2	12	8
Totals:		6	6	6	18
		7	5	9	7

On the pages following the ANOVA table, StatView provides the incidence table. The incidence table details the count and mean of each cell found in the analysis. The cell count is the top value in each cell; the cell mean is the lower value. Each row and column of the incidence table is provided with that row or column's total count and mean. This particular analysis contains six cells. Note that factor A (X1) is always the vertical component of the incidence table and any other factors are displayed horizontally.

There are no graphic views available.

Two Factor Factorial, Non-Repeated Measures — Unbalanced Model

The layout for this model is the same as in the preceding Two Factor Factorial, Non-Repeated Measures example.

- Open the Afifi & Azen 2 Factor Unbalanced dataset.

This data is from Afifi & Azen (1972, p. 166). The experiment evaluates the increase in systolic blood pressure resulting from four different treatments for three different diseases.

The two factors in the experiment are treatment and disease. The first factor in the experiment, factor A, is specified by the column Treatment. This is a category column containing four groups: Drug A, Drug B, Drug C, and Drug D. The second factor in the experiment, factor B, is specified by the Disease column. This is a category column containing three groups: Disease 1, Disease 2, and Disease 3. The blood pressure data is in the column Systolic Pressure.

This design is unbalanced because the cells have unequal sample sizes.

- Assign Y to Systolic Pressure.
- Assign X to Treatment and to Disease.
- Select ANOVA from the Compare menu.
- Click OK.

Unbalanced models take longer time to evaluate than balanced models. A dialog box informs you on the progress of the analysis. The following table will appear:

Anova table for a 2-factor Analysis of Variance on Y₁: Systolic Pressure

Source:	df:	Sum of Squares:	Mean Square:	F-test:	P value:
Treatment (A)	3	2997.472	999.157	9.046	.0001
Disease (B)	2	415.873	207.937	1.883	.1637
AB	6	707.266	117.878	1.067	.3958
Error	46	5080.817	110.453		

There were no missing cells found.

The view title specifies the experiment design and the Y variable analyzed. The ANOVA and Incidence table have the same format as the balanced two-level ANOVA.

The AB Incidence table on Y₁: Systolic Pressure

Disease:		Disease 1	Disease 2	Disease 3	Totals:
Treatment	Drug A	6 29.333	4 28.25	5 20.4	15 26.067
	Drug B	5 28	4 33.5	6 18.167	15 25.533
	Drug C	3 16.333	5 4.4	4 8.5	12 8.75
	Drug D	5 13.6	6 12.833	5 14.2	16 13.5
	Totals:	19 22.789	19 18.211	20 15.8	58 18.879

There are no graphic views available.

Single Factor Factorial — One Repeated Measure

Experiments in which the same subject is observed under each treatment are repeated measure experiments. *In StatView, the single factor repeated measures ANOVA is set up differently than the multiple factor repeated measures models. For single factor models, each treatment is assigned an X variable. It is a ManyX statistic. (See the table at the beginning of this chapter.)*

- Open the Winer 1 Factor Repeated Measure dataset.

This data is from Winer p. 268. Winer describes the experiments as follows: The experiment studies the effects of four drugs upon reaction time to a series of tasks.

Each subject was observed under each of the drugs. Each column reflects the mean reaction time of a subject to the series of tasks.

Dependent variables				
Drug 1	Drug 2	Drug 3	Drug 4	
X1	X2	X3	X4	
Four values in each row, one subject per row.				Subject 1
				Subject 2
				Subject 3
				Subject 4
				Subject 5

Note: For the single factor model only, the repeated measures factor is indicated by X variables. This example has four X variables assigned, indicating each subject's four measures on the dependent variable. Each subject will have four row entries involved in the analysis.

- Assign X to each of the four Drug columns.
- Select ANOVA from the Compare menu. Note that the default with many X assigned is Repeated Measure.
- Enter 99 as the significance level.
- Click OK, using the default, Repeated Measures, as the experiment type. The following table appears:

One Factor ANOVA-Repeated Measures for X1 -- X4

Source:	df:	Sum of Squares:	Mean Square:	F-test:	P value:
Between subjects	4	680.8	170.2	3.148	.0458
Within subjects	15	811	54.067		
treatments	3	698.2	232.733	24.759	.0001
residual	12	112.8	9.4		
Total	19	1491.8			

Reliability Estimates for- All treatments: .682 Single Treatment: .349

The view title specifies the experiment design and the number of X variables. The ANOVA table includes:

- the between-subjects degrees of freedom, sum of squares, mean square
- the within-subjects degrees of freedom, sum of squares, mean square
- the treatments degrees of freedom, sum of squares, mean square, F value and probability
- the residual degrees of freedom, sum of squares, mean square
- the total degrees of freedom, sum of squares

Reliability estimates are computed for the mean of all treatments and for a single treatment (Winer, p. 283).

- Click the down arrow in the scroll bar.

One Factor ANOVA-Repeated Measures for X₁ ... X₄

Group:	Count:	Mean:	Std. Dev.:	Std. Error:
Drug 1	5	26.4	8.764	3.919
Drug 2	5	25.6	6.542	2.926
Drug 3	5	15.6	3.847	1.72
Drug 4	5	32	8	3.578

For each treatment the following summary statistics are provided:

- name
- count of cases
- mean
- standard deviation
- standard error of the mean

Note that mean for Drug 3 is significantly lower than the other means. This indicates that Drug 3 is associated with the fastest reaction.

If there are more than five treatments this table is continued on successive pages

- Click the down arrow in the scroll bar to display the table of post hoc means tests.

One Factor ANOVA-Repeated Measures for X₁ ... X₄

Comparison:	Mean Diff.:	Fisher PLSD:	Scheffe F-test:	Dunnett t:
Drug 1 vs. Drug 2	.8	5.923	.057	.413
Drug 1 vs. Drug 3	10.8	5.923*	10.34*	5.57
Drug 1 vs. Drug 4	-5.6	5.923	2.78	2.888
Drug 2 vs. Drug 3	10	5.923*	8.865*	5.157
Drug 2 vs. Drug 4	-6.4	5.923*	3.631	3.301

* Significant at 99%

Comparisons between treatment means are provided in the last table. For each treatment comparison the following information is provided:

- difference between the means
- Fisher's PLSD test
- Scheffé F-test
- Dunnett t-test

If the Fisher and Scheffé test are significant at the level entered in the dialog box, an asterisk (*) will appear by the comparison value. No significance is computed for the Dunnett test. See the discussion of this table for the single factor - non-repeated measures example for more information.

There are no graphic views available.

Three Factor Factorial — One Repeated Measure

Multiway repeated measures models are experiments where there are repeated measures on one factor (the within-subjects factor) and one or more grouping factors. Grouping factors refer to effects that occur between groups while within-subjects factors refer to effects measured by differences within subjects.

For models containing two or more factors, the repeated measure (within-subjects factor) is specified by two or more dependent Y variables. The grouping factors are specified by X variables. The columns assigned as X variables must be Category or Integer columns and must be suitable as grouping variables. (Please read the warning at the beginning of the section on ANOVA in this chapter.) If there are two or more grouping factors the data must be balanced.

A multi-factor repeated measure design is a ManyXManyY statistic. (See the table at the beginning of this chapter.)

- Open the Winer 3 Factor Repeated Measure dataset.

This data is from Winer (1971, p. 564). The experiment evaluates the effect of anxiety and muscular tension on a learning task. Each factor has two levels. The experiment is repeated four times for each subject (Trial 1 through Trial 4) with the number of errors recorded.

Independent variables		Dependent variables				
Anxiety	Tension	Trial 1	Trial 2	Trial 3	Trial 4	
X1	X2	Y1	Y2	Y3	Y4	
Low	None	Four values in each row, one subject per row.				Subject 1
	High					Subject 2
High	None					Subject 3
	High					Subject 4
Low	None					Subject 5
	High					Subject 6
High	None					Subject 7
	High					Subject 8
Low	None					Subject 9
	High					Subject 10
High	None					Subject 11
	High					Subject 12

This model has three independent variables, one of which is a repeated measures factor. There are only three factors associated with this model (Anxiety, Tension, and Trial), but there will be six columns involved in the analysis, two X columns and four Y columns. The first X column will be associated with independent variable Anxiety, X₁, and will have two unique values, Low and High, one for each level of the variable. The second X column will be associated with independent variable Tension, X₂, and will have two unique values, None and High, one for each level of the variable. The third independent variable (the repeated measures factor) will have four Y columns associated with it, one for each subject's four measures on the dependent variable, Y₁, Y₂, Y₃, and Y₄. Each subject will have six row entries involved in the analysis.

- Assign Y to the four Trial columns.
- Assign X to Anxiety and to Tension.

- Select ANOVA from the Compare menu.
- Click Repeated Measures.
- Click OK.

The dialog box keeps you up to data on the progress of the analysis. The following table will appear:

Anova table for a 3-factor repeated measures Anova.

Source:	df:	Sum of Squares:	Mean Square:	F-test:	P value:
Anxiety (A)	1	10.083	10.083	.978	.3317
Tension (B)	1	8.333	8.333	.808	.3949
AB	1	80.083	80.083	7.766	.0237
subjects w. groups	8	82.5	10.312		
Repeated Measure (C)	3	991.5	330.5	152.051	.0001
AC	3	8.417	2.806	1.291	.3003
BC	3	12.167	4.056	1.866	.1624
ABC	3	12.75	4.25	1.933	.1477
C x subjects w. groups	24	52.167	2.174		

There were no missing cells found.

The view title specifies the experiment design. The ANOVA table contains:

- the sum of squares, mean square, degrees of freedom, F-test and p value for all the between subjects variation (in this example, these are the A, B and the AB interaction)
- the Subjects within Groups' sum of squares, mean square, and degrees of freedom. Note that this line indicates the error for the between factors A & B and serves as the denominator for the between subjects' F-tests
- the sum of squares, mean square, degrees of freedom, F-test and p value for all the within subjects variation (in this example, these are C, AC, BC, and ABC)
- the C x Subjects within Groups' sum of squares, mean square, and degrees of freedom. Note that this line indicates the error for the within factors and serves as the denominator for the within subjects' F-tests.

The note on the bottom of the page indicates whether or not any missing cells (cells with no observed values) were found.

- Click the down arrow in the scroll bar.

Just as in the factorial model, the ANOVA table is followed by the incidence table. Since this example has three factors, StatView displays four different incidence tables to give a comprehensive breakdown of the data. The tables given are the AB, AC, BC and ABC cell counts and means. Each table is clearly labelled.

There are no graphic views available.

Contingency Tables and Cross-tabs

A Contingency table computes both the Chi-Square Goodness of Fit (comparing the observed frequency of a single group sample with the expected frequency) and the Formation of Contingency (frequency, cross-tabulation) tables with summarizing statistics. The statistics include: Total Chi-Square, G Statistic, Cramer's V, Phi (2x2 only), Chi-Square with Continuity Correction (2x2 only). The Contingency tables can be tabulated from coded data or input as previously tabulated data.

StatView uses a relatively new procedure for identifying the cells of a contingency table responsible for a significant chi square, referred to as the *post hoc cell contributions*. An adjusted residual is computed for each cell of a contingency table associated with a chi square. This adjusted residual is determined from a standardized residual and is approximately normally distributed with a mean of 0 and a standard deviation of 1. Thus, an adjusted residual of 1.96 suggests that the deviation of the cell observed frequency from the cell expected frequency is significant at the .05 level. Usually a cell significance only accompanies a significant chi square.

- Select Contingency Table from the Compare menu, and the Contingency Table dialog box appears:

Select Test:

☐ Chi-Square (One Group)

☒ Contingency Table analysis

---Contingency Table Information---

The contingency table size is determined by your data.

