SAS® Press Series

JMP® Start Statistics

A Guide to Statistics and Data Analysis Using JMP®

Fourth Edition

John Sall
Lee Creighton
Ann Lehman

JMP® Start Statistics

A Guide to Statistics and Data Analysis Using JMP®

Fourth Edition

John Sall
Lee Creighton
Ann Lehman

THE POWER TO KNOW®

The correct bibliographic citation for this manual is as follows: Sall, John, Lee Creighton, and Ann Lehman. 2007. *JMP® Start Statistics: A Guide to Statistics and Data Analysis Using JMP®, Fourth Edition*. Cary, NC: SAS Institute Inc.

JMP® Start Statistics: A Guide to Statistics and Data Analysis Using JMP®, Fourth Edition

ISBN 978-1-59994-572-9

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, September 2007

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at support.sas.com/pubs or call 1-800-727-3228.

Table of Contents

Preface

JMP® is statistical discovery software. JMP helps you explore data, fit models, discover patterns, and discover points that don't fit patterns. This book is a guide to statistics using JMP.

The Software

The emphasis of JMP as statistical discovery software is to interactively work with data and graphics in a progressive structure to make discoveries.

- With graphics, you are more likely to make discoveries. You are also more likely to understand the results.
- With interactivity, you are encouraged to dig deeper and try out more things that might improve your chances of discovering something important. With interactivity, one analysis leads to a refinement, and one discovery leads to another discovery.
- With a progressive structure, you build a context that maintains a live analysis. You don't have to redo analyses and plots to make changes in them, so details come to attention at the right time.

Software's job is to create a virtual workplace. The software has facilities and platforms where the tools are located and the work is performed. JMP provides the workplace that we think is best for the job of analyzing data. With the right software workplace, researchers embrace computers and statistics, rather than avoid them.

JMP aims to present a graph with every statistic. You should always see the analysis in both ways, with statistical text and graphics, without having to ask for it. The text and graphs stay together.

JMP is controlled largely through point-and-click mouse manipulation. If you hover the mouse over a point, JMP identifies it. If you click on a point in a plot, JMP highlights the point in the plot, and highlights the point in the data table. In fact, JMP highlights the point everywhere it is represented.

JMP has a progressive organization. You begin with a simple report (sometimes called a *report surface* or simply *surface*) at the top, and as you analyze, more and more depth is revealed. The analysis is alive, and as you dig deeper into the data, more and more options are offered according to the context of the analysis.

In JMP, completeness is not measured by the "feature count," but by the range of possible applications, and the orthogonality of the tools. In JMP, you get a feeling of being in more control despite less awareness of the control surface. You also get a feeling that statistics is an orderly discipline that makes sense, rather than an unorganized collection of methods.

A statistical software package is often the point of entry into the practice of statistics. JMP strives to offer fulfillment rather than frustration, empowerment rather than intimidation.

If you give someone a large truck, they will find someone to drive it for them. But if you give them a sports car, they will learn to drive it themselves. Believe that statistics can be interesting and reachable so that people will want to drive that vehicle.

JMP Start Statistics, Fourth Edition

Many changes have been made since the third edition of *JMP Start Statistics*. Based on comments and suggestions by teachers, students, and other users, we have expanded and enhanced the book, hopefully to make it more informative and useful.

JMP Start Statistics has been updated and revised to feature JMP 7. Major enhancements have been made to the product, including new platforms for design (Split Plots, Computer Designs), analysis (Generalized Linear Models, Time Series, Gaussian Processes), and graphics (Tree Maps, Bubble Plots) as well as more report options (such as the Tabulate platform, Data Filter, Phase and T^2 control charts) unavailable in previous versions. The chapter on Design of Experiments (DOE) has been completely rewritten to reflect the popularity and utility of optimal designs. In addition, JMP has a new interface to SAS that makes using the products together much easier.

JMP 7 also focuses on enhancing the user experience with the product. Tutorials, Did you know tips, and an extensive use of tool tips on menus and reports make using JMP easier than ever.

Building on the comments from teachers on the third edition, chapters have been rearranged to streamline their pedagogy, and new sections and chapters have been added where needed.

SAS

JMP is a product from SAS, a large private research institution specializing in data analysis software. The company's principal commercial product is the SAS System, a large software system that performs much of the world's large-scale statistical data processing. JMP is positioned as the small personal analysis tool, involving a much smaller investment than the SAS System.

This Book

Software Manual and Statistics Text

This book is a mix of software manual and statistics text. It is designed to be a complete and orderly introduction to analyzing data. It is a teaching text, but is especially useful when used in conjunction with a standard statistical textbook.

Not Just the Basics

A few of the techniques in this book are not found in most introductory statistics courses, but are accessible in basic form using JMP. These techniques include logistic regression, correspondence analysis, principal components with biplots, leverage plots, and density estimation. All these techniques are used in the service of understanding other, more basic methods. Where appropriate, supplemental material is labeled as "Special Topics" so that it is recognized as optional material that is not on the main track.

JMP also includes several advanced methods not covered in this book, such as nonlinear regression, multivariate analysis of variance, and some advanced design of experiments capabilities. If you are planning to use these features extensively, it is recommended that you refer to the help system or the documentation for the professional version of JMP (included on the JMP CD or at http://www.jmp.com).

Examples Both Real and Simulated

Most examples are real-world applications. A few simulations are included too, so that the difference between a true value and its estimate can be discussed, along with the variability in the estimates. Some examples are unusual, calculated to surprise you in the service of emphasizing an important concept. The data for the examples are installed with JMP, with

step-by-step instructions in the text. The same data are also available on the internet at www.jmp.com. JMP can also import data from files distributed with other textbooks. See Chapter 3, "Data Tables, Reports, and Scripts" for details on importing various kinds of data.

Acknowledgments

Thank you to the testers for JMP and the reviewers of *JMP Start Statistics*: Michael Benson, Avignor Cahaner, Howard Yetter, David Ikle, Robert Stine, Andy Mauromoustkos, Al Best, Jacques Goupy, and Chris Olsen. Further acknowledgements for JMP are in the JMP documentation on the installation CD.

1 Preliminaries

What You Need to Know

…about your computer

Before you begin using JMP, you should be familiar with standard operations and terminology such as click, double-click, ⌘-click, and option-click on the Macintosh (Control-click and Alt-click under Windows or Linux), shift-click, drag, select, copy, and paste. You should also know how to use menu bars and scroll bars, move and resize windows, and open and save files. If you are using your computer for the first time, consult the reference guides that came with it for more information.

…about statistics

This book is designed to help you learn about statistics. Even though JMP has many advanced features, you do not need a background of formal statistical training to use it. All analysis platforms include graphical displays with options that help you review and interpret the results. Each platform also includes access to help that offers general help and appropriate statistical details.

Learning About JMP

…on your own with JMP Help

If you are familiar with Macintosh, Microsoft Windows, or Linux software, you may want to proceed on your own. After you install JMP, you can open any of the JMP sample data files and experiment with analysis tools. Help is available for most menus, options, and reports.

There are several ways to access JMP Help:

- If you are using Microsoft Windows, help in typical Windows format is available under the **Help** menu on the main menu bar.
- On the Macintosh, select **JMP Help** from the help menu.
- On Linux, select an item from the **Help** menu.
- You can click the **Help** button from launch dialogs whenever you launch an analysis or graph platform.
- After you generate a report, select the help tool (?) from the **Tools** menu or toolbar and click the report surface. Context-sensitive help tells about the items that you click on.

…hands-on examples

This book, *JMP Start Statistics*, describes JMP features, and is reinforced with hands-on examples. By following along with these step-by-step examples, you can quickly become familiar with JMP menus, options, and report windows.

Mouse-along steps for example analyses begin with the mouse symbol in the margin, like this paragraph.

…using Tutorials

Tutorials interactively guide you through some common tasks in JMP, and are accessible from the **Help > Tutorials** menu. We recommend that you complete the Beginner's tutorial as a quick introduction to the report features found in JMP.

…reading about JMP

The professional version of JMP is accompanied by five books—the *JMP Introductory Guide*, the *JMP User Guide*, *JMP Design of Experiments*, the *JMP Statistics and Graphics Guide*, and the *JMP Scripting Guide*. These references cover all the commands and options in JMP and have extensive examples of the **Analyze** and **Graph** menus. These books may be available in printed form from your department, computer lab, or library. They were installed as PDF files when you first installed JMP.

Chapter Organization

This book contains chapters of documentation supported by guided actions you can take to become familiar with the JMP product. It is divided into two parts:

The first five chapters get you quickly started with information about JMP tables, how to use the JMP formula editor, and give an overview of how to obtain results from the **Analyze** and **Graph** menus.

- Chapter 1, "Preliminaries," is this introductory material.
- Chapter 2, "JMP Right In," tells you how to start and stop JMP, how to open data tables, and takes you on a short guided tour. You are introduced to the general personality of JMP. You will see how data is handled by JMP. There is an overview of all analysis and graph commands, information about how to navigate a platform of results, and a description of the tools and options available for all analyses. The Help system is covered in detail.
- Chapter 3, "Data Tables, Reports, and Scripts," focuses on using the JMP data table. It shows how to create tables, subset, sort, and manipulate them with built-in menu commands, and how to get data and results out of JMP and into a report.
- Chapter 4, "Formula Editor Adventures," covers the formula editor. There is a description of the formula editor components and overview of the extensive functions available for calculating column values.
- Chapter 5, "What Are Statistics?" gives you some things to ponder about the nature and use of statistics. It also attempts to dispel statistical fears and phobias that are prevalent among students and professionals alike.

Chapters 6–21 cover the array of analysis techniques offered by JMP. Chapters begin with simple-to-use techniques and gradually work toward more complex methods. Emphasis is on learning to think about these techniques and on how to visualize data analysis at work. JMP offers a graph for almost every statistic and supporting tables for every graph. Using highly interactive methods, you can learn more quickly and discover what your data has to say.

- Chapter 6, "Simulations," introduces you to some probability topics by using the JMP scripting language. You learn how to open and execute these scripts.
- Chapter 7, "Univariate Distributions: One Variable, One Sample," covers distributions of continuous and categorical variables and statistics to test univariate distributions.
- Chapter 8, "The Difference between Two Means," covers *t*-tests of independent groups and tells how to handle paired data. The nonparametric approach to testing related pairs is shown.
- Chapter 9, "Comparing Many Means: One-Way Analysis of Variance," covers one-way analysis of variance, with standard statistics and a variety of graphical techniques.
- Chapter 10, "Fitting Curves through Points: Regression," shows how to fit a regression model for a single factor.

- Chapter 11, “Categorical Distributions,” discusses how to think about the variability in single batches of categorical data. It covers estimating and testing probabilities in categorical distributions, shows Monte Carlo methods, and introduces the Pearson and Likelihood ratio chi-square statistics.
- Chapter 12, “Categorical Models,” covers fitting categorical responses to a model, starting with the usual tests of independence in a two-way table, and continuing with graphical techniques and logistic regression.
- Chapter 13, “Multiple Regression,” describes the parts of a linear model with continuous factors, talks about fitting models with multiple numeric effects, and shows a variety of examples, including the use of stepwise regression to find active effects.
- Chapter 14, “Fitting Linear Models,” is an advanced chapter that continues the discussion of Chapter 12, moving on to categorical effects and complex effects, such as interactions and nesting.
- Chapter 15, “Bivariate and Multivariate Relationships,” looks at ways to examine two or more response variables using correlations, scatterplot matrices, three-dimensional plots, principal components, and other techniques. Outliers are discussed.
- Chapter 16, “Design of Experiments,” looks at the built-in commands in JMP used to generate specified experimental designs. Also, examples of how to analyze common screening and response level designs are covered.
- Chapter 17, “Exploratory Modeling,” illustrates two common data mining techniques—Neural Nets and Recursive Partitioning.
- Chapter 18, “Discriminant and Cluster Analysis,” discusses methods that group data into clumps.
- Chapter 19, “Statistical Quality Control,” discusses common types of control charts for both continuous and attribute data.
- Chapter 20, “Time Series,” discusses some elementary methods for looking at data with correlations over time.
- Chapter 21, “Machines of Fit,” is an essay about statistical fitting that may prove enlightening to those who have a mind for mechanics.

Typographical Conventions

The following conventions help you relate written material to information you see on your screen:

- Reference to menu names (**File** menu) or menu items (**Save** command), and buttons on dialogs (**OK**), appear in the **Helvetica bold** font.
- When you are asked to choose a command from a submenu, such as **File** > **Save As**, go to the **File** menu and choose the **Save As** command.
- Likewise, items on popup menus in reports are shown in the **Helvetica bold** font, but you are given a more detailed instruction about where to find the command or option. For example, you might be asked to select the **Show Points** option from the popup menu on the analysis title bar, or select the **Save Predicted** command from the Fitting popup menu on the scatterplot title bar. The popup menus will always be visible as a small red triangle on the platform or on its outline title bars, as circled in the picture below.

- References to variable names, data table names, and some items in reports show in Helvetica but can appear in illustrations in either a plain or boldface font. These items show on your screen as you have specified in your JMP Preferences.
- Words or phrases that are important, new, or have definitions specific to JMP are in *italics* the first time you see them.
- When there is an action statement, you can follow along with the example by following the instruction. These statements are preceded with a mouse symbol (🖱) in the margin. An example of an action statement is:

🖱 Highlight the Month column by clicking the area above the column name, and then choose **Cols > Column Info**.

- Occasionally, side comments or special paragraphs are included and shaded in gray, or are in a side bar.

2 JMP Right In

Hello!

JMP (pronounced “jump”) software is so easy to use that after reading this chapter you’ll find yourself confident enough to learn everything on your own. Therefore, we cover the essentials fast—before you escape this book. This chapter offers you the opportunity to make a small investment in time for a large return later on.

If you are already familiar with JMP and want to dive right into statistics, you can skip ahead to Chapters 6–21. You can always return later for more details about using JMP or for more details about statistics.