Select the source for contingency table data

☐ Coded Raw Data (Y-rows, H-columns)

☒ Tabulated Data (from data window)

Additional Frequency Tables for display:

☒ Row % ☒ Expected Values

☒ Column % ☒ Post-hoc Contributions

OK Cancel

The dialog box allows you to select which type of test you are performing.

Select Chi-Square (One Group) if you are comparing the frequency observed in a sample with the expected frequency derived from a theoretical model.

Select Contingency Table if you are analyzing two variables which are classified into a number of categories or attributes. If you are analyzing Contingency table data you need to specify additional information.

Data may be in coded raw form. In this case the program tabulates the contingency table for you. Data may already be in tabulated form in the dataset. The program reads the tabulated data.

All contingency table analyses include a summary page of statistics and the observed frequency table. You may also choose to display four additional tables:

- row percents
- column percents
- expected values
- post hoc cell contributions displaying the adjusted residual for each cell in the contingency table.

Chi-Square — One Group

The Chi-Square test compares the observed frequencies, located in an X column, to expected frequencies, located in a Y column. It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

For this statistic, we need to create a new dataset. Assume that we expect an even distribution across the four category elements of Alcohol use in Lipid Data. The expected frequency for each element is 23.75. We find the observed frequencies for the column using **Frequency Distribution** from the **Describe** menu. We create a new dataset with two columns, an Integer column titled Observed Alcohol use, and a Real column titled Expected Alcohol use. The first column contains the observed values: 32, 33, 28, and 2. The second column contains the expected values: 23.75 in each cell.

- Create a new dataset and enter the values described above.
- Assign X to Observed Alcohol use.
- Assign Y to Expected Alcohol use.
- Select **Contingency Table** from the **Compare** menu.
- Click **Chi-Square (One Group)**.
- Click **OK**. The following table appears:

One Group Chi-Square X1 : Observed Alcohol use Y1 : Expected Alcoh...

DF :	Chi-Square :	Probability :
3	27.147	.0001

There are no graphic views.

Contingency Table

When a Contingency table is to be tabulated by StatView, select **Coded raw data** and designate the row(s) with a Y variable and the column(s) with an X variable. Both variables must be either Category or Integer columns.

It is a OneXOneY statistic. (See the table at the beginning of this chapter.) If the resulting table has more than 8 rows or columns, you cannot compute multiple results.

When using a previously tabulated contingency table (entered in a dataset), select **Tabulated Data** and assign X variables to the dataset columns which make up the contingency table columns. The rows of the table are the included rows of the dataset. It is a ManyX statistic.

The maximum size of a contingency table is 1600 cells.

This example determines whether a relationship exists between patient gender and alcohol use. The data is coded (gender as male and female; alcohol use as none, < 2, 2 - 6, and > 6).

- Open and zoom Lipid Data.
- Assign X to Gender (columns of the contingency table being tabulated)
- Assign Y to Alcohol Use (rows of the table).
- Select **Contingency Table** from the **Compare** menu.
- Use the default settings to create a **Contingency Table** analysis from Coded raw data. The defaults will also display all four additional tables: Row Percents, Column Percents, Expected Values, and Post-hoc Contributions.
- Click OK. The following table appears:

Coded Chi-Square X1 : Gender Y1 : Alcohol use

Summary Statistics

DF:	3	
Total Chi-Square:	1.761	p = .6236
G Statistic:	1.665	
Contingency Coefficient:	.135	
Cramer's V:	.136	

The view title specifies the X and Y variables. The first page contains summary statistics for the contingency table.

- degrees of freedom
- total Chi-Square and probability
- G statistic
- contingency coefficient
- Cramer's V or Phi (2x2 only)
- Chi-Square with continuity correction and probability (2x2 only)

The pages after the summary contain the frequency tables. The maximum size table displayed on one page is 8x8. Larger tables are displayed on additional pages.

Pages 2 and 3 contain the observed frequency table and the percents of row totals table.

Observed Frequency Table

	male	female	Totals:
none	22	10	32
< 2	26	7	33
2 - 6	22	6	28
> 6	1	1	2
Totals:	71	24	95

Percents of Row Totals

	male	female	Totals:
none	68.75%	31.25%	100%
< 2	78.79%	21.21%	100%
2 - 6	78.57%	21.43%	100%
> 6	50%	50%	100%
Totals:	74.74%	25.26%	100%

Pages 4 and 5 contain the percents of column totals table and the expected values table.

Percents of Column Totals

	male	female	Totals:
none	30.99%	41.67%	33.68%
< 2	36.62%	29.17%	34.74%
2 - 6	30.99%	25%	29.47%
> 6	1.41%	4.17%	2.11%
Totals:	100%	100%	100%

Expected Values

	male	female	Totals:
none	23.92	8.08	32
< 2	24.66	8.34	33
2 - 6	20.93	7.07	28
> 6	1.49	.51	2
Totals:	71	24	95

Page 6 contains the post hoc cell contributions. These are the observed frequencies minus the expected values standardized to have a variance of 1 and mean of 0 when the data come from a multinomial distribution.

Post-Hoc Cell Contributions

	male	female
none	-.96	.96
< 2	.66	-.66
2 - 6	.56	-.56
> 6	-.81	.81

There are no graphic views available.

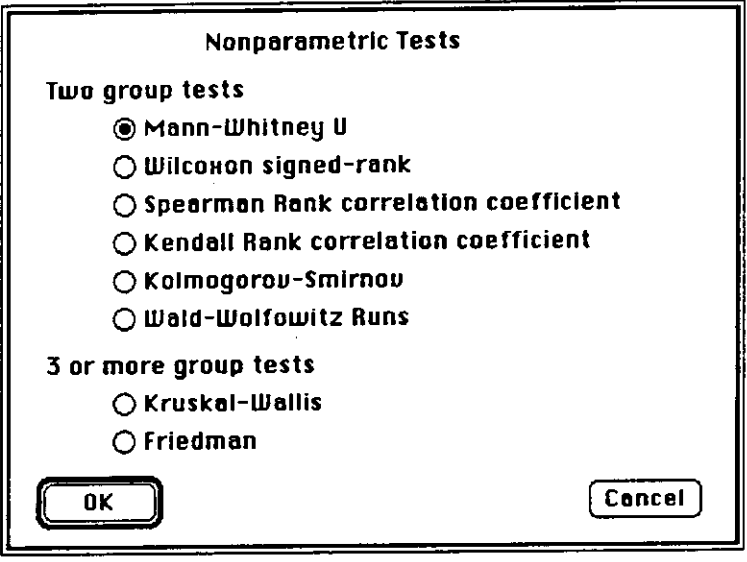
Nonparametrics

StatView computes the following nonparametric tests:

Two group tests	Three or more group tests
Mann-Whitney U Wilcoxon signed rank Spearman rank correlation coefficient Kendall rank correlation coefficient Kolmogorov-Smirnov Wald Wolfowitz runs	Kruskal-Wallis Friedman

There are no graphic views available for any of the nonparametric tests.

- Select **Nonparametrics** from the **Compare** menu. The Nonparametrics dialog box is displayed:



The dialog box is titled "Nonparametric Tests". It contains two sections. The first section, "Two group tests", has a list of six options with radio buttons: "Mann-Whitney U" (selected), "Wilcoxon signed-rank", "Spearman Rank correlation coefficient", "Kendall Rank correlation coefficient", "Kolmogorov-Smirnov", and "Wald-Wolfowitz Runs". The second section, "3 or more group tests", has two options with radio buttons: "Kruskal-Wallis" and "Friedman". At the bottom left is an "OK" button and at the bottom right is a "Cancel" button.

Mann-Whitney U

The Mann-Whitney U is the nonparametric version of the two group unpaired t-Test. It performs the test between two groups within a Y column. The groups in the Y column are specified by an X column. The column assigned as the X variable must be a Category or Integer column and must be suitable as a grouping variable. (Please read the warning at the beginning of the section on ANOVA in this chapter.) It is a OneXOneY Statistic. (See the table at the beginning of this chapter.)

This example compares the Cholesterol values of males and females in Lipid Data.

- Open Lipid Data.
- Assign X to Gender and Y to Cholesterol.
- Select **Nonparametrics** from the **Compare** menu.
- Click **OK**, since Mann-Whitney U is the default. The following table appears:

Mann-Whitney U X ₁ : Gender Y ₁ : Cholesterol			
	Number:	Σ Rank:	Mean Rank:
male	71	3398.5	47.866
female	24	1161.5	48.396

U	842.5
U-prime	861.5
Z	-.081 p = .9352
Z corrected for ties	-.081 p = .9351
* tied groups	24

The view title specifies the X and Y variables. For each group in the Y column, the top table shows:

- the number of observations in the group
- the sum of the ranks
- the mean rank

The summary statistics table shows:

- the Mann-Whitney U statistic
- the U-prime value
- the Z value and two-tail probability

If there are tied groups, this table includes the Z corrected for ties and two-tail probability, and the # of tied groups.

Wilcoxon Signed-Rank

The Wilcoxon Signed-Rank is the nonparametric version of the two group paired t-Test. It compares the values of paired X and Y columns. It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

This example compares the Cholesterol values of the patients in Lipid Data to the Chol-3yrs values.

- Open Lipid Data.
- Assign X to Cholesterol and Y to Chol-3yrs.
- Select Nonparametrics from the Compare menu.
- Click Wilcoxon signed-rank and click OK. The following table appears:

Wilcoxon signed-rank X ₁ : Cholesterol Y ₁ : Chol-3yrs			
	Number:	Σ Rank:	Mean Rank:
- Ranks	15	256	17.067
+ Ranks	26	605	23.269

note 2 cases eliminated for difference = 0.

Z	-2.261	p = .0237
Z corrected for ties	-2.262	p = .0237
* tied groups	10	

Note: 52 cases deleted with missing values.

The top table shows the following information for both the negative and positive ranks.

- number of each rank
- sum of the ranks
- mean of the ranks

If any cases are eliminated with a difference of 0, the number is noted.

The summary statistics show:

- Z and two-tail probability
- Z corrected for ties (if there are tied groups in the table) and two-tail probability
- the number of tied groups (if there are tied groups in the table)

Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient calculates a correlation based on the ranks of the values of an X and Y column. It is a OneXOneY statistic. (See the table at the beginning of this chapter.) This example compares the Cholesterol values with Weight in Lipid Data.

- Open Lipid Data.
- Assign X to Cholesterol and Y to Chol-3yrs.
- Select Nonparametrics from the Compare menu.
- Click Spearman Rank correlation coefficient and click OK. The following table appears:

Spearman Corr. Coef. X1: Cholesterol Y1: Chol-3yrs

N	43
ΣD^2	3382.5
Rho	.745
Z	4.826 p = .0001
Rho corrected for ties	.744
Z corrected for ties	4.825 p = .0001
*X tied groups: 5	*Y tied groups: 6

Note: 52 cases deleted with missing values.

The table shows the following statistics:

- number of cases
- sum of the squares of the difference of the ranks
- Spearman Rho
- Z and two-tail probability

If there are tied groups the table includes:

- Spearman Rho corrected for ties
- Z corrected for ties and two-tail probability
- number of tied groups in the X and Y variables

Kendall Rank Correlation Coefficient

The Kendall rank correlation coefficient calculates a correlation based on the ranks of the values of an X and Y column. It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

This example compares the base Cholesterol values of the Lipid Data patients with their Weight.

- Open Lipid Data. Assign X to Cholesterol.
- Assign Y to Chol-3yrs.
- Select Nonparametrics from the Compare menu.
- Click Kendall Rank correlation coefficient.
- Click OK. The following table appears:

Kendall Corr. Coef. X ₁ : Cholesterol Y ₁ : Chol-3yrs			
N	43		
Score	498		
Tau	.551		
Z	5.212	p = .0001	
Tau corrected for ties	.555		
Z corrected for ties	5.244	p = .0001	
*X tied groups: 5		*Y tied groups: 6	

Note: 52 cases deleted with missing values.

The view title names the X and Y variables. The table shows the following statistics:

- number of cases
- score
- the Kendall Tau
- Z and two-tail probability

If there are tied groups the table includes:

- the Kendall Tau corrected for ties
- Z corrected for ties and two-tail probability
- the number of tied groups in the X and Y variables

Kolmogorov-Smirnov Tests

The Kolmogorov-Smirnov tests the differences between two groups within a Y column. The groups in the Y column are specified by an X column. The column

assigned as the X variable must be a Category or Integer column and must be suitable as a grouping variable. (Please read the warning at the beginning of the section on ANOVA in this chapter.) It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

This example compares the distribution of male and female Cholesterol values in Lipid Data.

- Open Lipid Data. Assign X to Gender.
- Assign Y to Cholesterol.
- Select Nonparametrics from the Compare menu.
- Click Kolmogorov-Smirnov.
- Click OK. The following table appears:

Kolmogorov-Smirnov X₁: Gender Y₁: Cholesterol

DF	2
male cases	71
female cases	24
Maximum Difference	.156
K-S Chi Square	1.748
Z	.661
	p = .5085

The view title name the X and Y variables. The statistics include:

- degrees of freedom
- number of cases in each group
- maximum difference between the two cumulative distributions
- Kolmogorov-Smirnov Chi-Square
- Z and two-tail probability

Wald-Wolfowitz Runs

The Wald-Wolfowitz runs tests whether two groups within a Y column have been drawn from the same population. The groups in the Y column are specified by an X column. The column assigned as the X variable must be a Category or Integer column and must be suitable as a grouping variable. (Please read the warning at the beginning of the section on ANOVA in this chapter.) It is a OneXOneY statistic. (See the table at the beginning of this chapter.)

This example tests the male and female Weight values in Lipid Data.

- Open Lipid Data. Assign X to Gender.
- Assign Y to Cholesterol.
- Select Nonparametrics from the Compare menu.
- Click Wald-Wolfowitz Runs.
- Click OK. The following table appears:

Vald-Yolfowitz Runs X₁ : Gender Y₁ : Cholesterol

* Runs	36
male cases	71
female cases	24
Mean	36.874
Standard Deviation	3.648
Z	.102
	p = .9184

The view title names the X and Y variables. The statistics include:

- number of runs
- number of cases in each group
- mean
- standard deviation
- Z and two-tail probability

Kruskal-Wallis Test

The Kruskal-Wallis test is a one-way analysis of variance by ranks. It tests whether 3 or more independent groups within a Y column are from different populations. The groups in the Y column are specified by an X column. The column assigned as the X variable must be a Category or Integer column and must be suitable as a grouping variable. (Please read the warning at the beginning of the section on ANOVA in this chapter.)

It is a OneXOneY statistic, however, only one X variable may be assigned. If there are multiple Y variables assigned, a result is calculated for each Y variable.

This example uses Heart History as a grouping (X) column and the Cholesterol values as an independent (Y) column.

- Open Lipid Data. Assign X to Heart History.
- Assign Y to Cholesterol.
- Select Nonparametrics from the Compare menu.
- Click Kruskal-Wallis.
- Click OK. The following table appears:

Kruskal-Wallis X₁ : Heart History Y₁ : Cholesterol

DF	3
* Groups	4
* Cases	95
H	2.087
H corrected for ties	2.088
* tied groups	24
	p = .5546
	p = .5544

The first page contains summary information. The view title names the X and the Y variables. The summary statistics table shows:

- degrees of freedom

- number of different groups
- total number of cases read
- Kruskal-Wallis H statistic and probability

If there are tied groups the table includes:

- Kruskal-Wallis H corrected for ties and probability
- number of tied groups
- Click the down arrow in the scroll bar.

Kruskal-Wallis X₁ : Heart History Y₁ : Cholesterol

Group :	# Cases :	Σ Rank :	Mean Rank :
none	64	2980	46.562
<50	8	373.5	46.688
<60	7	298	42.571
over 60	16	908.5	56.781

The following pages show summary information for each group with a maximum of five groups per page. The information for each group includes:

- group name
- number of cases in that group
- sum of the ranks
- mean rank

Friedman Test

The Friedman test is a two-way analysis of variance by ranks for matched samples. It tests whether 3 or more matched samples, designated by X variables, are from the same population. The number of X variables that can be assigned is limited to 1,150. It is a Many X statistic. (See the table at the beginning of this chapter.)

This example tests four variables from Lipid Data.

- Open Lipid Data.
- Assign X to Weight, Cholesterol, Triglycerides and HDL.
- Select Nonparametrics from the **Compare** menu.
- Click **Friedman**.
- Click **OK**. The following table appears:

Friedman 4 X variables

DF	3
* Samples	4
* Cases	95
Chi-Squared	228.632 p = .0001
Chi corrected for ties	229.114 p = .0001
* tied groups	2

The first page contains summary information. The view title specifies the number of X columns. The summary statistics table shows:

- the degrees of freedom
- the number of samples
- the total number of cases read
- the Friedman Chi-Squared and probability

If there are tied groups the table includes:

- the Chi-Squared corrected for ties and probability
- the # of tied groups
- Click the down arrow in the scroll bar.

Friedman 4 X variables

Name:	Z Rank:	Mean Rank:
Weight	299.5	3.153
Cholesterol	351.5	3.7
Triglycerides	193.5	2.037
HDL	105.5	1.111

The following pages show summary information for each sample with a maximum of five samples per page. The summary information for each sample includes:

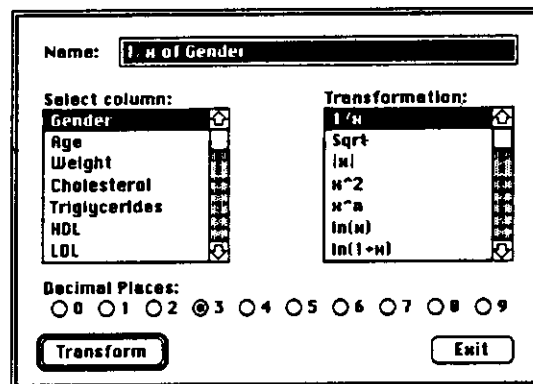
- sample name
- sum of the ranks
- mean rank

Chapter 7 — Advanced Column Creation

This chapter explains five of the commands in the Tools menu that create new columns. These commands allow you to manipulate existing data to create new columns or specify parameters for a new column defined by a series or distribution. The columns created in this process are appended to the right of the dataset. The commands described in this chapter are Transform, Formula, Recode, Series, and Split Columns.

Transform

This command creates a new column which is a transformation of data in an existing column. The transformation can be one of 35 functions listed in the dialog. The Transform dialog looks like:



(Note that StatView also has a Formula command that lets you transform and combine columns in one step. This is described in the next section.)

The column name defaults to "(transformation) of (column)", in this case "1/x of Gender", combining the name of the selected column with the selected transformation. The default name updates as the selections are changed. Of course, you can name the column whatever you want. Once you edit the column name, it will no longer be updated.

The "Select column" list (the list on the left) shows all columns in the dataset except string columns. The "Transformation" list (the list on the right) shows the 35 available transformations. The simple transformations are:

1/x	log(1+x)	tan(x)	sinh(x)
Sqrt	log2(x)	arcsin(x)	cosh(x)
x	e^x	arccos(x)	tanh(x)
x^2	10^x	arctan(x)	arcsinh(x)
ln(x)	2^x	sec(x)	arccosh(x)
ln(1+x)	sin(x)	csc(x)	arctanh(x)
log(x)	cos(x)	cot(x)	

The trigonometric functions require measurements in radians and the inverse trigonometric functions return values in radians. The more complex transformations are:

x^n	"n" is specified in a dialog after you click Transform .
Rank	Compute the rank of each value in the column.
Standard Score	Computed as (value - column mean)/(column standard deviation).
Running Sum	Sum of values from first record up to and including the cell.
Difference	Difference between the cell's original value and that of the cell in the row above.
Percentages	Cell value divided by the sum of the column and multiplied by 100
Lag	Lags the selected column by the number of rows specified in the dialog that appears when you click Transform .
Moving Averages	Displays the mean of a specified number of preceding rows. You select the period of averaging in the dialog that appears when you click Transform . With a period of X specified, X-1 rows at the top of the dataset will contain missing values.

Clicking **Transform** creates a new column which is appended to the right of the dataset. You can scroll to the bottom of the column name list to see that the newly created column name has been added to the list.

The **Transform** dialog does not close when you click **Transform**. You can change the selections in the **Transform** dialog and create additional new columns or click **Exit** to close the dialog box and return to the dataset.

Note: Since StatView requires that each column name be unique, you will get an alert if you try to create a column which has the same name as an existing column.

Formula

The **Formula** command allows you to create a new column that is a transformation of an existing column or that is an arithmetic combination of two columns or transformed columns. You may also add, subtract, multiply or divide a column by a constant or create a column that is the mean or sum of several columns.

The most common use of this command is the creation of a new column that is an arithmetic combination two untransformed columns. For instance, if your dataset has a column for number of men and a column for number of women, you can create a new column (the sum of these two) that is number of adults. In addition,

this command allows you to create a column whose cells are the Mean or Sum of the values in two or more columns. For instance if your dataset has columns that are the results of several samples of a patient's weight, you can create a new column containing the average of those weights.

The Formula dialog looks like:

The column name default is "Column x" when no name is typed in by the user, with "x" being the consecutive number higher than the preceding column.

The list below the column name shows the transformations which can be applied to columns in either of the "Operand" lists. The available transformations are:

None	$\log(x)$	$\tan(x)$	$\cosh(x)$
$1/x$	$\log(1+x)$	$\arcsin(x)$	$\tanh(x)$
Sqrt	$\log_2(x)$	$\arccos(x)$	$\operatorname{arcsinh}(x)$
$ x $	e^x	$\arctan(x)$	$\operatorname{arccosh}(x)$
$x*x$	10^x	$\sec(x)$	$\operatorname{arctanh}(x)$
x^n	2^x	$\csc(x)$	
$\ln(x)$	$\sin(x)$	$\cot(x)$	
$\ln(1+x)$	$\cos(x)$	$\sinh(x)$	

The trigonometric functions require measurements in radians and the inverse trigonometric functions return values in radians. In " x^n ", "n" is specified in a dialog after you click Transform.

The "None" choice in the transformations list indicates that no transformation is applied. You use this choice when you do not want to apply a transformation to either column.

The Operand 1 and Operand 2 lists show the names of all columns in the dataset except string columns. Select a column in one of the Operand lists and click either of Op1 or of Op2 to apply the selected transformation to the column. The heading of the Operand list changes to reflect the assigned transformation.