First Session

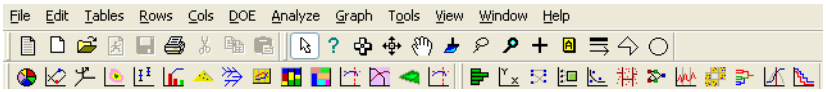
This first section just gets you started learning JMP. In most of the chapters of this book, you can follow along in a hands-on fashion. Watch for the mouse symbol (🖱) and perform the action it describes. Try it now:

- 🖱 To start JMP, double-click the JMP application icon.

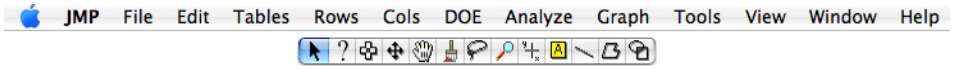
When the application is active, you see the JMP menu bar and the JMP Starter window. You may also see toolbars, depending on how your system is set up. (Macintosh toolbars are attached to each window, and are appropriate for their window, and therefore vary.)

Figure 2.1 The JMP Main Menu and the JMP Starter

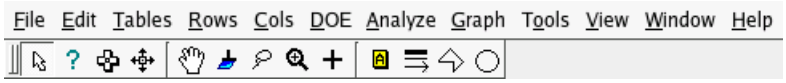
Windows menu and toolbar



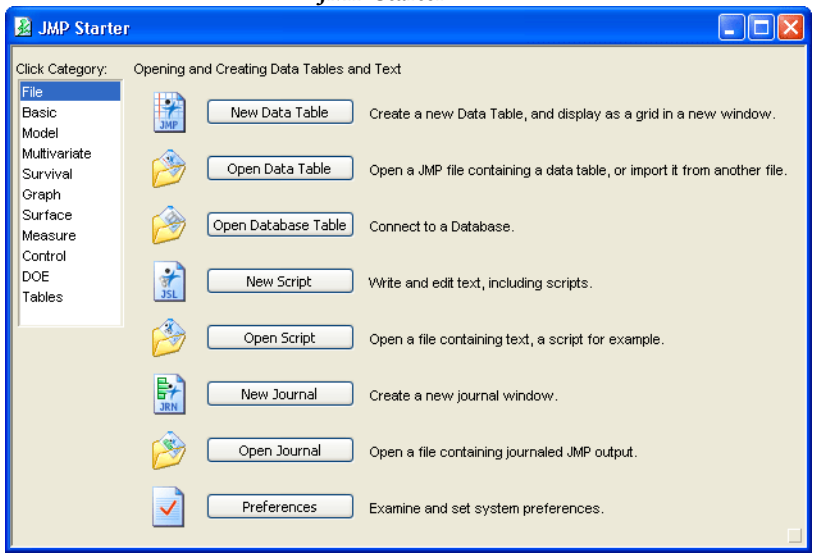
Macintosh menu and toolbar



Linux menu and toolbar



JMP Starter



As with other applications, the **File** menu (**JMP** menu on Macintosh) has all the strategic commands, like opening data tables or saving them. To quit JMP, choose the **Exit** (Windows and Linux) or **Quit** (Macintosh) command from this menu. (Note that the **Quit** command is located on the **JMP** menu on the Macintosh.)

Start by opening a JMP data table and doing a simple analysis.

Open a JMP Data Table

When you first start JMP, you are presented with the JMP Starter window, a window that allows quick access to the most frequently used features of JMP. Instead of starting with a blank file or importing data from text files, open a JMP data table from the collection of sample data tables that comes with JMP.

- Choose the **Open** command in the **File** menu (choose **File** > **Open**).
- When the Open File dialog appears, as shown in **Figure 2.2**, **Figure 2.3**, or **Figure 2.4**, select Big Class.jmp from the list of sample data files.

- Windows sample data is usually installed at C:\Program Files\SAS\JMP7\English Support Files\Sample Data.
- Macintosh Sample Data is usually installed at the root level at /Library/Application Support/JMP/Support Files English/Sample Data.
- Linux Sample Data is usually installed at /JMP7/Support Files English/Sample Data in the directory where you installed JMP (typically /opt).

- Select Big Class and click **Open** (Windows and Macintosh) or **Finish** (Linux) on the dialog.

There is also a categorized list of the sample data, accessible from **Help > Sample Data Directory**. The pre-defined list of files may help you when searching through the samples. The above procedure was meant to show you how to, in general, open a data table.

Figure 2.2 Open File Dialog (Windows)

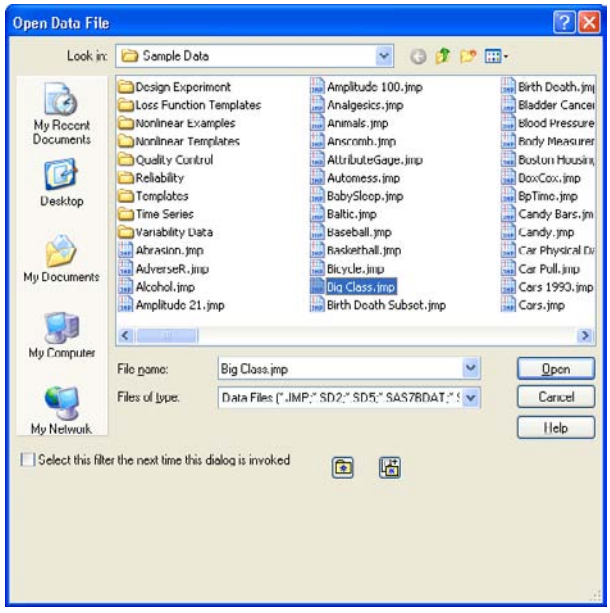


Figure 2.3 Open File Dialog (Macintosh)

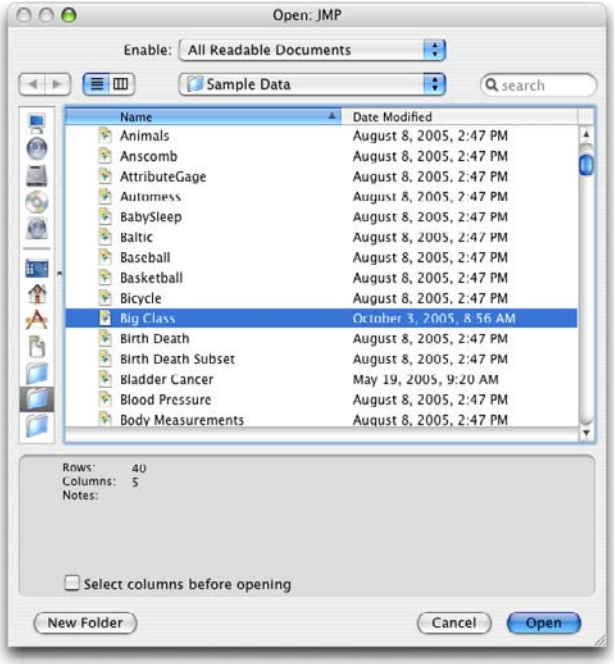
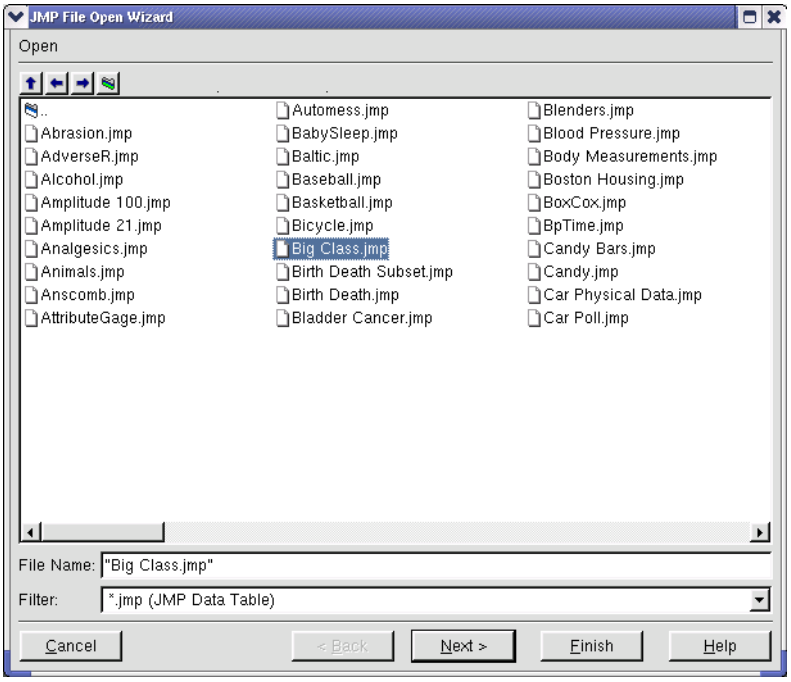


Figure 2.4 Open File Dialog (Linux)



You should now see a table with columns titled name, age, sex, height, and weight (shown in **Figure 2.5**).

In Chapter 3, "Data Tables, Reports, and Scripts" on page 27, you learn the details of the data table, but for now let's try an analysis.

Figure 2.5 Partial Listing of the Big Class Data Table

Big Class
Distribution
Bivariate
Oneway
Logistic
Contingency
Fit Model
Set Sex Value Labels
Set Age Value Labels
Columns (5/0)
name
age
sex
height
weight
Rows
All rows 40

	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M	60	84
7	JAMES	12	M	61	128
8	ROBERT	12	M	51	79
9	BARBARA	13	F	60	112
10	ALICE	13	F	61	107
11	SUSAN	13	F	56	67
12	JOHN	13	M	65	98
13	JOE	13	M	63	105
14	MICHAEL	13	M	58	95
15	DAVID	13	M	59	79

Launch an Analysis Platform

What is the distribution of the weight and age columns in the table?

- Click on the **Analyze** menu and choose the **Distribution** command.

This is called *launching* the Distribution platform. The launch dialog (**Figure 2.6**) now appears, prompting you to choose the variables you want to analyze.

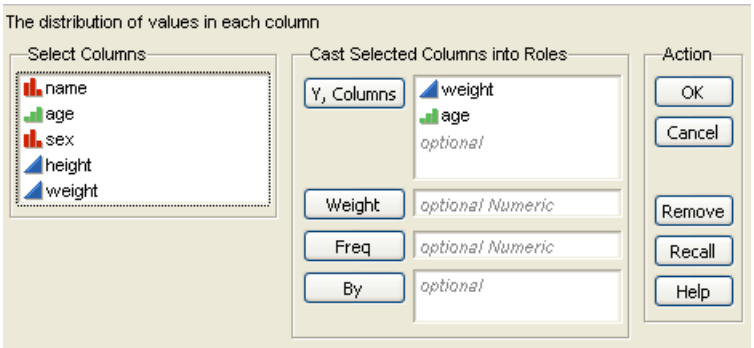
- Click on weight to highlight it in the variable list on the left of the dialog.
- Click **Y, Columns** to add it to the list of variables on the right of the dialog, which are the variables to be analyzed.
- Similarly, select the age variable and add it to the analysis variable list.

The term *variable* is often used to designate a column in the data table. Picking variables to fill roles is sometimes called *role assignment*.

You should now see the completed launch dialog shown in **Figure 2.6**.

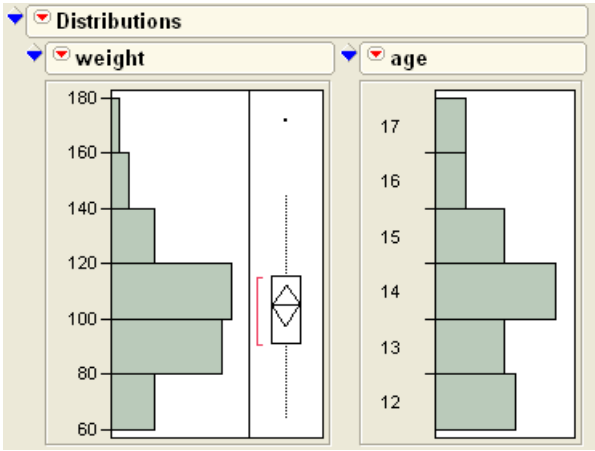
- Click **OK**, which closes the launch dialog and performs the Distribution analysis.

Figure 2.6 Distribution Launch Dialog



The resulting window shows the distribution of the two variables, weight and age, as in **Figure 2.7**.

Figure 2.7 Histograms from the Distribution Platform



Interact with the Surface of the Report

All JMP reports start with a basic analysis, which is then worked with interactively. This allows you to dig into a more detailed analysis, or customize the presentation. The report is a live object, not a dead transcript of calculations.

Row Highlighting

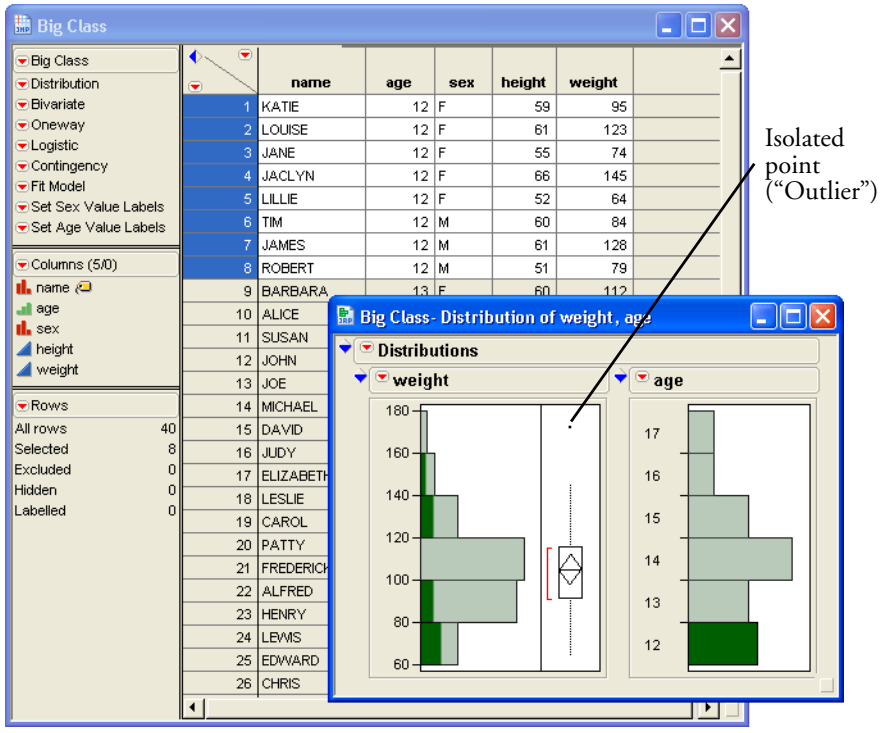
- Click on one of the histogram bars, for example, the age bar for 12-year-olds.

The bar is highlighted, along with portions of the bars in the other histogram and certain rows in the data table corresponding to the highlighted histogram bar. This is the dynamic

linking of rows in the data tables to plots. Later, you will see other ways of selecting and working with attributes of rows in a table.

On the Windows and Linux operating systems, if you have all the windows maximized, then you need to un-maximize them to see both windows at the same time.

Figure 2.8 Highlighted Bars and Data Table Rows



On the right of the weight histogram is a box plot with a single point near the top.

- Move the mouse over that point to see the label, LAWRENCE, appear in a popup box.
- Click on the point in the plot.

The point highlights and the corresponding row is highlighted in the data table.

Disclosure Icons

Each report title is part of the analysis presentation outline. Click on the diamond on the side of each report title to alternately open and close the contents of that outline level.

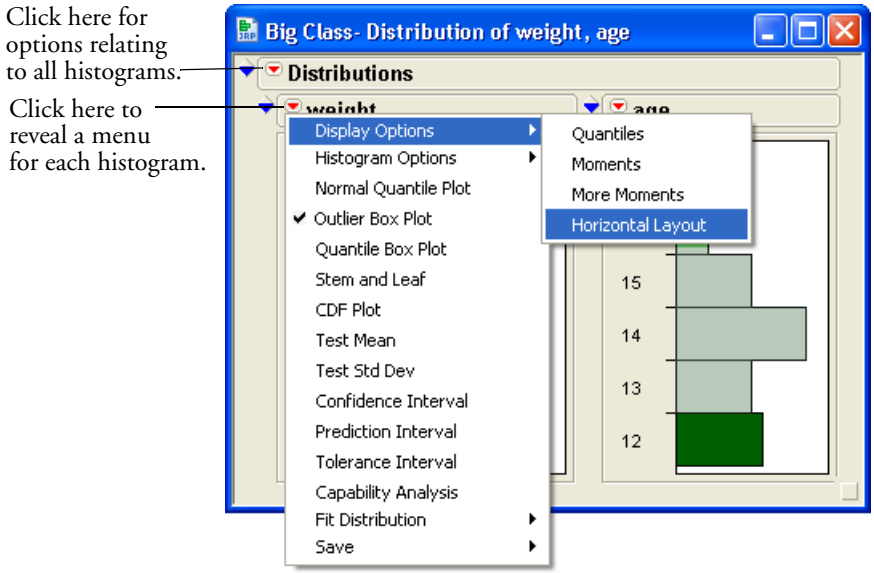
Figure 2.9 Disclosure Icons for Windows and Linux (left) and Macintosh (right)

Contextual Popup Menus

There is a small red triangle (a *hot spot*) on the title bar at the top of the analysis window that accesses popup menu commands for the analysis. This popup menu has commands specific to the platform. Hot spots on the title bars of each histogram contain commands that only influence that histogram. For example, you can change the orientation of the graphs in the Distribution platform by checking or unchecking **Display Options** > **Horizontal Layout** (**Figure 2.10**).

Click on one of the menus next to weight or age and select **Display Options** > **Horizontal Layout**.

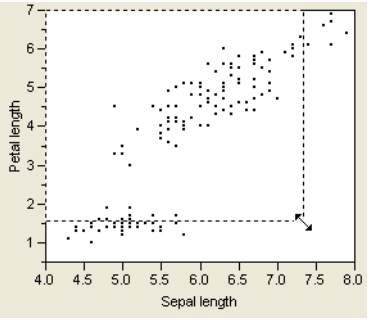
Figure 2.10 Display Options Menu



In this same popup menu, you find options for performing further analyses or saving parts of the analysis in several forms. Whenever you see a red triangle hot spot, there are more options available. The options are specific to the context of the outline level where they are located. Many options are explained in later sections of this book.

Resizing Graphs

If you want to resize the graph windows in an analysis, move your mouse over the side or corner of the graph. The cursor changes to a double arrow, which lets you to drag the borders of the graph to the position you want.



Special Tools

When you need to do something special, pick a tool in the tools menu or tool palette and click or drag inside the analysis.

The grabber ( ) is for grabbing objects.

- Select the grabber, then click and drag in a continuous histogram.

The brush ( ) is for highlighting all the data in an rectangular area.

- Try getting the brush and dragging in the histogram. To change the size of the rectangle, option-drag (Macintosh), Alt-drag (Windows) or Shift-Alt-drag on Linux.

The lasso ( ) is for selecting points by roping them in. We use this later in scatterplots.

The crosshairs ( ) are for sighting along lines in a graph.

The magnifier ( ) is for zooming in to certain areas in a plot. Hold down the ⌘ (Macintosh) Alt (Windows) or Shift+Alt (Linux) key and click to restore the original scaling.

The drawing tools ( ) let you draw circles, squares, lines and shapes to annotate your report. The annotate tool ( ) is for adding text annotations anywhere on the report.

The question mark ( ) is for getting help on the analysis platform surface.

- Get the question mark tool and click on different areas in the Distribution platform.

The selection tool ( )is for picking out an area to copy so that you can paste its contents into another application. Hold down the Shift key to select multiple report sections. Refer to the chapter “Data Tables, Reports, and Scripts” on page 27 for details.

In JMP, the surface of an analysis platform bristles with interactivity. Launching an analysis is just the starting point. You then explore, evaluate, follow clues, dig deeper, get more details, and fine-tune the presentation.

Modeling Type

Notice in the previous example that there are different kinds of graphs and reports for weight and age. This is because the variables are assigned different *modeling types*. The weight column has a *continuous* modeling type, so JMP treats the numbers as values from a continuous scale. The age column has an *ordinal* modeling type, so JMP treats its values as labels of discrete categories.

Here is a brief description of the three modeling types:

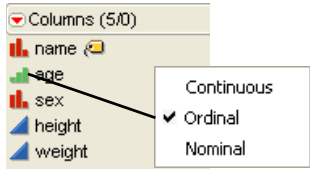
- *Continuous* ( )are numeric values used directly in an analysis.
- *Ordinal* ( )values are category labels, but their order is meaningful.
- *Nominal* ( )values are treated as unordered, categorical names of levels.

The ordinal and nominal modeling types are treated the same in most analyses, and are often referred to collectively as *categorical*.

You can change the modeling type using the Columns panel at the left of the data grid (**Figure 2.11**). Notice the beside the column heading for age. This icon is a popup menu.

Click on the to see the menu for choosing the modeling type for a column.

Figure 2.11 Modeling Type Popup Menu on the Columns Panel



Why does JMP distinguish among modeling types? For one thing, it's a convenience feature. You are telling JMP ahead of time how you want the column treated so that you don't have to say it again every time you do an analysis. It also helps reduce the number of commands you need to learn. Instead of two distribution platforms, one for continuous variables and a different one for categorical variables, a single command performs the anticipated analysis based on the modeling type you assigned.

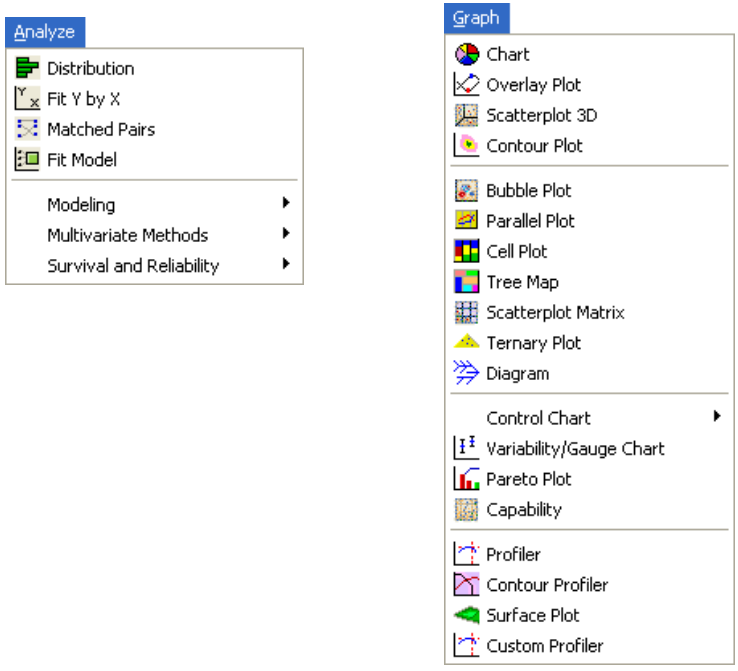
You can change the modeling type whenever you want the variable treated differently. For example, if you wanted to find the mean of age instead of categorical frequency counts, simply change the modeling type from ordinal to continuous and repeat the analysis.

The following sections demonstrate how the modeling type affects the kind of analysis from several platforms.

Analyze and Graph

The **Analyze** and **Graph** menus, shown here, launch interactive platforms to analyze data.

Figure 2.12 Analyze and Graph Menus



The **Analyze** menu is for statistics and data analysis. The **Graph** menu is for specialized plots. That distinction, however, doesn't prevent analysis platforms from being full of graphs, nor the graph platforms from computing statistics. Each platform provides a context for sets of related statistical methods and graphs. It won't take long to learn this short list of platforms. The next sections briefly describe the **Analyze** and **Graph** commands.

The Analyze Menu

Distribution is for univariate statistics, and describes the distribution of values for each variable, one at a time, using histograms, box plots, and other statistics.

Fit Y by X is for bivariate analysis. A bivariate analysis describes the distribution of a *y*-variable as it depends on the value of the *x*-variable. The continuous or categorical modeling type of

the *y*- and *x*- variables leads to one of the four following analyses: scatterplot with regression curve fitting, one-way analysis of variance, contingency table analysis, or logistic regression.

Matched Pairs compares means between two response columns using a paired *t*-test. Often the two columns represent measurements on the same subject before and after some treatment.

Fit Model launches a general fitting platform for linear models. Analyses found in this platform include multiple regression, analysis of variance models, generalized linear models, and logistic regression.

Modeling

Screening helps select a model to fit to a two-level screening design by showing which effects are large.

Nonlinear fits models that are nonlinear in their parameters, using iterative methods.

Neural Net implements a standard type of neural network.

Gaussian Process models the relationship between a continuous response and one or more continuous predictors. These models are common in areas like computer simulation experiments, such as the output of finite element codes, and they often perfectly interpolate the data. Gaussian processes can deal with these no-error-term models.

Time Series lets you explore, analyze, and forecast univariate time series taken over equally spaced time periods. The analysis begins with a plot of the points in the time series with autocorrelations and partial autocorrelations, and can fit ARIMA, seasonal ARIMA, transfer function models, and smoothing models.

Partition recursively partitions values, similar to CART™ and CHAID™.

Categorical tabulates and summarizes categorical response data, including multiple response data, and calculates test statistics. It is designed to handle survey and other categorical response data, including multiple response data like defect records, side effects, and so on.

Multivariate Methods

Multivariate describes relationships among variables, focusing on the correlation structure: correlations and other measures of association, scatterplot matrices, multivariate outliers, and principal components.

Cluster allows for *k*-means and hierarchical clustering. Normal mixtures and Self-Organizing Maps (SOMs) are found in this platform.

Principal Components derives a small number of independent linear combinations (principal components) of a set of variables that capture as much of the variability in the original variables as possible. JMP offers several types of orthogonal and oblique Factor-Analysis-Style rotations to help interpret the extracted components.

Discriminant fits discriminant analysis models, categorizing data into groups.

PLS implements partial least-squares analyses.

Item Analysis analyzes questionnaire or test data using Item Response Theory.

Survival and Reliability

Survival /Reliability models the time until an event, allowing censored data. This kind of analysis is used in both reliability engineering and survival analysis.

Fit Parametric Survival opens the Fit Model dialog to model parametric (regression) survival curves.

Fit Proportional Hazards opens the Fit Model dialog to fit the Cox proportional hazards model.

Recurrence Analysis analyzes repairable systems.

The Graph Menu

Chart gives many forms of charts such as bar, pie, line, and needle charts.

Overlay Plot overlays several numeric *y*-variables, with options to connect points, or show a step plot, needle plot, or others. It is possible to have two *y*-axes in these plots.

Scatterplot 3D produces a three-dimensional spinnable display of values from any three numeric columns in the active data table. It also produces an approximation to higher dimensions through principal components, standardized principal components, rotated components, and biplots.

Contour Plot constructs a contour plot for one or more response variables for the values of two *x*-variables. **Contour Plot** assumes the *x* values lie in a rectangular coordinate system, but the observed points do not have to form a grid.

Bubble Plot draws a scatter plot which represents its points as circles (bubbles). Optionally the bubbles can be sized according to another column, colored by yet another column, aggregated across groups defined by one or more other columns, and dynamically indexed by a time column.

Parallel Plot shows connected-line plots of several variables at once.

Cell Plot produces a "heat map" of a column, assigning colors based on a gradient (for continuous variables) or according to a discrete list of colors (for categorical variables).

Tree Map presents a two-dimensional, tiled view of the data.

Scatterplot Matrix produces scatterplot matrices.

Ternary Plot constructs a plot using triangular coordinates. The ternary platform uses the same options as the contour platform for building and filling contours. In addition, it uses a specialized crosshair tool that lets you read the triangular axis values.

Diagram is used to construct *Ishikawa charts*, also called *fishbone charts*, or *cause-and-effect diagrams*. These charts are useful when organizing the sources (causes) of a problem (effect), perhaps for brainstorming, or as a preliminary analysis to identify variables in preparation for further experimentation.

Control Chart presents a submenu of various control charts available in JMP.

Variability/Gage Chart is used for analyzing measurement systems. Data can be continuous measurements or attributes.

Pareto Plot creates a bar chart (Pareto chart) that displays the severity (frequency) of problems in a quality-related process or operation. Pareto plots compare quality-related measures or counts in a process or operation. The defining characteristic of Pareto plots is that the bars are in descending order of values, which visually emphasizes the most important measures or frequencies.