The mathematical operators define the relationship between two selected columns:

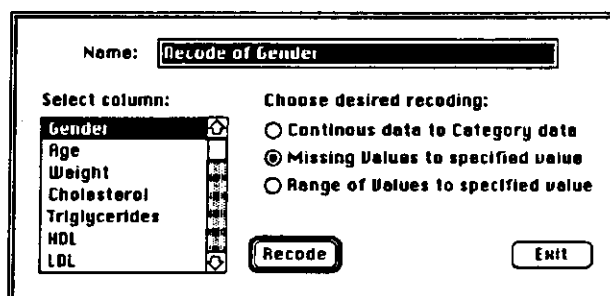
none	Achieves the same effect as the Transform command. The created column contains the data in the selected column, modified by the selected transformation. The Operand 2 list is inactive when this choice is selected.
+ - * ÷	Adds, subtracts, multiplies or divides the columns selected in the Operand lists. Transformations can be applied to either or both of the columns. If you want the second argument of the math to be a constant, fill that constant in the box labelled k. The Operand 2 list is inactive when this choice is selected; this choice is inactive when the Operand 2 list is used.
mean and sum	The created column contains the mean or the sum of the values in two or more selected columns in the Operand 2 list. Drag the cursor down the list to select contiguous columns or use the Command key to select discontinuous columns in the list. The Operand 1 list and the transformations list are inactive when either of these choices are selected.

Click **Calculate** to create a new column. You can scroll to the bottom of the Operand list to see that the newly created column name has been added to the list.

The Formula dialog does not close when you click **Calculate**. You can change the selections in the Formula dialog and create additional new columns or click **Exit** to close the dialog box and return to the dataset.

Recode

The **Recode** command creates a new column by manipulating data in an existing column. You use this command to change continuous data to category form, to convert missing values to a specified value, or to substitute a specified value for a range of values. The Recode dialog looks like:



Select the column to be recoded from the list which contains all the columns in the active dataset except string columns. Select the desired recoding from the three options at the right. Clicking **Recode** displays a dialog pertaining to the selected recode option. Click **Exit** to leave the dialog without creating a column.

Continuous Data to Category Data

Converting Numbers

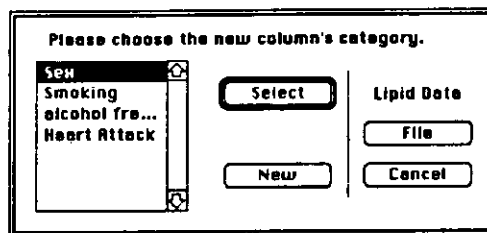
This choice recodes continuous data (real numbers, integers or long integers) into category data. (Chapter 2 describes categories in detail.) An existing category can

be selected or a new category defined. It is also possible to recode a category to different category in the same fashion, as described at the end of this section.

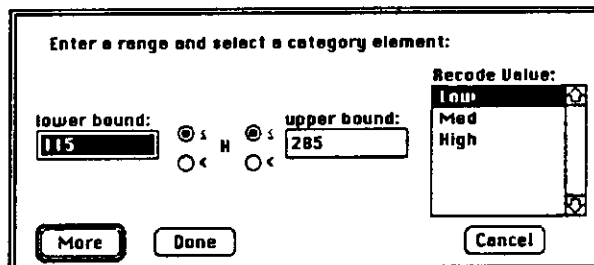
For example the Cholesterol column in Lipid Data contains the measured cholesterol values of the students in a lipid study. You can create a new column which categorizes these values as "low", "medium", and "high" as follows:

- Open Lipid Data.
- Select Recode from the Tools menu.
- Select the Cholesterol column.
- Select Continuous Data to Category Data.
- Click Recode.

The following dialog appears:



Now select a category from the list, choose a category from the library, or click New to create a new category. For this example, create a new category called "Chol levels" that has three elements: "Low," "Medium," and "High." Category creation is described in detail in Chapter 2. The following dialog appears:



This dialog sets the numerical boundaries which define each of the elements. The lower and upper bound values in the dialog are the lowest and highest values in the selected column. This is the range within which divisions are made which correspond to the defined elements (low, medium and high in this case). By clicking the appropriate radio button you can determine whether values can be greater than or equal to or simply greater than the lower bound and less than or equal to or simply less than the upper bound. The value of the lower bound must be greater than the value of the previous upper bound, unless you selected less than upper bound when you entered the previous upper bound.

In the following example, Cholesterol levels between 115 and 130 are recoded to "low"; levels between 131 and 160 are recoded to "medium"; levels greater than 160 are recoded to "high".

- Select the upper bound box and type 130. Make sure Low is selected in the Recode Values list.

- Click **More** or press **Enter**. You have just indicated that values greater than or equal to 115 and less than or equal to 130 will be recoded to **Low**. (Notice that after pressing **More** the second element, **Medium**, is automatically selected and that the lower bound value automatically updates to 130)
- Type 131 in the lower bound box and 160 in the upper bound box.
- Click **More** or press **Enter**. You have just indicated that values greater than or equal to 131 and less than or equal to 160 will be recoded to **Medium**. (The third element, **High**, is now highlighted and that the lower bound value automatically updates to 160)
- Click the **<** button next to the lower bound. You have just indicated that values greater than 160 and less than or equal to 285 will be recoded to **High**.
- Click **Done**.

The recode of Cholesterol is complete; the data appears in a new column in the dataset named "Recode of Cholesterol". Click **Exit** to leave the Recode dialog.

Converting Categories

Category elements are stored internally as ordinal numbers. Thus, you can map elements from one category to another using the **Recode** command similar to the description above. You cannot enter the names of the elements in the upper and lower bounds boxes: you must instead enter their ordinal numbers.

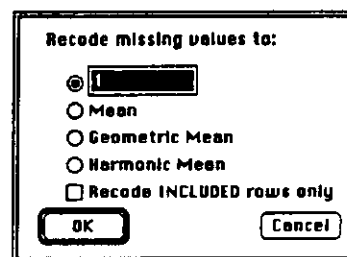
For example you may have a column called **Ratings** which uses a category with five elements (very poor, poor, good, very good, excellent) that you wish to recode to a column called **New Ratings** which contains a category with three elements (poor, good, and excellent). You would use the **Continuous Data to Category Data** dialog box to recode the **Ratings** column. In the dialog box discussed above you would recode values between 1 and 2 (very poor and poor) to the element poor; the values between 3 and 4 (good and very good) to good; and the value 5 to excellent.

Missing Values to Specified Value

This choice replaces missing values in a continuous data column with a specified value or with the column's mean, geometric mean or harmonic mean. The mean values can be computed from all rows in the column or just from *included* rows. (Including and excluding rows is discussed in Chapter 3.) This choice also recodes missing values in a category column to a specified element of the category .

Continuous Data

When a continuous column is selected, clicking **Recode** displays the following dialog:



Set the value to which missing values will recode by selecting one of the Means or by selecting the text entry box and entering a value. Checking **Recode INCLUDED rows only** will cause only missing values located in included rows to be recoded and will also cause the means to be calculated from included values only. When you click OK or Cancel, you return to the Recode dialog.

Category Data

When a category column is selected, clicking Recode displays a dialog containing a list of the elements in that category. For example, if you had selected the Smoking History column in Lipid Data, you would see:

A dialog box titled "Recode missing values to:". It contains a list box with the following items: "no", "quit", "cigarettes", "cigars", and "pipes". Below the list box is a checkbox labeled "Recode INCLUDED rows only". At the bottom are two buttons: "OK" and "Cancel".

Select the element to which missing values will be recoded from the list. Here again the option exists to recode included rows only. When you click OK or Cancel, you return to the Recode dialog.

Range of Values to Specified Value

This choice recodes defined ranges of data to a specified value. The dialog is similar to the one which defines the recoding of continuous data to category data. The following dialog appears:

A dialog box titled "Enter a range and a value to recode to:". It has four main input areas: "lower bound:" with a text box containing "115", "upper bound:" with a text box containing "285", and "Recode to:" with a text box containing "1". Between the lower and upper bound text boxes are two sets of radio buttons. The first set has two options: a selected radio button for " \leq " and an unselected radio button for "<". The second set has two options: an unselected radio button for " \leq " and a selected radio button for "<". At the bottom are three buttons: "More", "Done", and "Cancel".

The lower and upper bounds of each range are specified and paired with the recode value for that range. When a recode value is not specified, the default value begins at 1 and increases by 1 for each successive range that is defined. The range displayed initially in the dialog uses the lowest and highest values in the column as its lower and upper bounds.

You can also use dialog to redefine category data by mapping a group of elements to a single integer. This requires familiarity with the ordinal numbers associated with each element of the category. To recode in this manner, define the upper and lower bounds using the ordinal numbers, and specify the integer associated with each range.

For example, assume that the three elements in a category are "Light," "Medium," and "Dark." and that you want to recode all of the "Medium" values to "Light." In this case, you would enter 1 for the lower bound, 2 for the upper bound, \leq for both operators, and 1 for the recoded value. This recodes both "Light" and "Medium" to "Light."

Series

This option generates a new column whose values are generated from a uniform or normalized random number distribution or from a linear time series. The Series dialog looks like:

Create a new column using specified values:

Name:

Number of values to be created:

Column contains:

☒ Uniform Random ☐ Normalized Random

☐ Time Series Start: Step:

Decimal Places:

☐ 0 ☐ 1 ☐ 2 ☒ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

If the desired number of dataset rows is not entered, the number of values to be created defaults to the number of rows in the dataset. If fewer rows are created, the remaining cells in the column will contain missing values. If the specified number of values is greater than the number of rows in the dataset, all other columns will contain missing values in the added rows.

Three series are available:

- **Uniform Random** generates a column of random numbers greater than zero but less than one.
- **Normalized Random** generates a column of random numbers with a standard deviation that approximates 1 and a mean that approximates 0; this describes a normal curve.
- **Time Series** generates a column that begins at the value specified in the Start box and increases by intervals specified in the Step box.

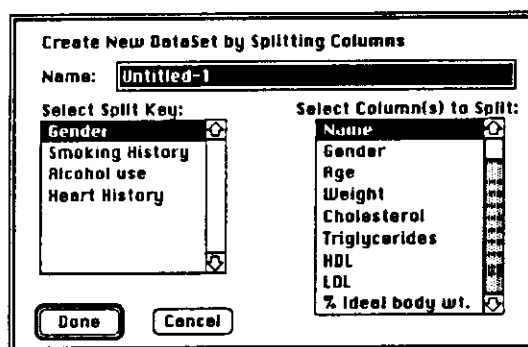
The time series can be used to return a sorted dataset to its original order if the time series is made before the dataset is sorted. Use the default value of 1 for both Start and Step to create a column which has the same numbers as the record numbers. Resorting the dataset, in ascending order, on the time series column will return the dataset to the original order of the records.

The Series command is useful for creating a column with year values without having to enter them by hand. If you are creating a new dataset where one of the columns will be years, use the Series command to create that column and fill in the values for you.

Splitting Columns

The Split Columns command is a convenient way to create a new dataset where values that had been in categories in the original dataset appear in separate columns in the new dataset. For instance, Lipid Data has a Gender column which indicates the gender of the students in the lipid study. You may wish to graphically compare the cholesterol values for males and values. You would use the Split Columns command to create a dataset where the Male cholesterol values appear in one column and the Female cholesterol values appear in another column and then use a Box plot or Error Bar plot to compare the means. (See "Descriptive Comparisons Around the Mean" in Chapter 5 for an example of using split columns in such an analysis.) In addition the Split Columns commands allows you to generate interaction plots (see the example below).

With Lipid Data active the Split Columns dialog looks like:



Note that the name at the top of the dialog is for the new dataset, not for the columns. The Select Split Key list shows the category columns in the original dataset. The Select Column(s) to Split list shows the string, integer, real, and long integer columns in the active dataset. Select one category column from the list on the left and one or more columns from the list on the right.

When you click Done, the new dataset appears in a new data window. Each variable from the Select Column(s) list is now split into as many columns as there are elements in the selected category. If missing values are present in the column to be split, "missing values" will be treated like a distinct category element and a column will be created for these values.

You could use the Split Columns command to prepare an interaction plot. For example, assume that you have an agricultural dataset which examines how nitrogen addition affects crop yield.