Capability measures the conformance of a process to given specification limits. Using these limits, you can compare a current process to specific tolerances and maintain consistency in production. Graphical tools such as the goalpost plot and box plot give you quick visual ways of observing within-spec behaviors.

Profiler is available for tables with columns whose values are computed from model prediction formulas. Usually, profiler plots appear in standard least squares reports, where they are a menu option. However, if you save the prediction equation from the analysis, you can access the prediction profile independent of a report from the **Graph** menu and look at the model using the response column with the saved prediction formula.

Contour Profiler works the same as the **Profiler** command. It is usually accessed from the Fit Model platform when a model has multiple responses. However, if you save the prediction

formulas for the responses, you can access the Contour Profiler at a later time from the **Graph** menu and specify the columns with the prediction equations as the response columns.

Surface Plot draws up to four three-dimensional, rotatable surfaces. It can also produce density shells and isosurfaces.

Custom Profiler is an advanced feature for optimization and simulation.

Navigating Platforms and Building Context

The first few times JMP is used, most people have navigational questions: How do I get a particular graph? How do I produce a histogram? How do I get a *t*-test?

The strategy for approaching JMP analyses is to build an analysis context. Once you build that context, the graphs and statistics become easily available—often they happen automatically, without having to ask for them specifically.

There are three keys for establishing the context:

- Designating the *Modeling Type* of the variables in the analysis as either continuous, ordinal, or nominal.
- Assigning *X or Y Roles* to identify whether the variable is a response (Y) or a factor (X).
- Selecting an *analysis platform* for the general approach and character of the analysis.

Once you settle on a context, commands appear in logical places.

Contexts for a Histogram

Suppose you want to display a histogram. In other software, you might find a histogram command in a graph menu, but in JMP you need to think of the context. You want a histogram so that you can see a distribution of values. So, launch the **Distribution** platform in the **Analyze** menu. Once launched, there are many graphs and reports available for focusing on the distribution of values.

Occasionally, you may want the histogram as a presentation graph. Then, instead of using the Distribution platform, use the Chart platform in the **Graph** menu.

Contexts for the *t*-Test

Suppose you want a *t*-test. Other software might have a *t*-test command on a main menu. JMP has many *t*-test commands, because there are many contexts in which this statistic is used. So first, you have to build the context of your situation.

If you want the *t*-test to test a single variable's mean against a hypothesized value, you are focusing on a univariate distribution, and therefore launch the Distribution platform (**Analyze > Distribution**). On the title bar of the distribution report is a popup menu with the command **Test Mean**. This command gives you a *t*-test, as well as the option to conduct a nonparametric test.

If you want the *t*-test to compare the means of two independent groups, then you have two variables in the context—perhaps a continuous *Y* response and a categorical *X* factor. Since the analysis deals with two variables, use the Fit Y By X platform. If you launch the Fit Y by X platform, you'll see the side-by-side comparison of the two distributions, and you can use the **t test** or **Means/Anova/Pooled t** command from the popup menu on the analysis title bar.

If you want to compare the means of two continuous responses that form matched pairs, there are several ways to build the appropriate context. You can make a third data column to form the difference of the responses, and use the Distribution platform to do a *t*-test that the mean of the differences is zero. Alternatively, you can use the **Matched Pairs** command to launch the Matched Pairs platform for the two variables. In Chapter 8, "The Difference between Two Means," you will learn more ways to do a *t*-test.

Contexts for a Scatterplot

Suppose you want a scatterplot of two variables. The general context is a bivariate analysis, which suggests using the Fit Y by X platform. With two continuous variables, the Fit Y By X platform produces a scatterplot. You can then fit regression lines or other appropriate items with this scatterplot from the same report.

You might also consider the **Overlay Plot** command in the **Graph** menu when you want a presentation graph. As a **Graph** menu platform, it does not compute regressions, but can overlay multiple Y's in the same graph and support two *y*-axes.

If you have a whole series of scatterplots for many variables in mind, your context is many bivariate associations, which is part of either the Multivariate or Scatterplot 3D platforms. A matrix of scatterplots appears automatically for Multivariate platform analyses.

Contexts for Nonparametric Statistics

There is not a separate platform for nonparametric statistics. However, there are many standard nonparametric statistics in JMP, positioned by context. When you test a mean in the Distribution platform, you have the option to do a (nonparametric) Wilcoxon signed-rank test. When you do a *t*-test or one-way ANOVA in the Fit Y by X platform, you also have three optional nonparametric tests, including the Wilcoxon rank sum (which is equivalent to the

Mann-Whitney *U*-test). If you want a nonparametric measure of association, like Kendall's τ or Spearman's correlation, look in the Multivariate platform.

The Personality of JMP

Here are some reasons why JMP is different from other statistical software:

Graphs are in the service of statistics (and vice versa). The goal of JMP is to provide a graph for every statistic, presented with the statistic. The graphs shouldn't appear in separate windows, but rather should work together. In the analysis platforms, the graphs tend to follow the statistical context. In the graph platforms, the statistics tend to follow the graphical context.

JMP encourages good data analysis. In the example presented in this chapter, you didn't have to ask for a histogram because it appeared when you launched the Distribution platform. The Distribution platform was designed that way, because in good data analysis you always examine a graph of a distribution before you start doing statistical tests on it. This encourages responsible data analysis.

JMP allows you to make discoveries. JMP was developed with the charter to be "Statistical Discovery Software." After all, you want to find out what you didn't know, as well as try to prove what you already know. Graphs attract your attention to an outlier or other unusual feature of the data that might prove valuable to discovery. Imagine Marie Curie using a computer for her pitchblende experiment. If software had given her only the end results, rather than showing her the data and the graphs, she might not have noticed the discrepancy that led to the discovery of radium.

JMP bristles with interactivity. In some products, you have to specify exactly what you want ahead of time because often that is your only chance while doing the analysis. JMP is interactive, so everything is open to change and customization at any point in the analysis. It is easier to remove a histogram when you don't want it than decide ahead of time that you want one.

You can see your data from multiple perspectives. Did you know that a *t*-test for two groups is a special case of an *F*-test for several groups? With JMP, you tend to get general methods that are good for many situations, rather than specialty methods for special cases. You also tend to get several ways to test the same thing. For two groups, there is a *t*-test and its equivalent *F*-test. When you are ready for more, there are three nonparametric tests to use in the same situation. You also can test for different variances across the groups and get appropriate results. And there are two graphs to show you the separation of the means. Even after you perform statistical tests, there are multiple ways of looking at the results, in terms of the *p*-

value, the confidence intervals, least significant differences, the sample size, and least significant number. With this much statistical breadth, it is good that commands appear as you qualify the context, rather than having to select multiple commands from a single menu bar. JMP unfolds the details progressively, as they become relevant.

3 Data Tables, Reports, and Scripts

Overview

JMP data are organized as rows and columns of a grid referred to as the *data table*. The columns have names and the rows are numbered. An open data table is kept in memory and displays a data grid with panels of information about the data table. You can open as many data tables in a JMP session as memory allows.

People often ask about the largest data table that JMP can handle. This question is harder to answer than it sounds, since JMP can store some types of data in a more compact form than others. In general, we recommend that the largest data table that JMP should be used for is one that is the same size as the actual memory resident in the machine. So, for example, this book is being written on a machine with 2GB of memory. JMP could therefore comfortably manage a data table that is (approximately) 2 GB.

Commands in the **File**, **Edit**, **Tables**, **Rows**, and **Cols** menus give you a broad range of data handling operations and file management tasks, such as data entry, data validation, text editing, and extensive table manipulation. The Edit menu contains the **Run Script** command, used for executing scripts.

In particular, the **Tables** menu has commands that perform a wide variety of data management tasks on JMP data tables.You can also create summary tables for group processing and summary statistics.

The purpose of this chapter is to tell you about JMP data tables and give a variety of hands-on examples to help you get comfortable handling table operations and scripts.

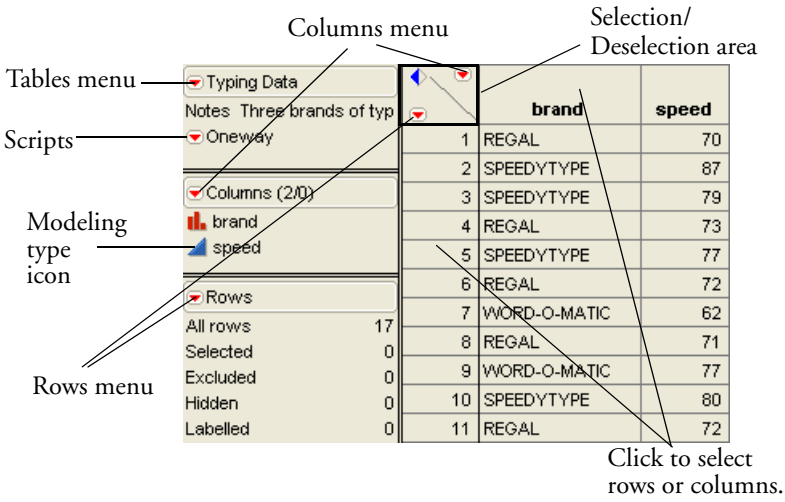
The Ins and Outs of a JMP Data Table

JMP displays data as a data grid, often called a data table. From the data table, you can do a variety of table management tasks such as editing cells; creating, rearranging or deleting rows and columns; subsetting the data; sorting; or combining tables. **Figure 3.1** identifies active areas of a JMP data table.

There are a few basic things to keep in mind:

- Column names can use any keyboard character, including spaces. The size and font for names and values is a setting you control through JMP Preferences.
- If the name of the column is long, you can drag column boundaries to widen the column.
- There is no set limit to the number of rows or columns in a data table; however, the table must fit in memory.

Figure 3.1 Active Areas of a JMP Spreadsheet



Selecting and Deselecting Rows and Columns

Many actions from the **Rows** and **Cols** menus operate only on selected rows and columns. To select rows and columns, highlight them.

- To highlight a row, click the space that contains the row number.

- To highlight a column, click the background area above the column name.

These areas are shown in **Figure 3.1**.

To extend a selection of rows or columns, drag (in the selection area) across the range of rows or columns or Shift-click the first and last row or column of the range. Ctrl-click (⌘-click on the Macintosh) to make a discontiguous selection. To select both rows and columns at the same time, drag across table cells in the spreadsheet.

To deselect a row or column, Ctrl-click (⌘-click on the Macintosh) on the row or column. To deselect all rows or columns at once, click the triangular rows or columns area in the upper-left corner of the spreadsheet.

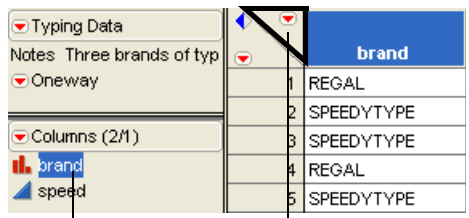
Mousing Around a Spreadsheet: Cursor Forms

To navigate in the spreadsheet, you need to understand how the cursor works in each part of the spreadsheet.

To experiment with the different cursor forms, open the Typing.jmp sample table, move the mouse around on the surface (as illustrated in **Figure 3.1**), and see how the cursor changes to the forms listed next.

Arrow cursor ( )

When a data table is the active window, the cursor is a standard arrow when it is anywhere in the table panels to the left of the data grid, except when it is on a red triangle popup menu icon or a diamond-shaped disclosure icon. It is also a standard arrow when it is in the upper-left corner of the data grid. It is used for the selection of items.

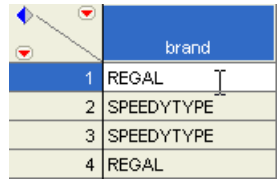


Click to deselect columns.

Click to select columns.
Select and double-click to edit column name.

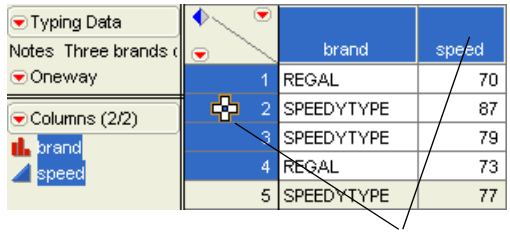
I-beam cursor ( )

The cursor is an I-beam when it is over text in the data grid, highlighted column names in the data grid, or column panels. It signals that the text underneath it is editable. To edit text in the data grid, position the I-beam next to characters and click to highlight the cell. Once the cell is highlighted, simply begin typing to edit the cell's contents. Alternatively, double-click and start typing to replace the existing text.



Selection cursor ( )

Move the blinking cursor around. The cursor becomes a large thick plus sign when you move it into a column or row selection area. It is used to select items in JMP data tables and reports. Use the selection cursor to select a single row or column. Shift-click a beginning and ending row (or a beginning and ending column) to select an entire range. Ctrl-click (⌘-click on the Macintosh) to select multiple rows or columns that are not contiguous.



Click and drag to select.

The selection cursor appears when you select it from the tools menu. It is used to select areas of reports to copy and paste to other locations. See “Copy, Paste, and Drag Data” on page 44 for details on using it for cut and paste operations.

Double Arrow cursor ( )

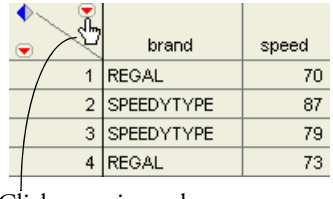
The cursor changes to a double arrow when placed on a column boundary. To change the width of a spreadsheet column, drag this cursor left or right.

List Check and Range Check cursors ( )

The cursor changes when it moves over values in columns that have data validation in effect (automatic checking for specific values). It becomes a small, downward-pointing arrow on a column with *list checking* and a large double I–beam on a column with *range checking*. When you click, the value highlights and the cursor becomes the standard I–beam; you enter or edit data as usual. However, you can only enter data values from a list or range of values you pre-specify. In addition, you can right-click in columns with list checks to see a menu listing the possible entries for the column.

Popup Pointer cursor ( )

The cursor changes to a finger pointer over any menu icon or diamond-shaped disclosure icon. It signifies that you are over a clickable item. Click a disclosure icon to open or close a window panel or report outline; click red triangle buttons to open menus.



Click to activate the menu.

- After you finish exploring, choose **File** > **Close** to close the Typing.jmp table.

Creating a New JMP Table

Hopefully, most of the data you analyze is already in electronic form. However, if you have to key in data, a JMP data table is like a spreadsheet with familiar data entry features. A short example shows you how to start from scratch.

Suppose data values are blood pressure readings collected over six months and recorded in a notebook page as shown in **Figure 3.2.**

Figure 3.2 Notebook of Raw Study Data Used to Define Rows and Columns

	Blood Pressure Study				
	Month	*Control*	*Placebo*	*300mg*	*450mg*
	March	*165*	*163*	*166*	*168*
	April	*162*	*159*	*165*	*163*
	May	*164*	*158*	*161*	*153*
	June	*162*	*161*	*158*	*151*
	July	*166*	*158*	*160*	*148*
	August	*163*	*158*	*157*	*150*

Define Rows and Columns

JMP data tables have rows and columns, which represent *observations* and *variables* in statistical terms. This differs from the way other software that also presents data as a data grid works. In JMP, the rows always represent observations, and the columns always represent variables. The raw data in **Figure 3.2** are arranged as five columns (month and four treatment groups) and six rows (months March through August). The first line in the notebook describes each column of values, and these descriptions can be used as column names in a JMP data table. To enter this data into JMP, you first need a blank data table.

- Choose **File** > **New** > **(New)**[1] **Data Table** to create a new empty data table. This new, untitled table has one column and no rows.

1. On the Macintosh, the menu item is **New Data Table**, while on other platforms, it is simply **Data Table**. We use this notation to represent all platforms.

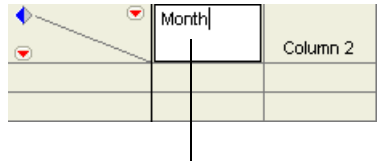
Add Columns

We now want to add five columns to the data table to hold the data from the study.

- Choose **Cols** > **Add Multiple Columns** and respond to the **How many columns to add** question by requesting five new columns.

The default column names are Column 1, Column 2, and so on, but you can change them by typing in the editable column fields.

To edit a column name, first click the column selection area to highlight the column. Then, begin typing the name of the column. (The column name area highlights, but doesn't obtain "edit focus" until you begin typing). The column name starts changing, with the insertion point showing as a blinking vertical bar. Use the Tab key to move from one column to the next.



Highlight the column, then begin typing.

- Type the names from the data journal (Month, Control, Placebo, 300 mg, and 450 mg) into the columns headers of the new table.

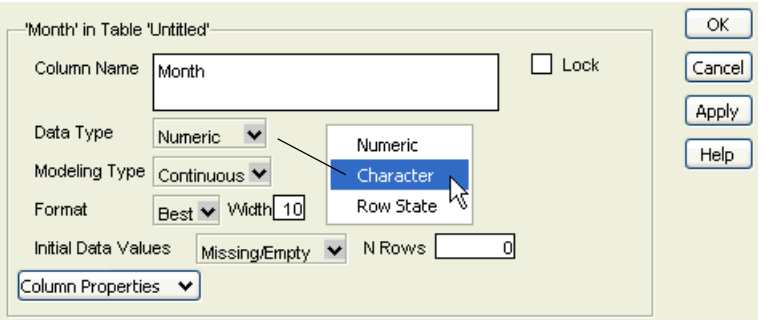
Set Column Characteristics

Columns can have different characteristics, such as modeling type or numeric format. By default, their modeling type is continuous, so they expect numeric data. However, in this example, the Month column holds a non-numeric character variable.

- Right-click the column name area for Month and select **Column Info** to activate the column info dialog.

- In the Column Info dialog, use the Data Type popup menu to change Month to a character variable (**Figure 3.3**), then click **OK**.

Figure 3.3 Column Info Dialog



Add Rows

Adding new rows is easy.

- Choose **Rows** > **Add Rows** and ask for six new rows.

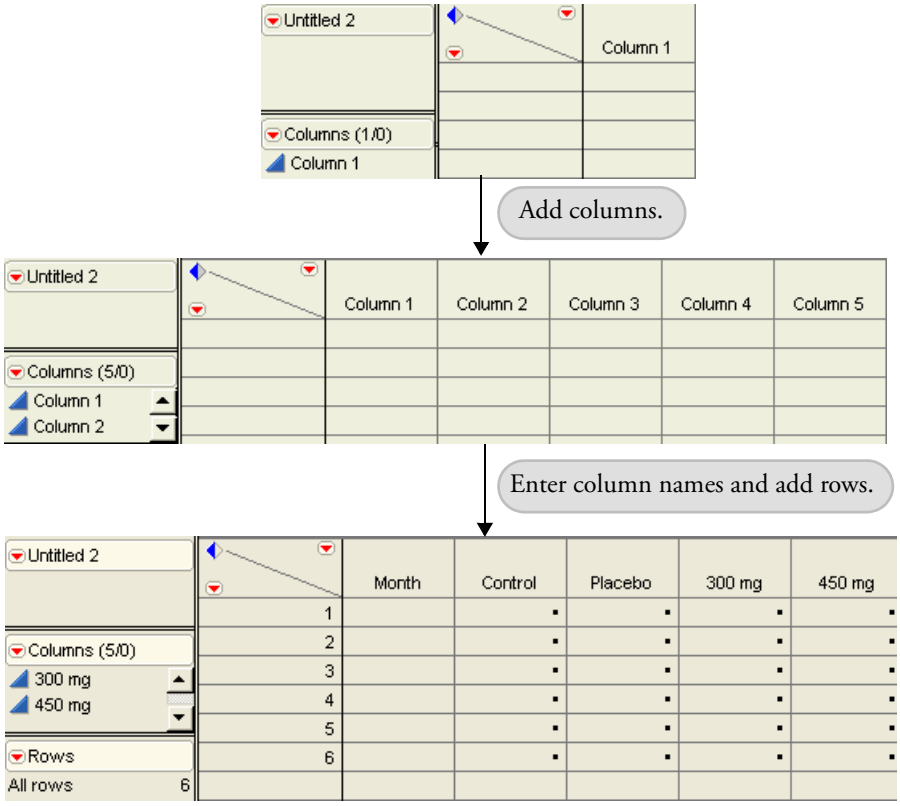
Alternatively, if you double-click anywhere in the body of the data table, the data table automatically fills with new rows through the position of the cursor.

The last step is to give the data table a name and save it.

- Choose **File** > **Save As** to name the data table BP Study.jmp. You may also navigate to another folder if you want to save this data table somewhere else.

The data table is now ready to hold data values. **Figure 3.4** summarizes the table evolution so far.

Figure 3.4 JMP Data Table with New Rows, Columns, and Names



Enter Data

Entering data into the data table requires typing values into the appropriate table cells. To enter data into the data table, do the following:

- Move the cursor into a data cell and double-click to begin editing the cell.

A blinking vertical bar appears in the indicating that you can begin typing.

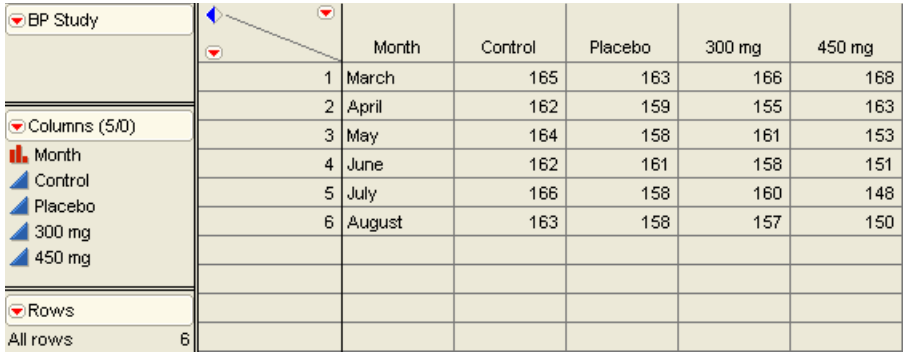
- Type the appropriate data from the notebook (**Figure 3.2**).

If you make a mistake, drag the I–beam across the incorrect entry to highlight it and type the correction over it. The Tab and Return keys are useful keyboard tools for data entry:

- Tab moves the cursor one cell to the right. Shift-Tab moves the cursor one cell to the left. Moving the cursor with the Tab key automatically wraps it to the beginning of the next (or previous) row. Tabbing past the last table cell creates a new row.
- Enter (or Return) either moves the cursor down one cell or one cell to the right, based on the setting in JMP Preferences.

Your results should look like the table in **Figure 3.5**.

Figure 3.5 Finished Blood Pressure Study Table



	Month	Control	Placebo	300 mg	450 mg
1	March	165	163	166	168
2	April	162	159	155	163
3	May	164	158	161	153
4	June	162	161	158	151
5	July	166	158	160	148
6	August	163	158	157	150

The New Column Command

In the first part of this example, you used the **Add Multiple Columns** command from the **Cols** menu to create several new columns in a data table. Often you only need to add a single new column with specific characteristics.

Continuing with the current example, suppose you learn that the blood pressure readings were taken at one lab, called “Accurate Readings Inc.” during March and April, but at another location called “Most Reliable Measurements Ltd.” for the remaining months of the study. You want to include this information in the data table.

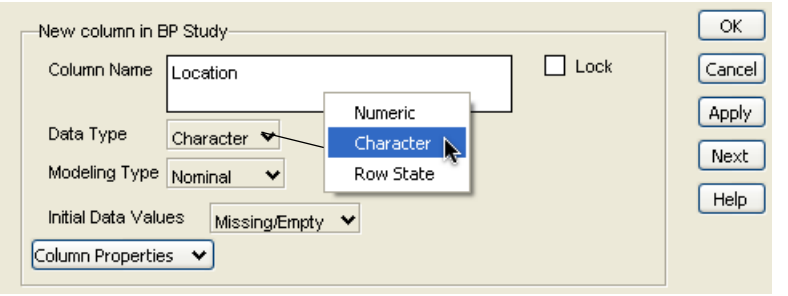
- Begin by choosing **Cols > New Column**, which displays a New Column dialog like the one shown in **Figure 3.6**.

The New Column dialog lets you set the new column’s characteristics.

- Type a new name, Location, in the **Column Name** area.
- Because the actual names of the location are characters, select **Character** from the Data Type menu as shown in **Figure 3.6**.

Notice that the **Modeling Type** then automatically changes to **Nominal**.

Figure 3.6 The New Column Dialog



When you click **OK**, the new column appears in the table, where you can enter the data as previously described.

Plot the Data

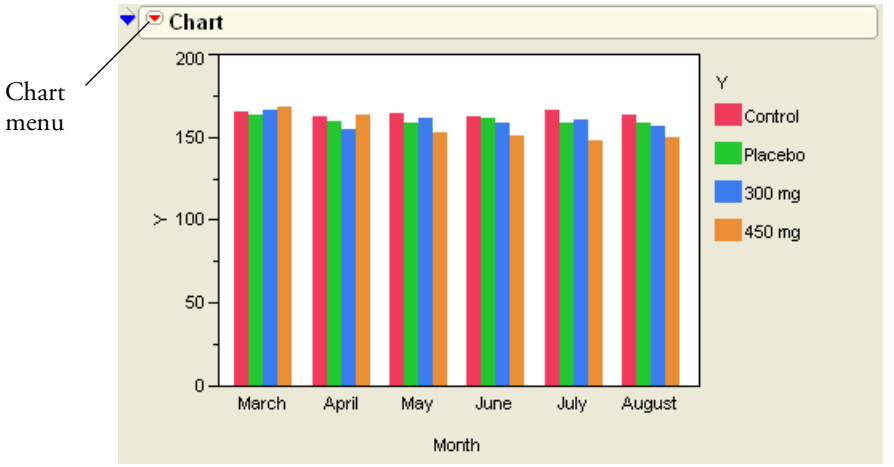
There are many ways to check the data for errors. One way is to plot the data to check for obvious anomalous values. Let's experiment with the **Chart** command in the **Graph** menu.

To plot the months on the horizontal (*x*) axis and the columns of blood pressure statistics for each treatment group on the vertical (*y*) axis, follow these steps:

- Choose **Graph** > **Chart**.
- Designate Month as the **Categories, X, Level** variable.
- Highlight Control, Placebo, 300 mg, and 450 mg, and select **Data** from the **Statistics** drop-down list.

When you click **OK**, you see a bar chart.

Figure 3.7 Initial Bar Chart

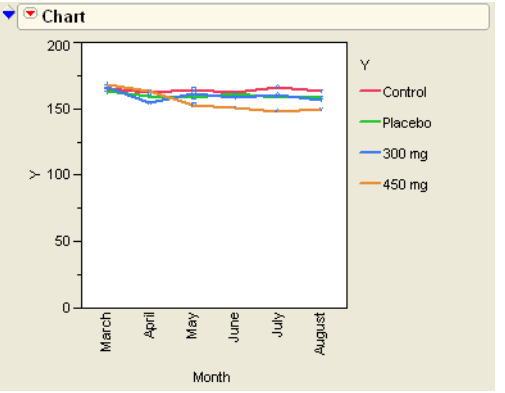


Now, use some options.

- Click the menu icon on the title bar of the chart to see a list of options.
- Make sure the **Overlay** option is checked, and then select **Y Options** > **Line Chart** to see the chart shown in **Figure 3.8**.

The plot doesn't appear to have much to say yet because it is difficult to read at the current scaling.

Figure 3.8 Line Chart for Blood Pressure Values over Month



By default, *y*-axis scaling begins at zero. To present easier-to-read information, the *y*-axis needs to be rescaled.

- Double-click anywhere in the *y*-axis area to bring up the Y Axis Specification dialog box.