- Create a dataset with seven columns: the plot type (a category column with three elements: Strip Plot, Square Plot, and Bunch Plot), crop yield at 0 lb/acre of nitrogen, 50 lb/acre, 100 lb/acre, 150 lb/acre, 200 lb/acre and 250 lb/acre.
- Enter the following data:

	Plot	0 lb.	50 lb.	100 lb.	150 lb.	200 lb.	250 lb.
1	Strip Plot	10.8	13.5	15.6	15.1	16.0	16.5
2	Strip Plot	11.4	14.2	14.7	15.0	14.4	16.4
3	Strip Plot	11.6	14.5	15.6	15.8	15.5	15.7
4	Strip Plot	12.3	15.4	15.4	16.1	15.6	16.6
5	Strip Plot	11.9	14.9	15.6	15.0	16.4	16.0
6	Square Plot	9.2	10.4	12.4	12.5	13.3	13.1
7	Square Plot	9.5	9.6	11.2	11.4	11.1	12.6
8	Square Plot	9.7	11.1	11.1	13.0	13.3	12.2
9	Square Plot	9.7	12.1	10.2	12.2	13.0	12.9
10	Square Plot	8.5	10.6	12.2	11.6	13.0	13.4
11	Bunch Plot	7.5	6.9	7.7	7.5	8.0	7.9
12	Bunch Plot	7.1	6.4	8.3	6.2	6.7	7.6
13	Bunch Plot	6.8	7.2	7.1	8.8	8.0	7.3
14	Bunch Plot	6.8	7.2	6.0	7.3	7.8	7.7
15	Bunch Plot	6.4	6.4	7.3	7.0	7.8	10.1

To prepare an interaction plot which compares the mean values for each plot type at each nitrogen level:

- Select **Split Columns** from the **Tools** menu.
- Type **Split Fertilizer** for the name of the new dataset
- Select **Plot** as the **Select Split Key**.
- Select **0 lb.**, **50 lb.**, **100 lb.**, **150 lb.**, **200 lb.** and **250 lb.** as the columns to split.
- Click **Done**. StatView creates 18 columns, one for each plot/fertilizer pair. Part of the new dataset looks like:

	Strip Plot - 0 lb.	Square Plot - 0 lb.	Bunch Plot - 0 lb.	Strip Plot - 50 lb.	Square Plot - 50 lb.	Bunch Plot - 50 lb.
1	10.8	9.2	7.5	13.5	10.4	
2	11.4	9.5	7.1	14.2	9.6	
3	11.6	9.7	6.8	14.5	11.1	
4	12.3	9.7	6.8	15.4	12.1	
5	11.9	8.5	6.4	14.9	10.6	

- Select all 18 columns in the new dataset.
- Select **Choose X** from the **Variables** menu.
- Select **Mean**, **Std. Dev.**, etc... from the **Describe** menu.
- Select **Table** in the **View** menu. You see:

X1: Strip Plot - 0 lb.					
Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
11.6	.574	.257	.33	4.949	5
Minimum:	Maximum:	Range:	Sum:	Sum of Sqr.:	* Missing:
10.8	12.32	1.52	58	674.118	0

You use the values from this dataset to create the interaction plot dataset.

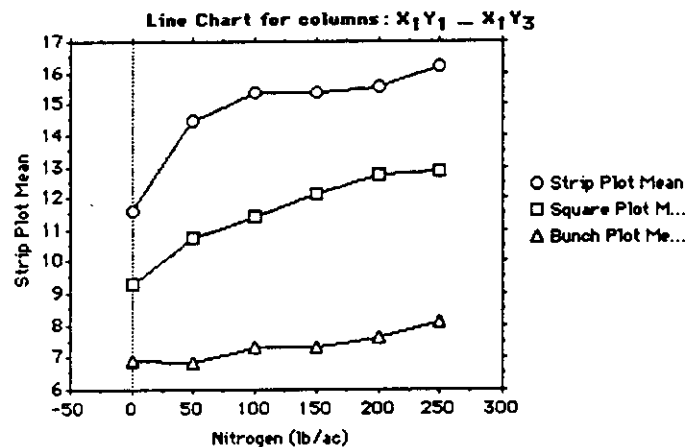
- Select **New** from the **File** menu. (Follow the directions in Chapter 2 to create the new dataset.)


This dataset will have seven columns: Nitrogen (lb/acre), Strip Plot Mean, Strip Plot Error, Square Plot Mean, Square Plot Error, Bunch Plot Mean, and Bunch Plot Error. The Nitrogen (lb/acre) column will be an integer type and the other six columns will be type real. The dataset will have six rows, one for each amount of fertilizer.

- Select the mean and standard deviation cells in the table and select Copy from the Edit menu.
- Select the cell in the dataset and select Paste from the Edit menu. (You can also type in the new values, but copying and pasting is less prone to error.)
- Click the scroll bar to see the next table and repeat the above steps. The new dataset will look like:

	Nitrogen (lb/ac)	Strip Plot Mean	Strip Plot Error	Square Plot Mean	Square Plot Error	Bunch Plot Mean	Bunch Plot Error
1	0	11.600	.574	9.301	.488	6.920	.404
2	50	14.500	.718	10.776	.896	6.826	.433
3	100	15.380	.390	11.400	.896	7.296	.849
4	150	15.400	.515	12.128	.626	7.337	.950
5	200	15.380	.750	12.720	.909	7.632	.345
6	250	16.240	.378	12.864	.457	8.118	1.105



- Assign X to Nitrogen (lb/acre).
- Assign Y to Strip Plot Mean, Square Plot Mean and Bunch Plot Mean, in that order.
- Select Line Chart from the View menu. You see:



- Click the error bar tool . You see:

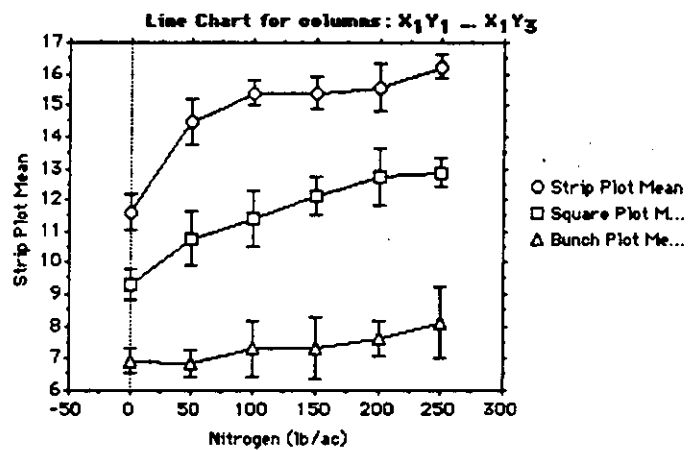
Error Bars for: Strip Plot Mean

☒ No Error bars for this column
☐ Use a fixed error of
☐ Use % of the data value
☐ The selected column contains the error

Nitrogen (lb/ac) 
 Strip Plot Mean
 Strip Plot Error
 Square Plot Mean
 Square Plot Error
 Bunch Plot Mean
 Bunch Plot Error 

- Click The selected column contains the error to indicate that the errors are coming from columns.
- Select Strip Plot Error for the error of the Strip Plot Mean column.
- Click Next.
- Select Square Plot Error for the error of the Square Plot Mean column.
- Click Next.
- Select Bunch Plot Error for the error of the Bunch Plot Mean column.
- Click Done.

The resulting chart shows the interaction plot:



Appendix A — StatView Memory Limits

Dataset Size Considerations

Dataset size in StatView is limited by RAM. The more RAM available, the larger the file size can be.

StatView allocates memory in a dynamic manner allowing datasets with fewer columns to have more rows than datasets with a larger number of columns. Additionally, columns of different types (real, integer, long integer, category, and string) require different amounts of memory. These factors, while they add to the utility of the program, make it difficult to provide hard and fast dataset size limits. The following information, however, can help you plan the size of datasets.

When a column is created, StatView allocates a fixed number of rows for that column whether or not the rows actually exist (have data entered in them). For category columns 1524 rows are allocated; for integer columns 762 rows are allocated; for long integer columns 381 rows are allocated; for string columns 381 rows are allocated; for real columns 127 rows are allocated. When rows exceeding the original allocation are entered, StatView allocates another block as large as the first.

Therefore, 127 rows of real data takes no more memory than 1 row; but 128 rows of real data takes twice as much memory as 127 rows. 254 ($2 * 127$) rows of data takes no more memory than 128 rows, but 255 ($2 * 127 + 1$) rows of real data takes three times as much as 127 rows of real data. 382 ($3 * 127 + 1$) rows of real data take four times as much as 127 rows; 509 ($4 * 127 + 1$) rows of real data take five times as much as 127 rows, and so on.

To increase the amount of RAM available try one or more of the following:

- Do not use MultiFinder.
- If you use MultiFinder, allocate more memory to StatView in the Get Info dialog: Quit the program, select the program, choose **Get Info** from the File menu, and increase the application memory size.
- Close desk accessories.
- Close all StatView datasets other than the one in use.
- Clear the Clipboard by selecting it from the Windows menu and choosing **Delete Clipboard** from the File menu.
- Turn RAM cache off. (In **Control Panel** under the Apple menu.)
- Avoid pasting large color pictures into the view window.

Memory Alert

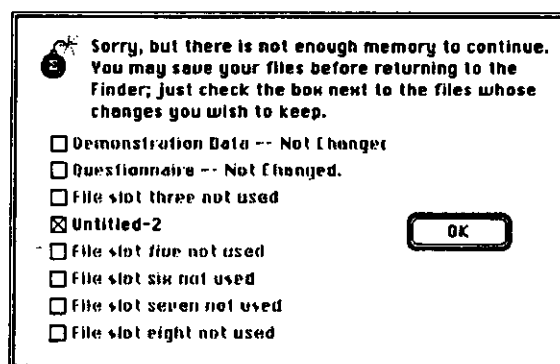
The absolute maximum size a StatView dataset can be is 32,765 rows by 8,192 columns.

A good indication that StatView dataset size is crowding memory is sluggish performance. If editing and statistical operations take unusually long to perform, the program is running out of memory.

Two dialog boxes appear warning that you are approaching memory limits. The first box states "Sorry, but StatView is running low on memory. Please close any unnecessary files and desk accessories." If memory problems persist, a second box appears stating "StatView is running dangerously low on memory. It is imperative that you close any unnecessary files and desk accessories."

If StatView does run out of memory, you are still able to save data files open on the desktop. If there are no files open that have unsaved changes, StatView displays a bomb box that states, "Sorry, but there isn't enough memory to continue. Since you didn't make any changes to your files, clicking OK will end StatView." At this juncture, you have no choice but to click OK and return to the desktop.

If you have made changes to your files, StatView displays this dialog box:



Since there can be eight StatView data files open at any one time, there are eight save-file check boxes in the box. If a file is open, and changes have been made to it, StatView displays the file's name next to a check box. Check the box to save the file. After you click OK, the standard Macintosh name-save file dialog box appears.

If a file was open but no changes had been made to it, its name is displayed in gray next to a check box with the notation "not changed" appended to it.

Appendix B — Formulae and References

Computational Considerations

All calculations are performed in 80 bit extended arithmetic which ensures approximately 18 decimal places of accuracy.

Sum of Squares Calculations

Several StatView statistics require calculation of the sum of squared deviations (or the sum of squares):

$$\sum (X - \bar{x})^2$$

StatView uses an algorithm which provides more accurate results for the sum of squared deviations than the Monroe Calculator variance formula:

$$\sum X^2 - \frac{(\sum X)^2}{n}$$

StatView uses the following algorithm for the sum of squared deviation :

$$\sum (X - k)^2 - n(k - \bar{x})^2$$

where k is the first non-missing, non-excluded value for the variable, and \bar{x} is the calculated variable mean.