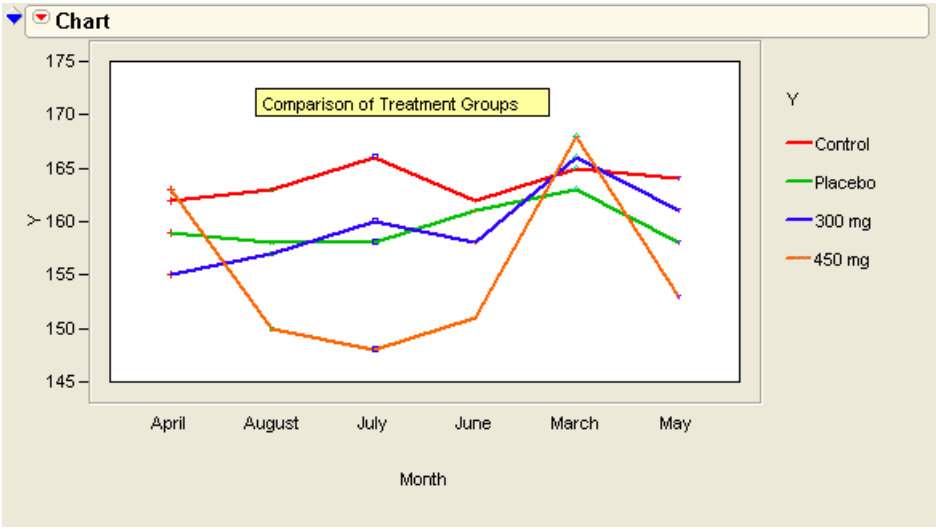
Based on what you can see in **Figure 3.8**, the plotted values range from about 145 to 175.

- Type these values into the Axis Rescale dialog as the minimum and maximum.
- Change the increment for the tick marks from 50 to 5, which divides the range into six intervals (145, 150,..., 160).
- Click **OK**.

Use the Annotate tool ( ) to annotate the chart with captions as shown in **Figure 3.9**.

- Select the Annotate tool and click in the chart where you want to insert the caption.
- Type “Comparison of Treatment Groups” and click outside the caption.
- Resize the caption by clicking and dragging on its corner. Move the caption by clicking and dragging in its interior.

Figure 3.9 Line Chart with Modified *y*-Axis



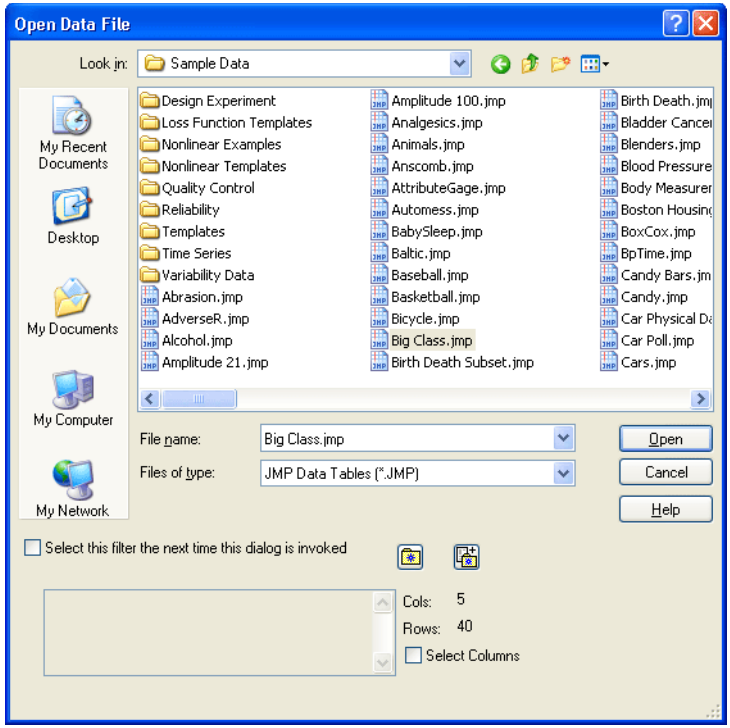
Importing Data

The **File** > **Open** command displays a specialized dialog that lets you locate the file you want to open and tell JMP the format of the incoming file. The **Open** command then reads the file into a JMP data table.

JMP directly reads JMP data tables, JMP journals, JMP scripts, SAS transport files, text files with any column delimiter, Excel files, and flat-file database files. To open database files, you must have an appropriate Open Database Connectivity (ODBC) driver installed on your system. In addition, the Windows version of JMP can read and write SAS data sets.

If you indicate what kind of file to expect with an appropriate **Files of Type** (Windows), **Show** (Macintosh), or **Filter** (Linux) selection, JMP gives helpful information when possible. The example in **Figure 3.10** shows an Open Data File dialog when the **Files of Type** drop-down list is changed from the default (**Data Files**) to **JMP data tables**. The dialog shows the table notes (if there are any).

Figure 3.10 Using the File Open Dialog to Read a JMP Table



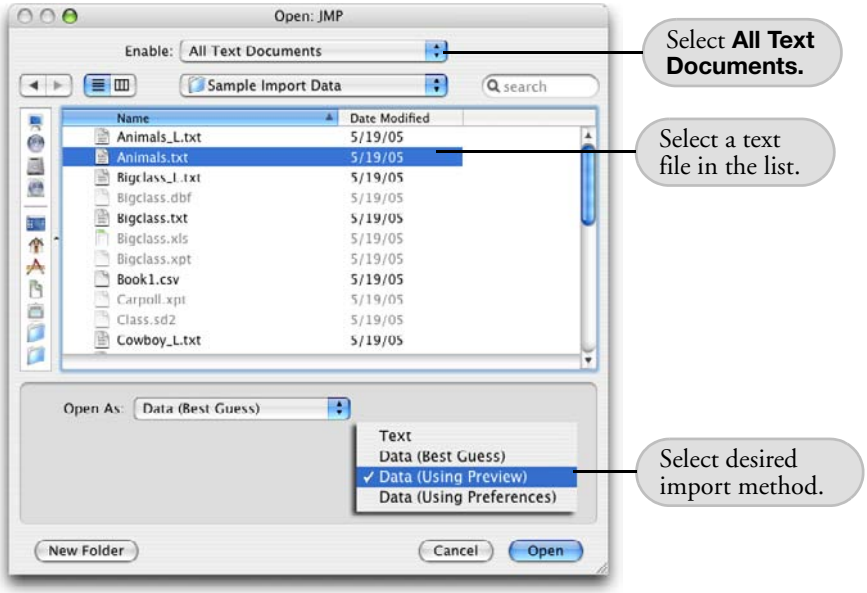
If the incoming file is not a JMP data table and you choose to look at all files, JMP determines the file type by the three-character extension appended to its file name and opens it accordingly. To examine all files, choose * (Linux), *.* (Windows), or **All Readable Files** (Macintosh). This works as long as the file has the structure indicated by its name.

Importing Text Files

To import a text file, select one of the following:

- On Windows, select **Text Import** or **Text Import Preview** in the **Files of type** list.
- On the Macintosh, select **All Text Documents** from the **Enable** list, select a file, then select **Data (Using Preview)** from the **Open As** list (**Figure 3.11**).
- On Linux, select **.txt (Fixed Width)** or **.txt (Delimited)** from the **Filter** menu. Note that you are immediately presented with the dialog in **Figure 3.12**.

Figure 3.11 Macintosh Text Import



If you select **Text Data** (Windows) or **Data (Best Guess)** (Macintosh), JMP attempts to discern the arrangement of text data. This is adequate for rectangular text files with no missing fields, a consistent field delimiter, and an end-of-line delimiter.

Note: If double-quotes are encountered when importing text data, JMP changes the delimiter rules to look for a matching end double-quote. Other text delimiters, including spaces embedded within the quotes, are ignored and treated as part of the text string.

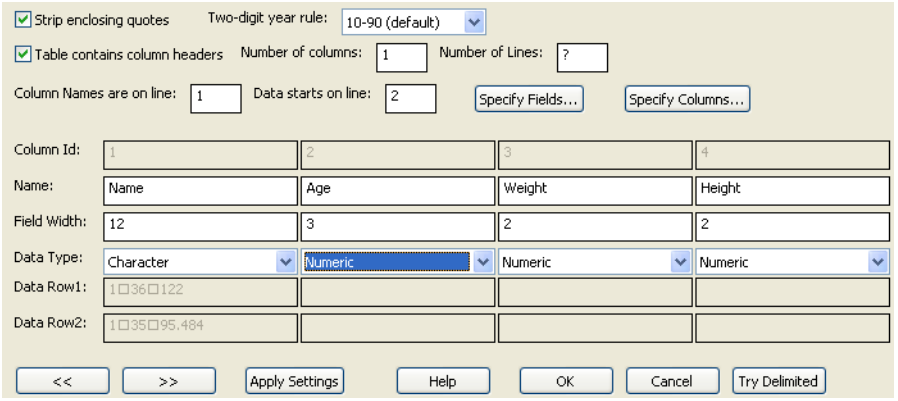
If you want to see a preview of an incoming text file, choose **Text Import Preview** (Windows) or **Data (Using Preview)** (Macintosh). If you want JMP to import the data based on your text import preferences (**File > Preferences** on Windows, **JMP > Preferences** on

Macintosh), click **OK** (on Windows, where this is the default method) or select **Data (Using Preferences)** (Macintosh).

Linux works slightly differently. If you want to open a text file immediately, select the **Finish** button on the Open dialog. To see a preview, select the **Next** button.

Regardless of your operating system, you are then presented with the import options shown in **Figure 3.12**.

Figure 3.12 Import Delimited Field Text File



This dialog is filled in automatically with settings from your Preferences file. It also shows the column names, data types, and the first two rows of data. In **Figure 3.12**, preferences are set that indicate the incoming table contains column headers to be used as the column names; the column names are name, age, weight, and height. If no column names are indicated, the **Name** fields are called Column 1, Column 2, and so on.

You can also identify one or more end-of-field delimiters, end-of-line delimiters, choose the option to **Strip enclosing quotes**, and see how many rows and columns will be read.

If your data has fixed-width fields, press the **Try Fixed Width** button. This alters the dialog so that you can specify the widths of the fields in the input data set.

Importing Microsoft Excel Files

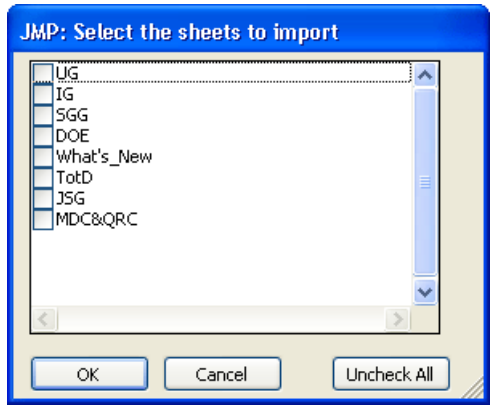
JMP (Windows and Macintosh versions) has the ability to directly import Microsoft Excel worksheets and workbooks. By default, all Excel worksheets are imported as separate JMP tables.

- On Windows, choose **Excel Files(*.xls)** from the **Files of Type** drop-down list.

- On the Macintosh, select **All Readable Documents** from the **Enable** drop-down list.

JMP for Windows also allows you to select among individual worksheets contained in a single Excel workbook.

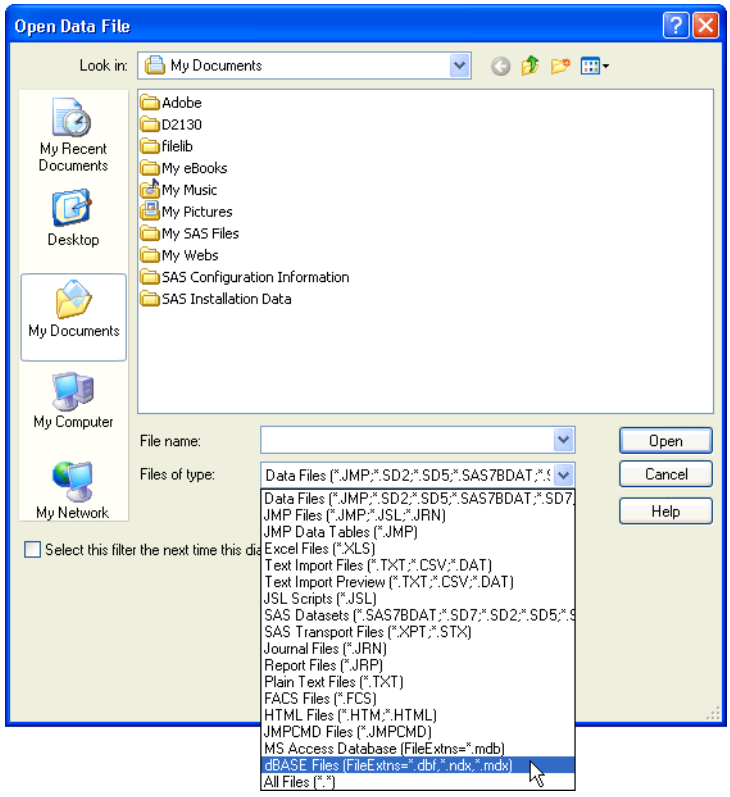
After selecting **Excel Files(*.xls)** from the **Open** dialog, select **Allow individual worksheet selection** and click **OK**. JMP presents the dialog shown in the figure to the right. This dialog shows all the worksheets in the selected workbook. Select the one to open and click **OK** to import the data into JMP.



Using ODBC

JMP can open files for any format that has a corresponding ODBC driver on your system.

On Windows, use the standard **File > Open** command to access flat-file databases like Microsoft Access, Microsoft FoxPro, and dBase. Installed ODBC drivers appear at the end of the **Files of Type** list.



Use the **Database > Open Table** command to import data from relational databases, like dBII and Oracle, and for all databases on Macintosh or Linux operating systems. Details of using the **Database > Open Table** command are in the *JMP User Guide*.

Opening Other File Types

JMP can open other file types aside from those discussed above. The *JMP User Guide* gives details on how to open the following types:

- SAS transport (.xpt, .stx) files
- FACS (.fcs)
- Microsoft Access Database (.mdb) on Windows with a V2+ compliant ODBC driver
- Database (dBASE) (.dbf, .ndx, .mdx) on Windows with a V2+ compliant ODBC driver
- MySQL, Oracle, and PostgreSQL on Linux or Macintosh with a compliant ODBC driver

- OpenOffice spreadsheets (.sxc) on Linux
- Data (.dat) files on Windows
- HTML (.htm, .html) on Windows and Macintosh
- SAS versions 5–9 (.sd2, .sd5, .sd7, .sas7bdat) on Windows
- SAS version 6 (.sas7bdat, .ssd, .ssd01, .saseb$data) on Macintosh and Linux

Copy, Paste, and Drag Data

You can use the standard copy and paste operations to move data and graphical displays within JMP and from JMP to other applications. The following commands in the **Edit** menu let you move data around:

Copy

The **Copy** command in the **Edit** menu copies the values of selected data cells from the active data table to the clipboard. If no rows are selected, **Copy** copies all rows. Likewise, you can copy values from specific columns by selecting them. If no columns are selected, all columns are copied. If you select both rows and columns, **Copy** copies the highlighted cells. Data you cut or copy to the clipboard can be pasted into JMP tables or into other applications.

If you want to copy part of an analysis window, use the selection tool ( ) from the **Tools** menu. Click on the area you want to copy to select and highlight it. Shift-click to extend the selection. If nothing is selected, the **Copy** command copies the entire window to the clipboard.

Paste

The **Paste** command copies data from the clipboard into a JMP data table or report. **Paste** can also be used with the **Copy** command to duplicate rows, columns, or any subset of cells defined by selected rows and columns.

To transfer data from another application into a JMP data table, first copy the data to the clipboard from within the other application. Then use the **Paste** command to copy the values to a JMP data table. Rows and columns are automatically created as needed. If you choose **Paste** while holding down the Shift (Windows and Linux) or Option (Macintosh) key, the first line of information on the clipboard is used as column names in the new JMP data table.

To duplicate an entire row or column:

1. Select and **Copy** only the row or column to be duplicated.
2. Select an existing row or column to receive the values.
3. Select **Paste**.

To duplicate a subset of values defined by selecting specific cells, follow the previous steps, but select an identical arrangement of cells to receive the pasted values. If you paste data with fewer rows into a destination with more rows, the source values repeat until all receiving rows are filled.

Drag

You can also move or duplicate rows and columns by dragging. Hold the mouse down in the selection area (above the column name) of one or more selected rows or columns and drag them to a new position in the data table. Use Ctrl-drag (Option-drag on the Macintosh) to duplicate rows and columns instead of moving them.

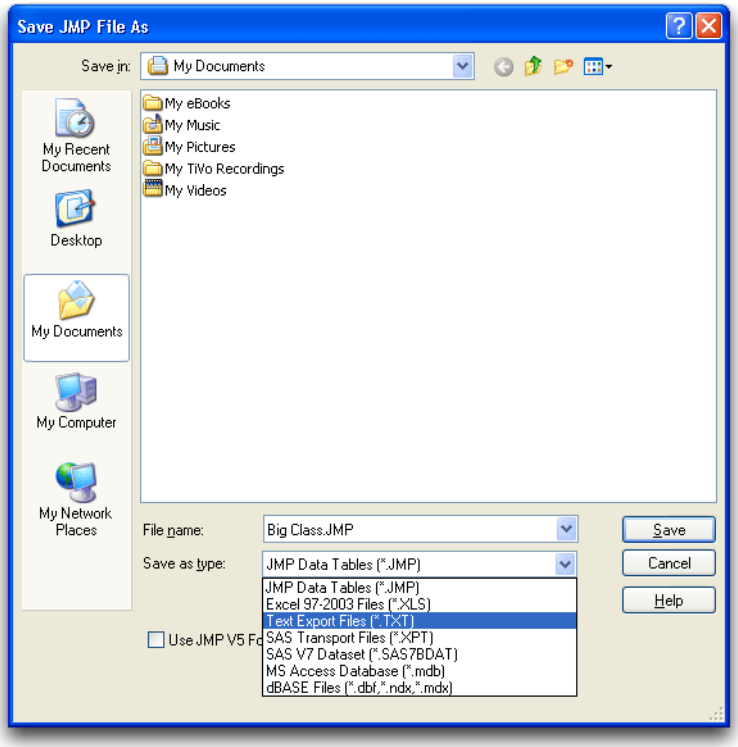
Moving Data Out of JMP

The **Save As** command saves the active data table to a file after prompting you for a name and file type. JMP can save a data table as a JMP file, convert it to a SAS Transport file, Excel file, text file, or save it in any database format available on your system.

Windows

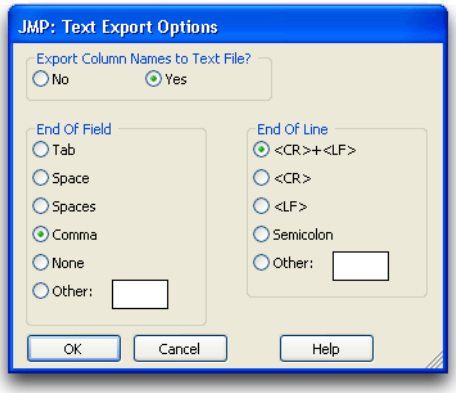
JMP can save data in any of the following formats:

Figure 3.13 Save As Dialog (Windows)



- **JMP Data Tables** saves the table in JMP format. This is the default **Save As** option.
- **Excel 97-2003 Files** saves data tables in Microsoft Excel .xls format. The resulting file is directly readable by most versions of Excel, including Excel 97, 98, 2000, and 2003.
- **Text Export Files** converts data from a JMP file to a standard text format, with rows and columns.

Figure 3.14 Text Export Options (Windows)



The **Options** button in the **Save As** dialog displays choices to describe specific text arrangements:

Export Column Names to Text File has **Yes** and **No** radio buttons to request that JMP column names be written as the first record of the text file, or that no labels or header information be saved with the data.

End of Field and **End of Line** designate the characters to identify the end of each field and end of line in the saved text file. These options are described previously in the section "Importing Data" on page 38.

- **SAS Transport Files** converts a JMP data table to SAS transport file format and saves it in a SAS transport library. The **Append To** option appends the data table to an existing SAS transport library. If you don't use **Append To**, a new SAS transport library is created using the name and location you provide. If you do not specify a new file name, the SAS transport library replaces the existing JMP data table.
- **SAS V7 Data Set** saves the data as a SAS 7 data set, readable by SAS 7 or later.
- JMP can also save data in database formats that have ODBC drivers installed on your system, such as **Microsoft FoxPro**, **dBase**, and **Microsoft Access**.

Macintosh

JMP can save data as a JMP data table, SAS Transport file, text file, or an Excel file.

- To save data as a JMP data table, choose **File > Save As**.
- To save data as text, Excel, or SAS Transport, select **File > Export**, then select the appropriate format from the sheet that appears.

Figure 3.15 Macintosh Export Menu

Linux

JMP can save data as a JMP data table, text file, or SAS Transport file. For example, to save a text file, select **File > Save As**, then select **.txt (Delimited File)** from the **Save As Type** menu.

Working with Graphs and Reports

You can use standard copy and paste operations to move graphical displays and statistical reports from JMP to other applications. Although the **Edit** menu includes both **Cut** and **Copy** commands, they both perform the same tasks in report windows. **Cut** copies all or selected parts of the active report window into the clipboard, including the images.

Copy and Paste

When you copy from a report (results) window, the information is stored on the clipboard. If you want to copy part of a report window, use the selection tool(✥) from the **Tools** menu or toolbar. Click on the area you want to copy, shift-click to extend the selected area, use the **Copy** command to copy the selected area to the clipboard, then use the **Paste** command to paste the results into a JMP journal or another application.

Drag Report Elements

Any element in a JMP report window that can be selected can be dragged. When you drag report elements within the same report window, they are copied to the destination area where you drop them. As you drag an element, a visual cue shows where the element would be dropped.

You can copy and paste any report element to other applications, and drag and drop JMP reports and graphs to any other application that supports drag-and-drop operations.

The format used when pasting depends on the application you paste into. If the application has a **Paste Special** command, you can select among paste formats such as rich text, which includes pictures (RTF), unformatted text (TXT), picture (PICT or WMF), bitmap (BMP), and enhanced picture (EMF).

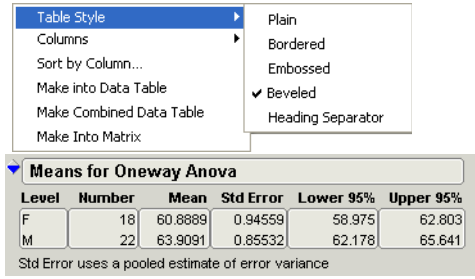
To delete a copied report element, select it and press the delete key on the keyboard.

Context Menu Commands

Right-click using the arrow tool (Control-click on the Macintosh) on a report window to see the context menu shown in the following examples. The context menu changes depending on where you click (hence its name, "context menu"). If you are not over a display element with its own context menu, the menu for the whole platform is shown.

Context Commands for Report Tables

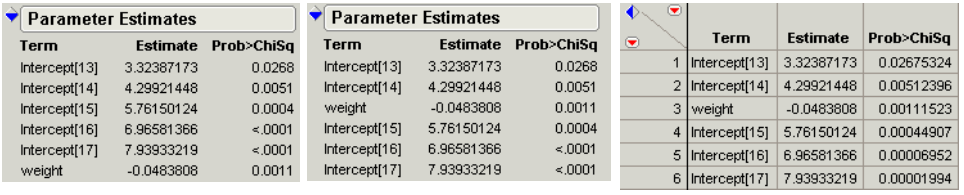
By default, the tables in JMP reports have no formatting to separate rows and columns. Some (or, in many cases, all) available columns for the report are showing. Context menu items for report tables let you tailor the appearance and content of the tables as follows:



- **Table Style** lets you enhance the appearance of a table by drawing borders or other visual styles to the table rows and columns. The example shown to the right has beveled column separators.
- **Columns** lets you choose which columns you want to show in the analysis table. Analysis tables often have many columns, some of which may be initially hidden. The leftmost table in **Figure 3.16** is a Parameter Estimates table showing only the estimate name, the estimate itself, and the probability associated with the estimate. The standard error and chi-square values (shown by default) are hidden.

- **Sort by Column** lets you sort the rows of a report table. This command displays a list of visible columns in a report and lets you choose one or more columns to sort by. The middle table in **Figure 3.16** is the Parameter Estimates table on the left sorted by **Prob>ChiSq**.
- **Make into Data Table** lets you create a JMP data table from any analysis table. The rightmost data table in **Figure 3.16** is the JMP data table created from the sorted Parameter Estimates table with hidden columns.

Figure 3.16 Results of Context Commands for Analysis Tables



Parameter Estimates

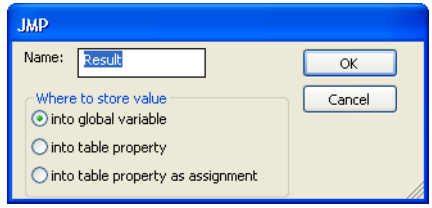
Term	Estimate	Prob>ChiSq
Intercept[13]	3.32387173	0.0268
Intercept[14]	4.29921448	0.0051
Intercept[15]	5.76150124	0.0004
Intercept[16]	6.96581366	<.0001
Intercept[17]	7.93933219	<.0001
weight	-0.0483808	0.0011

Parameter Estimates

Term	Estimate	Prob>ChiSq
Intercept[13]	3.32387173	0.0268
Intercept[14]	4.29921448	0.0051
weight	-0.0483808	0.0011
Intercept[15]	5.76150124	0.0004
Intercept[16]	6.96581366	<.0001
Intercept[17]	7.93933219	<.0001

	Term	Estimate	Prob>ChiSq
1	Intercept[13]	3.32387173	0.02675324
2	Intercept[14]	4.29921448	0.00512396
3	weight	-0.0483808	0.00111523
4	Intercept[15]	5.76150124	0.00044907
5	Intercept[16]	6.96581366	0.00006952
6	Intercept[17]	7.93933219	0.00001994

- **Make into Matrix** lets you store a report table as a matrix that is useful when you are using the JMP scripting language (JSL). When selected, the dialog shown here appears, allowing you to designate the name of the matrix and where it should be stored.



Juggling Data Tables

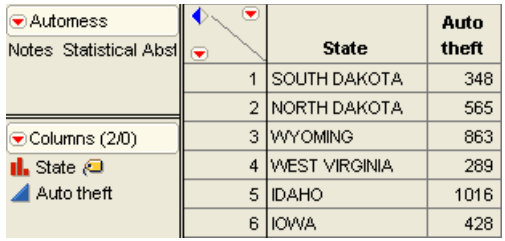
Each of the following examples uses commands from the **Tables**, **Rows**, or **Cols** menus.

Data Management

Suppose you have the following situation. Person A began a data entry task by entering state names in order of ascending auto theft rates. Then, Person B took over the data entry, but mistakenly entered the auto theft rates in alphabetical order by state.

To see the result, open the Automess.jmp sample table.

Could this ever really happen? Never underestimate the diabolical convolution of data that can appear in an electronic table, and hence a circulated report. Always check your data with common sense.



	State	Auto theft
1	SOUTH DAKOTA	348
2	NORTH DAKOTA	565
3	WYOMING	863
4	WEST VIRGINIA	289
5	IDAHO	1016
6	IOWA	428

To put the data into its correct order, you need to make a copy of the Automess table, sort it in ascending order, and join the sorted result with the original table.

- With Automess.jmp active, choose **Tables** > **Subset** and press **OK**.

This automatically creates a duplicate table since no rows or columns were selected in the original table.

The non-descriptive table name is Subset of Automess.JMP, but you don't need to give this table a descriptive name because it is only temporary.

- With this subset table active, choose **Tables** > **Sort**.
- When the Sort dialog appears, choose Auto theft as the sort variable and click **Sort**.

There is now an untitled table that is sorted by auto theft rates in ascending order.

- Close Subset of Automess.JMP, as it is no longer needed.

Now you want to join the incorrectly sorted Automess.JMP table with the correctly sorted Untitled table. You do this as follows:

- Choose **Tables** > **Join**.
- When the Join dialog appears, note which table is listed next to the word Join (either Automess or Untitled), then click the other table's name in the list of tables.
- Because you don't want all the columns from both tables in the final result, click the **Select Columns** button.

The variables from both tables appear in list boxes.

- Select State from the Automess table, Auto theft from the Untitled table, and click **Add**.
- Click **Done** to close the Select Columns dialog.

- When the Join dialog appears again, click the **Join** button.