In addition, several statistics require that the sum of deviation cross products be calculated:

$$\sum (X - \bar{x})(Y - \bar{y})$$

StatView uses the following algorithm for the sum of deviation cross products:

$$\sum (X - a)(Y - b) - n(a - \bar{x})(b - \bar{y})$$

where a is the first non-missing, non-excluded value for the X variable, b is the first non-missing, non-excluded value for the Y variable, \bar{x} is the X variable mean, and \bar{y} is the Y variable mean.

Descriptive Statistics

Matrix Inversions

Several statistics require matrix inversions. StatView uses the Sweep Operator procedure to invert matrices.

n = number of non-missing, non-excluded values

count = n

Mean = $\frac{\sum X}{n}$ (referred to below as \bar{x} or \bar{y})

Variance (s^2) = $\frac{\sum (X - \bar{x})^2}{n - 1}$

Standard Deviation (s) = $\sqrt{s^2}$

Standard Error of the Mean ($s\bar{x}$) = $\frac{s}{\sqrt{n}}$

Coefficient of Variation = $\frac{100s}{\bar{x}}$

Minimum = smallest value among X

Maximum = largest values among X

Range = Maximum - Minimum

Sum = $\sum X$

Sum of squares = $(\sum X^2)$

missing = Count of the missing values

Confidence interval of the mean: t distribution

error = $t_{a(n-1)} * s_{\bar{x}}$ where $a = 100 - \text{selected \%}$

Confidence interval of the mean: normal distribution

error = $\frac{Z_a * \text{user std. dev.}}{\sqrt{n}}$

where $a = 100 - \text{selected \%}$

Z_a = value such that the probability of normal deviation is less than $1 - \text{selected \%}$

lower = $\bar{x} - \text{error}$

upper = $\bar{x} + \text{error}$

Percentiles: The p th percentile using linear interpolation is:

$(1-f)x_k + f * x_{k+1}$

where x_k and x_{k+1} are the k th and $k+1$ th non-missing, non-excluded values in the column.

Comparative Statistics

k and f are determined from the value v shown below. k is the integer part of v and f is the fractional part:

$$v = \frac{np}{100} + 0.5$$

where n is the count, p is the desired percentile.

$$\text{Geometric Mean} = \sqrt[n]{\pi_x}$$

$$\text{Harmonic Mean} = \left(\frac{\sum \frac{1}{X}}{n} \right)^{-1}$$

$$\text{Kurtosis} = \frac{m_4}{m_2^2} - 3$$

$$\text{Skewness} = \frac{m_3}{m_2 \sqrt{m_2}}$$

Where:

$$m_2 = \frac{\sum (X - \bar{x})^2}{n}, \quad m_3 = \frac{\sum (X - \bar{x})^3}{n}, \quad m_4 = \frac{\sum (X - \bar{x})^4}{n}$$

Comparative Percentiles

For the X and Y column the following percentiles are computed:

1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, 100

See descriptive statistics above for the computation of percentiles.

One Sample t-Test

N = number of x observations

U = population mean, entered by user

$$T = \frac{\bar{x} - U}{\sqrt{\frac{\sum (X - \bar{x})^2 - \frac{(\sum X)^2}{N}}{N(N-1)}}$$

$$DF = N - 1$$

Paired t-Test

N = number of paired x,y observations

$$D_i = x_i - y_i$$

$$T = \frac{\sum D_i}{\sqrt{\frac{\sum (D_i)^2 - \frac{(\sum D)^2}{N}}{N(N-1)}}$$

$$DF = N - 1$$

UnPaired t-Test

N_1 = number of observations in group 1

N_2 = number of observations in group 2

\bar{x}_1 is the mean of the group 1 observations

\bar{x}_2 is the mean of the group 2 observations

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\left[\sum (X_1)^2 - \frac{(\sum X_1)^2}{N_1} \right] + \left[\sum (X_2)^2 - \frac{(\sum X_2)^2}{N_2} \right]}{N_1 + N_2 - 2} \cdot \left(\frac{N_1 + N_2}{N_1 N_2} \right)}$$

where X_1 designates the first group and X_2 the second group.

$$DF = N_1 + N_2 - 2$$

Correlation Coefficient and All Regressions

Covariances and correlations are computed in StatView using provisional means. See the section Computation Considerations in Chapter 2 for details on this procedure.

StatView applies the Sweep Operator to the XX' matrix of cross product deviations in order to calculate regression coefficients. Sweeping operations are discussed in Draper and Smith (1981), Hocking (1985) and Goodnight (1979). The sweeping operation is used to add and delete variables from the regression equation. Beta coefficients, partial correlations, multiple correlation, partial Fs and residual sum of squares are computed as each variable enters (or leaves) the regression equation.

The calculation of confidence bands for the mean and confidence intervals for the slope of a simple regression is discussed in Draper and Smith (1981) and Sokal and Rohlf (1981).

ANOVA

Single factor Factorial model

For a single factor factorial model StatView uses the procedures outlined by Winer (1971) in Chapter Three.

The Model II estimate of between component variance is discussed both by Winer and Afifi and Azen (1979). The formula is as follows:

$$\frac{(\text{Mean Square}_{\text{between groups}} - \text{Mean Square}_{\text{within groups}})(1 - 1)}{\sum_i J_i - \frac{\sum_i J_i^2}{\sum_i J_i} \quad i = 1..I}$$

$$\sum_i J_i - \frac{\sum_i J_i^2}{\sum_i J_i} \quad i = 1..I$$

where J_i is the count of non-missing non-excluded values for the i th group.

Multiple comparisons are discussed in Winer (1971) and Milliken and Johnson (1984). The formulas used are listed below:

For all tests:

k is the number of groups

MD is the mean difference between the group means

MS_r is the between groups mean square

t_a is the two tailed t value at the user entered significance level at the within groups df.

$$r = \frac{1}{N_a} + \frac{1}{N_b}$$

where N_a is the count of group a and N_b is the count of group b

Fisher's Protected Least Significant Difference (PLSD)

$$t_a * \sqrt{r * MS_r}$$

Scheffé F test

$$\frac{\frac{(MD * MD)}{r * MS_r}}{k - 1}$$

Dunnett's t

$$\frac{|MD|}{\sqrt{r * MS_r}}$$

Single factor Repeated Measures model

For a single factor repeated measures model StatView uses the procedures outlined by Winer (1971) in Chapter 4.

For a single factor model reliability estimates are computed (Winer, p. 283). Reliability estimates are given for the mean of all treatments and for a single treatments. The formulae are provided below:

$$\hat{\theta} = \frac{MS_{\text{between groups}} - MS_{\text{within groups}}}{k * MS_{\text{within groups}}}$$

$$\text{mean of all treatments} = \frac{k\hat{\theta}}{1 + k\hat{\theta}}$$

$$\text{mean of single treatment} = \frac{\hat{\theta}}{1 + \hat{\theta}}$$

k = number of treatments (X columns)

Multiple comparisons are discussed above. The calculations are the same except that MS_{residual} is substituted for $MS_{\text{between groups}}$.

Two and more factor Factorial and Repeated Measures models

The technique used for calculating the sums of squares for the various tests reported by StatView is the reduction technique as described by Searle (1971, pp. 246 - 248). The basic idea of the reduction technique is as follows. First a model is fit with all possible main effects and interactions (the full model), and the residual sum of squares, RSS_{full} is calculated. Then for each main effect or interaction to be tested, another model is fit, containing all the terms in the model except the one currently being considered. Once again, the residual sum of squares is calculated. Let the residual sum of squares for the model excluding only effect X (where X is any main effect or interaction in the model) be denoted RSS_X . Then the sum of squares for testing the hypothesis that effect X has no influence on the dependent variable is calculated as:

$$SS_X = RSS_X - RSS_{\text{full}}$$

This calculation is carried out for each term in the model.

The reduction sums of squares are calculated using a method described in detail by Hocking (1985, pp. 146 - 148). First, the matrix $X^T X$ is calculated, using a full rank parameterization for the design matrix X. In this parameterization, the first element of each row of the design matrix is a 1 (for the intercept), and for a main effect with k levels, there are $k-1$ columns in the design matrix. For all but the last level of the factor, a 1 is placed in the column corresponding to the level of that factor for a given observation (row), while for observations with the last level of the factor, all $k-1$ columns are filled with (-1)s. The columns corresponding to interaction terms for a particular row are formed as the Kronecker product of the columns corresponding to all main effects contained in the interaction. Finally, the value of the dependent variable is stored as the last column in the design matrix.

The residual sum of squares for the full model, (SS_{full}) can be found in the lower right hand corner of the matrix $X^T X$ after it has been swept on all of its columns except the last one, i.e. the one corresponding to the dependent variable. (See Goodnight (1979), for a description of the sweep operator). Due to the reversibility of the sweep operator, RSS_X for any effect X can be calculated by re-sweeping the columns corresponding to the effect in question in the fully swept matrix, and extracting the lower right hand element. The sums of squares for each effect are then calculated as described above.

For factorial models with no missing cells, this technique produces sums of squares which generally agree with such programs as BMDP4V or SAS GLM (Type III SS), even for models in which the cell sizes are not equal (unbalanced data). If missing cells are present, then at some point in the sweeping described above, a pivot

element will become too small to allow the sweeping process to continue. By recording the point at which the algorithm fails and then attempting to successfully sweep those columns in the course of resweeping the matrix to calculate the sums of squares for each of the effects, the algorithm described here produces what Hocking (1985) terms *effective hypotheses*. These hypotheses often have fewer degrees of freedom than would otherwise be expected, because they only consider those parts of the hypothesis for which there is sufficient data to calculate the necessary contrasts. For example, in a 2-way analysis with factors A and B each having two levels, if there are no observations for the cell (A=2,B=1), then each of the effects A, B and the A*B interaction will lose one degree of freedom, since, for example, the sum of squares for factor A will be calculated without considering level 2 of A. Essentially, the algorithm described here is attempting to test "as much as possible" of the usual hypotheses which would be tested if there were no missing cells, and ignores those levels of the factors for which any values are missing. Thus, this algorithm may report zero degrees of freedom for some of the effects in the model, whereas other programs may still produce a sum of squares with non-zero degrees of freedom. Programs which do produce sums of squares in these cases are generally testing hypotheses involving weighted averages of the cell means involved in the effect. These hypotheses may or may not be of interest for any particular data set.

Other tests

Goodness of Fit Chi-Square

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where X column contains the Observed Values and Y column the Expected Values

Contingency Table Analysis

N = number of observations

r = # rows of contingency table — determined from the groups of the Y column

c = # columns of contingency table — determined from the groups of the X column

$$DF = (r-1)(c-1)$$

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where:

E = CR/N, the expected values

C = column total

R = row total

O = observed value

N = grand total

$$G \text{ Statistic} = 2 \left[\left[\sum O \ln O \right] - \left[\sum R \ln R \right] - \left[\sum C \ln C \right] + N \ln N \right]$$

$$\text{Contingency Coefficient} = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

$$\text{Phi} = \sqrt{\frac{X^2}{N}}$$

$$\text{Cramer's } V = \sqrt{\frac{X^2}{N(q-1)}}$$

Note: when $r=c=2$, V is the same as Phi where $q = \min(r,c)$.

Chi-Square with continuity correction ($r=c=2$ only)

$$\chi^2 = \frac{N \left| AD - BC - \frac{N}{2} \right|^2}{(A+B)(C+D)(A+C)(B+D)}$$

where:

A=observed value in row1 column1

B=observed value in row1 column2

C=observed value in row2 column1

D=observed value in row2 column2

$$\text{Post-hoc cell contribution} = \frac{O - E}{\sqrt{E \left(1 - \frac{R}{N}\right) \left(1 - \frac{C}{N}\right)}}$$

Mann-Whitney U

n_1 = number of observations in group 1

n_2 = number of observations in group 2

$N = n_1 + n_2$

$R_1 = \sum \text{Rank of first group}$

$R_2 = \sum \text{Rank of second group}$

$$U_1 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$U_2 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$U = \min(U_1, U_2)$

$U' = n_1 n_2 - U$

$$\text{Mean} = \frac{n_1 n_2}{2}$$

$$\text{Standard Deviation} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

$$Z = \frac{U - \text{Mean}}{\text{Standard Deviation}}$$

Correction for Ties:

$$\text{Standard Deviation becomes} = \sqrt{\left[\frac{n_1 n_2}{N(N-1)} \left[\frac{N^3 - N}{12} - \sum T \right] \right]}$$

where $T = \frac{t^3 - t}{12}$ and t is the number of observations ties for a given rank.

Wilcoxon Signed-Rank

$D = X - Y$ for each matched pair

N = number of matched pairs excluding those with a D of zero

R = Rank of $|D|$

$R^+ = \sum R$ with $D > 0$

$R^- = \sum R$ with $D < 0$

$T = R^+$ if $R^+ \leq R^-$ else R^-

$$\text{Mean} = \frac{N(N+1)}{4}$$

$$\text{Standard Deviation} = \sqrt{\frac{N(N+1)(2N+1)}{24}}$$

$$Z = \frac{T - \text{Mean}}{\text{Standard Deviation}}$$

Correction for Ties:

$$\text{Standard Deviation} = \sqrt{\frac{N(N+1)(2N+1) - \frac{\sum T}{2}}{24}}$$

Where $T = t^3 - t$ and t is the number of observations tied for a given rank.

Spearman Rank Correlation Coefficient

N = number of matched pairs

R_x = Rank of X_i

R_y = Rank of Y_i

$D = R_x - R_y$ for each matched pair

$$\text{Rho } (\rho) = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

$$Z = \rho \sqrt{N-1}$$

Correction for Ties:

$$\text{Rho } (\rho) \text{ becomes} = \frac{\sum x^2 + \sum y^2 - \sum D^2}{2\sqrt{(\sum x^2)(\sum y^2)}}$$

$$\sum x^2 = \frac{N^3 - N}{12} - \sum T_x$$

$$\sum y^2 = \frac{N^3 - N}{12} - \sum T_y$$

$$T_x = \frac{t^3 - t}{12}$$

where t is the number of X observations tied for a given rank.

$$T_y = \frac{t^3 - t}{12}$$

where t is the number of Y observations tied for a given rank.

Kendall Correlation Coefficient

N = number of matched pairs

C = Kendall Statistic determined as follows:

Rank the observations on the X variable from 1 to N . Rank the observations on the Y variable from 1 to N . Arrange the list of N subjects so that the X ranks of the subjects are in their natural order, i.e. 1, 2, 3, ..., N . For each Y rank, count the number of ranks below it which are larger. Then subtract the number of ranks below it which are smaller. The Sum of this for each Y is C .

$$t = \frac{C}{\frac{1}{2}N(N-1)}$$

$D = R_x - R_y$ for each matched pair

$$\text{Standard Deviation} = \sqrt{\frac{2(2N+5)}{9N(N-1)}}$$

$$z = \frac{t}{\text{Standard Deviation}}$$

Correction for Ties:

$$t \text{ becomes } \frac{C}{\sqrt{\left[\frac{1}{2}N(N-1) - \sum T_x\right] \left[\frac{1}{2}N(N-1) - \sum T_y\right]}}$$

$$T_x = \frac{t^2 - t}{2}$$

where t is the number of X observations tied for a given rank

$$T_y = \frac{t^2 - t}{2}$$

where t is the number of Y observations tied for a given rank

Kolmogorov-Smirnov

See Siegel pp. 127-136 and Hollander.

Wald-Wolfowitz Runs Test

N_1 = number of group 1 observations.

N_2 = number of group 2 observations.

R = # of Runs. A run is any sequence of scores from the same column.

$$\text{Mean} = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$\text{Std. Deviation} = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$

$$Z = \frac{|R - \text{Mean}| - .5}{\text{Std. Deviation}}$$

Note that there are no correction for ties. Ties may invalidate the results.

Kruskal-Wallis Test

k = number of groups

n_j = number of cases in j^{th} group

$N = \sum n_j$, the number of cases in all groups combined

R_i = sum of ranks in the j^{th} group

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1)$$

Correction for Ties:

$$H \text{ becomes } \frac{\frac{12}{N(N+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1)}{1 - \frac{\sum T}{N^3 - N}}$$

where:

$T = t^3 - t$ (when t is the number of tied observations in a tied group of scores) and $\sum T$ directs on to sum over all groups of ties.

Friedman Test

k = number of X columns

N = number of rows

$R_i = \sum R$ for each column where R is the score ranked by row, $i=1 \dots k$

$$\chi_r^2 = \left[\frac{12}{Nk(k+1)} [\sum R_i^2] \right] - 3N(k+1)$$

Correction for Ties:

$$\chi_r^2 = \frac{12 \sum (R_i - N(\frac{k+1}{2}))^2}{Nk(k+1)} - \frac{\sum T}{k-1}$$

References

- Afifi, A. and Azen, S. (1979). *Statistical Analysis: A Computer Oriented Approach*. Academic Press, New York.
- Chambers, John M. Cleveland, William S. Kleiner, Beat Tukey, Paul A. (1983) *Graphical Methods for Data Analysis*. Wadsworth Statistics/Probability Series, Belmont, California.
- Cleveland, William S. (1985). *The Elements of Graphing Data*. Wadsworth Advanced Book Program, Monterey, California.
- Draper, N. and Smith, H. (1981) *Applied Regression Analysis, Second Edition*. John Wiley & Sons, New York, New York.
- Everitt, B.S. (1977) *The Analysis of Contingency Tables*. Chapman and Hall Ltd. London.
- Goodnight, J. H. (1979) "A Tutorial on the SWEEP Operator," *The American Statistician*, 33, 149 - 158.
- Hocking, R. R. (1985). *The Analysis of Linear Models*. Brooks/Cole, Monterey, California.
- Hollander, M. and Wolfe, D. (1973). *Nonparametric Statistical Methods*. John Wiley & Sons, New York.
- Kendall, M. and Stuart, A. (1977) *Volume 1: Advanced Theory of Statistics*. Charles Griffin & Company, London.
- Kleinbaum, D.G. & Kupper, L.L. (1978) *Applied Regression Analysis and Other Multivariate Methods*. Duxbury Press, Wadsworth Publishing Company, Belmont, California.
- Milliken, G. A. and Johnson, D. E. (1984). *Analysis of Messy Data Volume 1: Designed Experiments*. Lifetime Learning Publications, Belmont, California.
- Montgomery, D. & Peck, E. (1982). *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York.
- Searle, S.R. (1971), *Linear Models*. John Wiley & Sons, New York
- Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.
- Simpson, G.G., Roe, A. and Lewontin, R.C. (1960) *Quantitative Zoology*. Revised Edition. Harcourt, Brace & Co.. New York
- Snedecor, G. and Cochran, W. (1980). *Statistical Methods*. Iowa State University Press, Ames, Iowa.
- Sokal, Robert R. and Rohlf, F. James. (1981) *Biometry*. W. H. Freeman and Company, New York, New York.
- Winer, B. J. (1971). *Statistical Principles in Experimental Design*. McGraw-Hill, New York, New York.

Factor Analysis

- Armstrong, J.S. and Soelberg, P. "On the Interpretation of Factor Analysis." *Psychological Bulletin*: 70(5):361, 1968
- Bartlett, M.S. "A Further Note on Tests of Significance in Factor Analysis." *British Journal of Psychology*: 4(1):1, 1951
- Carroll, J.B. "Approximating Simple Structure in Factor Analysis." *Psychometrika*:18:23, 1953
- Cattell, R.B. and Jaspers, J.A. "A General Plasmode (No. 30-10-5-2) for Factor Analytic Exercises and Research." *Multivariate Behavioral Research Monographs*:67(3): 1967,211 pages
- Cattell, R.B. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research*:1(2):245, 1966
- Gorsuch, R. *Factor Analysis*. Lawrence Erlbaum Publishers, Hillsdale,N.J., 1983
- Guttman, L. "Some Necessary Conditions for Factor Analysis." *Psychometrika*: 19: 149, 1954
- Harman, H. *Modern Factor Analysis (3rd edition)*. Chicago, University of Chicago Press: 1976
- Harris, C.W. "Some Rao-Guttman Relationships." *Psychometrika*: 27:247, 1962
- Harris,C.W. "On Factors and Factor Scores." *Psychometrika*: 32:363, 1967
- Hofmann, R.J. "Brief Report: On the Proportionate Contributions of Transformed Factors to Common Variance." *Multivariate Behavioral Research*: 10(4):507, 1975
- Hofmann, R.J. "Complexity and Simplicity as Objective Indices Descriptive of Factor Solutions." *Multivariate Behavioral Research*: 13(1):247, 1978-b
- Hofmann, R.J. "Indices Descriptive of Factor Complexity." *The Journal of Psychology*: 96:103, 107
- Hofmann, R.J. "The Orthotran Solution." *Multivariate Behavioral Research*: 13(1):99, 1978-a
- Hotelling, H. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology*: 24: 417 and 498, 1933
- Kaiser, H.F. "A Second Generation Little Jiffy." *Psychometrika*: 35:401, 1970
- Kaiser, H.F. "Psychometric Approaches to Factor Analysis." *Proceedings of the 1964 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 37, 1965
- Kaiser, H.F. "The Varimax Criterion for Varimax Rotation in Factor Analysis." *Psychometrika*: 23: 187, 1958
- Mulaik, S. *The Foundations of Factor Analysis*; New York: McGraw Hill, 1972
- Saunders, D.R. "Trans-Varimax." *American Psychologist*. 17:395, 1962

Thurstone, L.L. *Multiple Factor Analysis*. Chicago: University of Chicago Press, 1947

Timm, N. *Multivariate Analysis with Applications in Education and Psychology*, New York:Brooks/Cole, 1975

Wilkenson, J.H. *The Algebraic Eigenvalue Problem*, London: Oxford University Press, 1965.

Suggested Reading

Snedecor, George W. and Cochran, William G. *Statistical Methods*, Ames: Iowa State University Press, 1989.

Steel, Robert G. D. and Torrie, James H. *Principles and Procedures of Statistics: a Biometrical Approach*, New York : McGraw-Hill, 1980.

Index

- 50th percentile 103
- 68881 3
- 68882 3
- abscissa 107
- adding columns 24
- Align to Grid command 91
- altering datasets 24
- ANOVA 177
 - assigning variables 178
 - capabilities 177
 - Dunnett t test 177
 - regression 146
 - single factor 178
 - single factor one repeated measure 184
 - three factor one repeated measure 187
 - two factor balanced 181
 - two factor unbalanced 183
- arccos(x) 203
- arccosh(x) 203
- arcsin(x) 202, 203
- arcsinh(x) 203
- arctan(x) 203
- arctanh(x) 203
- Arrow Head command 17, 93
- ASCII files 4, 34
- assigning variables 43
- axes 93
 - bounds 94
 - changing 93
 - tick marks 94
- balanced design 177, 181, 182
- bar chart 66
 - comparative 68
 - histogram 67, 126
 - percentiles 117
 - univariate 57
 - with error bars 70
 - z-score distribution 110
 - z-score histogram 68
- bar control 70, 81
- Bartlett Test of Sphericity 168
- beta coefficient table 143, 147, 151
- bounds 94
- box plot 71, 111, 123
 - notched 72, 81, 112
- calculation 50
- categories 26
 - as grouping variables 178
 - converting 206
 - converting to 204
 - editing 28
 - new 27
 - specifying 22
- category columns 22
- cellulation 75
- Center Justify command 91
- central tendency 103
- changing columns 24
- characteristic roots 163
- Chi-Square 189
- Choose X command 11, 43, 44
- Choose Y command 43, 44
- Chooser desk accessory 42
- Clear command 31, 88
- Clear Range command 51
- Clear X&Y command 43, 44
- Clipboard 4, 7, 31, 39
- Close command 10
- coefficient of variation 104
- coincidence 75
- color 3, 85, 86
- Color command 91, 93
- Color Picker 86
- column percents 190

- column totals table 192
- column type 22
 - category 22
 - long 22
 - real 22
 - string 22
- columns 7, 22, 43
 - adding 24
 - changing 24
 - changing type 25
 - inserting 24
 - removing 24
 - selecting 11
- Command key 44
- communality summary 170
- Comp menu 3
- comparative bar chart 68
- Compare menu 3, 8, 133
- Compare Percentiles command 64, 133
 - graphic view 134
 - table view 134
- comparison percentile chart 63
- composite mode 82
- confidence bands 62, 80, 148
- confidence intervals 108
- Confidence Intervals command 12
- Contingency Coefficient 189
- Contingency table 189
 - chi-square 190
 - column totals 192
 - contingency coefficient 189
 - continuity correction 189
 - Cramer's V 189
 - expected values 192
 - G statistic 189
 - observed frequency table 191
 - Phi 189
 - post hoc cell contributions 192
 - row totals 191
 - tabulating 190
- Continuity Correction 189
- continuous data 204
- Copy command 40, 88
- Copy View command 40, 89
- copying 32
 - columns and rows 30
 - graphs 89
 - values 89
 - views 40
- correlation coefficient 139
- correlation matrix 141, 162, 167
- cos(x) 202, 203
- cosh(x) 203
- cot(x) 202, 203
- Cramer's V 189
- csc(x) 202, 203
- cumulative frequency curve 58, 113
- currency marks 36, 38
- curvilinear regression 152
- Custom Rulers command 90
- customizations 96
- Cut command 31, 88
- data window 21
- datasets 7, 8, 21, 22
 - altering 24
 - creating 23
 - opening 23
- decimal places 45
- Delete Clipboard command 32
- Delete command 24, 32
- Desc menu 3
- Describe menu 3, 7, 101
- direct proportionate contribution 174
- drawing objects 92
- drawing tools 55
- Dunnett t-test 177, 181, 186
- Edit Categories command 28
- Edit Palette command 86
- Edit Range command 51
- eigenvalues 163, 164, 165, 169
- eigenvectors 170
- equal axes control 64, 81
- equamax 162, 165, 171
- error bar tool 211
- error bars 70, 107, 110, 119
- error bars control 79
- Excel 7, 38
- excluding rows 50
- expected values 190
- expected values table 192

- exporting 39
- factor analysis 162
 - Bartlett Test of Sphericity 168
 - eigenvalues 163, 164, 169
 - eigenvectors 170
 - factor extraction method 163
 - factor loadings 164
 - factor scores 166
 - oblique solution 176
 - orthogonally rotated solution 176
 - partial correlation 167
 - principal components 164
 - unrotated solution 175
- factor scores. 166
- factoranalysis
 - intercorrelation matrix 174
- factorial model 177
- Fill command 93
- fills 95
- final communality estimate 170
- Fisher's PLSD test 177, 181, 186
- fitted values 144, 157, 158
- floating-point math coprocessor 2
- Font command 91
- Format command 24
- Formula command 202
- formulae 215
- FPU 2
- frame control 74
- frequency distribution 125
 - histogram 67
 - pie chart 69
- Friedman Test 199
- full interaction model 177
- G Statistic 189
- geometric mean 106
- Graph menu 93
- graphing data 46
- graphs 13, 56
 - copying 89
 - resizing 84
- grid 89
- grid line 94
- grouping variable 178
- harmonic mean 107
- Harris image analysis 162, 164
- Heywood case 171
- Hide Legend command 88, 95
- histogram 67, 128
 - z-score 68
- Horizontal Legend command 95
- HOW method 169
- Import command 35
- importing
 - ASCII files 4, 34
- including rows 50
- initial factoring procedure 163
- input row 10, 29
- inserting columns 24
- inserting rows 25
- installation 7
- integer
 - as grouping variable 178
- integercolumns 22
- interaction plot 209
- intercept 144
- inverting matrices 216
- joint proportionate contribution 174
- Kaiser image analysis 162, 164
- Kendall rank correlation coefficient 196
- keyboard 30
- Kolmogorov-Smirnov Tests 196
- Kruskal-Wallis Test 198
- kurtosis 104
- lag transformation 202
- large screen 3
- LaserWriter 42
- latent roots 164
- Left Justify command 91
- legend 95
- leptokurtic 104
- line chart 64
 - back cover example 210
 - compare percentiles 64
 - error bars 70
 - subset specify 78
- Line Width command 93
- Lipid Data 5
- ln(x) 202, 203
- log(x) 202, 203

- log2(x) 202, 203
- long columns 22
- MacDraw 4
- MacWrite 4
- Mann-Whitney U 193
- matrix inversion 216
- maximum 103
- mean 58, 103, 117
- Mean, Std. Dev., etc... command 12, 210
- median 102, 103, 106
- memory alerts 214
- memory limits 213
- mesokurtic 104
- Microsoft Word 4
- minimum 103
- missing cells 177
- missing values 29, 49, 206
- mode 103
- Model estimate of between component variance 180
- monochrome monitor 86
- mouse 30
- moving averages transformation 202
- multiple linear regression 157
- multiple regression 149
- multiway repeated measures 187
- negatively-skewed distribution 116
- New Column command 24, 29
- New command 8, 23
- no symbols control 66, 82
- non-color 3
- None command 47
- nonparametrics 192
 - Friedman test 199
 - grouping variables 178
 - Kendall rank correlation coefficient 196
 - Kolmogorov-Smirnov tests 196
 - Kruskal-Wallis test 198
 - Mann-Whitney U 193
 - Spearman rank correlation coefficient 195
 - Wald-Wolfowitz runs 197
 - Wilcoxon Signed-Rank 194
- normal distribution 115
- notch control 72, 81, 112, 124
- numeric co-processor 3
- objects 92
- oblique solution 172, 176
- old QuickDraw 3
- one group t-Test 136
- Open Axis command 89, 93
- Open command 11, 23
- opening datasets 23
- Option-8 29
- ordinate 107
- orthogonally rotated solution 176
- orthotran 166
- outlier control 73, 81
- overlap control 74
- Page Setup command 42
- PageMaker 4
- paging mode 82
- paging tool 82
- paired two group t-Test 137
- palette 55, 73
- partial correlation 167
- partial F-test 151
- Paste command 32, 88
- Paste Transposed command 32
- patterns 95
- Pearson's correlation coefficient 139, 143
- Pen command 93
- percentages transformation 202
- percentile plot
 - comparison percentile chart 63
 - cumulative frequency curve 58
 - percentile control 60, 81
- percentiles 105, 111, 113, 133
- Percentiles command 58, 71
- Phi 189
- PICT 4, 40
- PICT format 89
- pictures 40
- pie chart 69
- Pie Chart command 129
- Pixel Paint 4
- platykurtic 104, 115
- plotting symbols 82
- point overlap 74
 - Bigger Points 75
 - Sunflowers 75
- Point Size command 93

- point type 85
- Point Type command 93
- polynomial regression 61, 62, 151
- positively-skewed distribution 115
- post hoc cell contributions 189, 190, 192
- predicted values 144, 157, 158
- Preferences command 22, 31, 45, 84
- primary pattern solution 172
- principal components analysis 162, 164
- Print command 41
- printing 41
- quartimax 162, 165, 172
- Quick Assignment command 43, 44
- QuickDraw colors 86
- radians 202, 203
- RAM 2
- ranges 50
- real columns 22
- Recode command 26, 29, 204
- rectangle tool 92
- reference structure solution 172
- References 226
- regression 143
 - ANOVA table 143, 146, 150, 160
 - beta coefficient table 143, 147, 151
 - confidence interval table 143
 - multiple 149
 - partial F-test 151
 - polynomial 151
 - predicted values 143
 - residual statistics 143
 - simple 145
 - stepwise 157
 - summary statistics 145
- Regression command 14, 61
- removing columns 24
- repeated measures model 177
- residuals 144, 157, 158
- resizing
 - legend 95
 - objects 88, 92
 - view window 84
- resolution 75
- Right Justify command 91
- Rotate Left command 91
- Rotate Right command 91
- Row percents 190
- row totals table 191
- rows 7, 23
 - excluding and including 50
 - inserting 25
- ruler 89
- running sum transformation 202
- Save As command 39
- scattergram 60
 - compare percentiles 64
 - confidence bands 80
 - correlation coefficient 141
 - error bars 70, 79
 - regression 61, 152
 - standard deviation error bars 107
 - subset specify 78
 - univariate 108, 120
- Scheffé F-test 177, 181, 186
- sec(x) 202, 203
- Select All Columns command 30
- Select All command 40, 88
- Select All Rows command 30
- Select Background command 88
- Select Range command 51
- selecting 30, 87
 - columns 11
- Series command 208
- Show Legend command 95
- Show Rulers command 89
- Show Selection command 30
- sigmoid 114
- Simple linear regression 61
- simple regression 145
- sin(x) 202, 203
- sinh(x) 203
- Size command 91
- skewness 103, 104
- small screen 3
- SMC 171
- sorting 41, 208
- Spearman rank correlation coefficient 195
- Split Columns command 117, 118, 209
- Sqrt 202, 203
- squared multiple correlations 171

- standard deviation 58, 103, 107
- standard deviation control 58, 80, 108
- standard error 106
- standard score transformation 202
- standardized residuals 157
- statistics texts, recommended 228
- stepwise regression 157
- string columns 22
- Style command 91
- subscript 43
- subset specify control 78
- sum of squares 215
- SuperANOVA 144, 157, 177
- symbols 82
- system requirements 2
- t-Test 135
 - one group 136
 - paired two group 137
 - unpaired two group 138
- tabulated data 191
- tan(x) 202, 203
- tanh(x) 203
- technical support 5
- text
 - editing 91
 - resizing 91
- text files 34, 39
- Text menu 16, 17
- tick marks 94
- tools palette 55, 56, 73
- Transform command 18, 201
- transformation method 165
- trigonometric functions 202, 203
- Turn Grid On command 90
- two group t-Test 137
- unbalanced design 183
- univariate chart 56
 - bar chart 57
 - line chart 57
 - scattergram 56, 108
- unpaired two group t-Test 138
- unrotated factor matrix 170
- unrotated solution 175
- variable complexity 173
- variable sampling adequacy 168

- variables
 - assigning 43, 47
 - distinguish by color 85
 - distinguish by point type 85
 - grouping 178
- Variables menu 3
- variance 103
- varimax 162, 165, 172
- Vars menu 3
- view controls 55, 73
- View menu 8, 45
- view window 8, 88
 - resizing 84
- Wald-Wolfowitz Runs 197
- Wilcoxon Signed-Rank 194
- Window menu 3, 21
- yin-yang 50
- z-score 122
 - distribution 110
 - histogram 68, 111
- zero line 94
- Zoom Up command 21, 46
- Σ menu 3