Check the new joined data table: the first row is South Dakota with a theft rate of 110, and the last row is the District of Columbia with a rate of 1336. If you want to keep this table, use **Save As** and specify a name and folder for it.

Give New Shape to a Table: Stack Columns

A typical situation occurs when response data are recorded in two columns and you need them to be stacked into a single column. For example, suppose you collect three months of data and enter it in three columns. If you then want to look at quarterly figures, you need to change the data arrangement so that the three columns stack into a single column. You can do this with the **Stack** command in the **Tables** menu.

An example of stacking columns follows:

- Open the Chezsplt.jmp sample data to see the table on the left in **Figure 3.17**.

This sample data (McCullagh and Nelder, 1983) has columns for four kinds of cheese, labeled A, B, C, and D. In a taste test, judges ranked the cheeses on an ordinal scale from 1 to 9 (1 being awful, and 9 being wonderful). The Response column shows these ratings. The counts for each cheese and for each ranking of taste are the body of the table. Its form looks like a two-way table, but to analyze it JMP needs the cheese categories in a single column. To rearrange the data:

- Choose **Tables** > **Stack**.
- In the dialog that appears, select the cheeses (A, B, C, and D) from the **Select Columns** list and add them to the **Stack Columns** list. Leave everything else as is.
- Click **OK** to see the table on the right in **Figure 3.17**.

Figure 3.17 Stack Columns Example

Chezsplt
Notes This sample da

Columns (5/0)
Response
A
B
C
D

	Response	A	B	C	D
1	1	0	6	1	0
2	2	0	9	1	0
3	3	1	12	6	0
4	4	7	11	8	1
5	5	8	7	23	3
6	6	8	6	7	7
7	7	19			
8	8	8			
9	9	1			

Untitled
Source

Columns (3/0)
Response
Label
Data

	Response	Label	Data
1	1	A	0
2	1	B	6
3	1	C	1
4	1	D	0
5	2	A	0
6	2	B	9
7	2	C	1
8	2	D	0
9	3	A	1
10	3	B	12

The Label column shows the cheeses, and the Data column is now the count variable for the response categories.

- Right-click (Control-click on the Mac) in the Data column, and select **Preselect Role** > **Freq** to change the Data column to represent frequency.

This causes the values in the Data column in the Untitled table to be interpreted by analyses as the number of times that row's response value occurred.

To see how response relates to type of cheese:

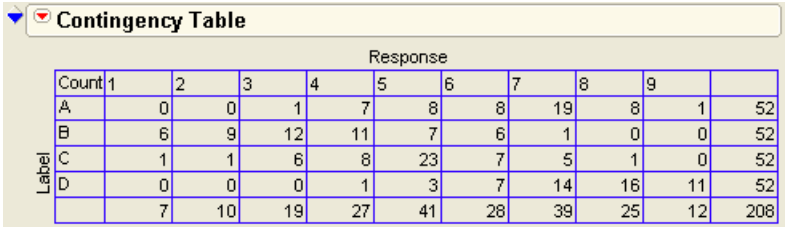
- Choose **Analyze** > **Fit Y by X**.
- In the Fit Y by X launch dialog select Response as **Y, Response** and Label as **X, Factor**.

If you preselected its role, **Data** is already assigned as a **Freq** variable. If not, make it so.

When you click **OK**, the contingency table platform appears with a Mosaic plot, Crosstabs table, Tests table, and menu options. To find more information about the platform components, you can use the Help tool (labeled with a question mark) in the Tools menu and

click on the platform surface. A simplified version of the Crosstabs table showing only counts is shown in **Figure 3.18**. (Right-click on the contingency table to modify what it displays.) The Cheese data is used again later for further analysis.

Figure 3.18 Contingency Table for the Cheese Data



Contingency Table

		Response									
	Count	1	2	3	4	5	6	7	8	9	
Label	A	0	0	1	7	8	8	19	8	1	52
	B	6	9	12	11	7	6	1	0	0	52
	C	1	1	6	8	23	7	5	1	0	52
	D	0	0	0	1	3	7	14	16	11	52
		7	10	19	27	41	28	39	25	12	208

Extra Credit: For practice, see if you can use the **Split** command on the stacked data table to reproduce a copy of the Chezsplt table.

The Summary Command

One of the most powerful and useful commands in the **Tables** menu is the **Summary** command.

Summary creates a JMP window that contains a summary table. This table summarizes columns from the active data table, called its *source table*. It has a single row for each level of a grouping variable you specify. A grouping variable divides a data table into groups according to each of its values. For example, a gender variable can be used to group a table into males and females.

When there are several grouping variables (for example, gender and age), the summary table has a row for each combination of levels of all variables. Each row in the summary table identifies its corresponding subset of rows in the source table. The columns of the summary table are summary statistics that you request.

Create a Table of Summary Statistics

The example data used to illustrate the **Summary** command is the JMP table called Companies.jmp (see **Figure 3.19**).

Open the Companies.jmp sample table.

It is a collection of financial information for 32 companies (Fortune 1990). The first column (Type) identifies the type of company with values “Computer” or “Pharmaceut.” The second column (Size Co) categorizes each company by size with values “small,” “medium,” and “big.” These two columns are typical examples of grouping information.

Figure 3.19 JMP Table to Summarize

Companies
Notes Selected Data

Columns (8/0)
- Type
- Size Co
- Sales ($M)
- Profits ($M)
- # Employ
- profit/emp
- Assets
- %profit/sales

Rows
All rows 32

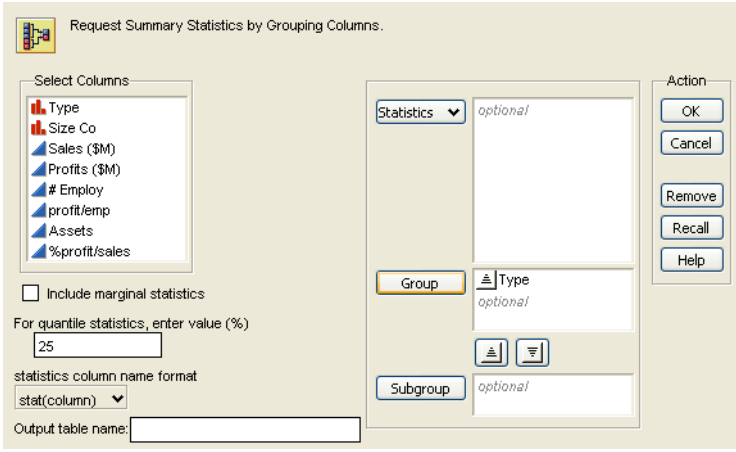
	Type	Size Co	Sales ($M)	Profits ($M)	# Employ	profit/emp	Assets	%profit/ sales
1	Computer	small	855.1	31.0	7523	4120.70	615.2	3.63
2	Pharmaceut	big	5453.5	859.8	40929	21007.11	4851.6	15.77
3	Computer	small	2153.7	153.0	8200	18658.54	2233.7	7.10
4	Pharmaceut	big	6747.0	1102.2	50816	21690.02	5681.5	16.34
5	Computer	small	5284.0	454.0	12068	37620.15	2743.9	8.59
6	Pharmaceut	big	9422.0	747.0	54100	13807.76	8497.0	7.93
7	Computer	small	2876.1	333.3	9500	35084.21	2090.4	11.59
8	Computer	small	709.3	41.4	5000	8280.00	468.1	5.84
9	Computer	small	2952.1	-680.4	18000	-37800.0	1860.7	-23.05
10	Computer	small	784.7	89.0	4708	18903.99	955.8	11.34
11	Computer	small	1324.3	-119.7	13740	-8711.79	1040.2	-9.04
12	Pharmaceut	medium	4175.6	939.5	28200	33315.60	5848.0	22.50

- Choose **Tables** > **Summary**.
- When the Summary dialog appears (**Figure 3.20**), select the variable Type in the Columns list of the dialog and click **Group** to add it to the grouping variables list.

In practice, you can select as many grouping variables as you want.

- Click **OK** to see the summary table.

Figure 3.20 Summary Dialog and Summary Table



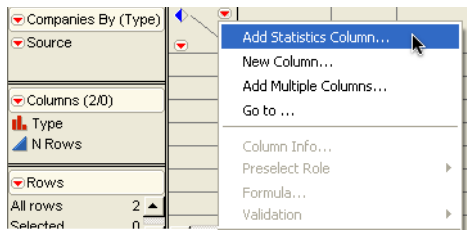
	Type	N Rows
1	Computer	20
2	Pharmaceut	12

The new summary table appears in an active window. This table is linked to its source table. When you highlight rows in the summary table, the corresponding rows are also highlighted in its source table.

Initially, a summary displays frequency counts (N Rows) for each level of the grouping variables. This example shows 20 computer companies and 12 pharmaceutical companies. However, you can add columns of descriptive statistics to the table. The **Statistics** popup menu in the Summary dialog lists standard univariate descriptive statistics.

To add summary statistics to an existing summary table, follow these steps:

- Select **Add Statistics Column** command from the menu in the upper left-hand corner of the summary table.



This command displays the Summary dialog again.

- Select any numeric column (for example, Profit $M) from the source table columns list.

- Select the statistic you want (for example, **Sum**) from the **Statistics** menu on the dialog.
- If desired, repeat to add more statistics to the summary table.
- Click **OK** to add the columns of statistics to the summary table.

The table in **Figure 3.21** shows the sum of Profits ($M) in the summary table grouped by Type.

Figure 3.21 Expanded Summary Table

	Type	N Rows	Sum(Profits ($M))
1	Computer	20	4817.3
2	Pharmaceut	12	8280.9

Another way to add summary statistics to a summary table is with the **Subgroup** button in the Summary dialog. This method creates a new column in the summary table for each level of the variable you specify with **Subgroup**. The subgroup variable is usually nested within all the grouping variables.

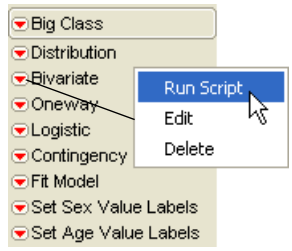
Working with Scripts

JMP contains a full-fledged scripting language, used for automating repetitive tasks and scripting instructional simulations. Several scripts are featured throughout this book to demonstrate statistical concepts.

Scripts are stored in two formats:

- attached to a data table
- as a stand-alone scripting file

Scripts that are attached to a data table are displayed in the Tables panel, as shown to the right. This sample is from the Big Class.jmp sample data table, showing six scripts that have been saved with it.



To run an attached script:

- Click the button beside the script's name.

- Select **Run Script** from the menu that appears.

Stand-alone scripts are stored as simple text files with the JSL extension. They may be opened and run independently of a data table.

Opening and Running Scripts on Windows and Linux

To open and run a stand-alone script on Windows:

- Select **File > Open**.
- From the **Files of Type** drop-down list, select **JSL Scripts (*.JSL)**, ***.JSL (JSL Script)** or **JMP Files (*.JMP, *.JSL, *.JRN)**.
- Double-click the name of the script to open.

The script opens in a script editor window. To execute the script:

- Select **Edit > Run Script** or press the shortcut key **Ctrl-R**.

Opening and Running Scripts on the Macintosh

To open and run a stand-alone script on the Macintosh:

- Select **File > Open.**
- Make sure the **Show** list displays **All Readable Documents**.
- Double-click the script to open.

The script opens in a script editor window. To execute the script:

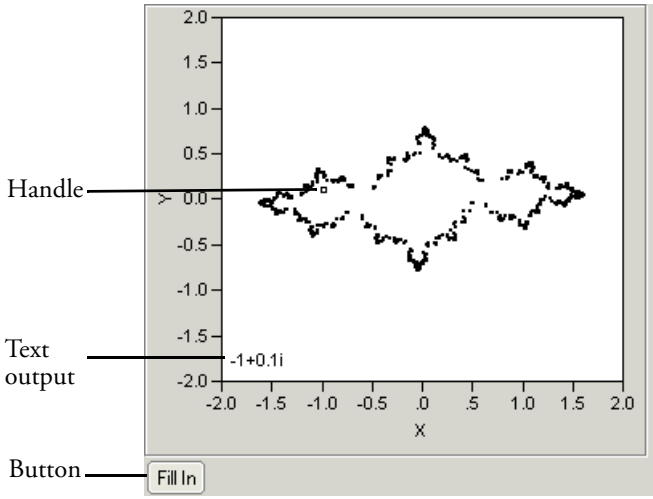
- Select **Edit > Run Script** or press the shortcut key **⌘-R**.

As an example, use the Julia Sets.jsl script stored in the Sample Scripts folder.

- Open and run the Julia Sets.jsl script.

The resulting window (shown in **Figure 3.22**) illustrates several key features of a typical instructional script.

Figure 3.22 Julia Set Script Window



- A *handle* is a draggable script element that updates the display as it is dragged. In this case, the handle represents the seed value for the Julia set, whose shape changes based on the value of the handle's coordinates.
- *Text output* is sometimes drawn directly on the graphics screen rather than displayed as reports below the window. This script shows the seed value (an imaginary number) in the lower left corner of the window.
- *Buttons* reveal options, set conditions, or trigger actions in the script. In this example, the **Fill In** button draws more details of the Julia set.

To practice with these elements:

- Click and drag the handle to different places in the window and observe how the Julia set changes.
- When the Julia set has an interesting shape, press the **Fill In** button to reveal its details.

4 Formula Editor Adventures

Overview

Each column has access to the *Formula Editor*. The JMP Formula Editor is a powerful tool for building formulas that calculate values for each cell in a column. The Formula Editor window operates like a pocket calculator with buttons, displays, and an extensive list of easy-to-use features for building formulas.

JMP formulas can be built using other columns in the data table, built-in functions, and constants. Formulas can be simple expressions of numeric, character, or row state constants or can contain complex evaluations based on conditional clauses. Once created, the formula remains with the column until the formula is deleted. It is visible in both the Column Info dialog and in the formula editor itself.

A column whose values are computed using a formula is both *linked* and *locked*. It is linked to (or dependent on) all other columns that its formula refers to. Its values are automatically recomputed whenever you edit the values in these columns. It is also locked, so its data values cannot be edited, which would invalidate its formula.

This chapter describes Formula Editor features and gives a variety of examples. See the online *JMP User Guide* for a complete list of Formula Editor functions.

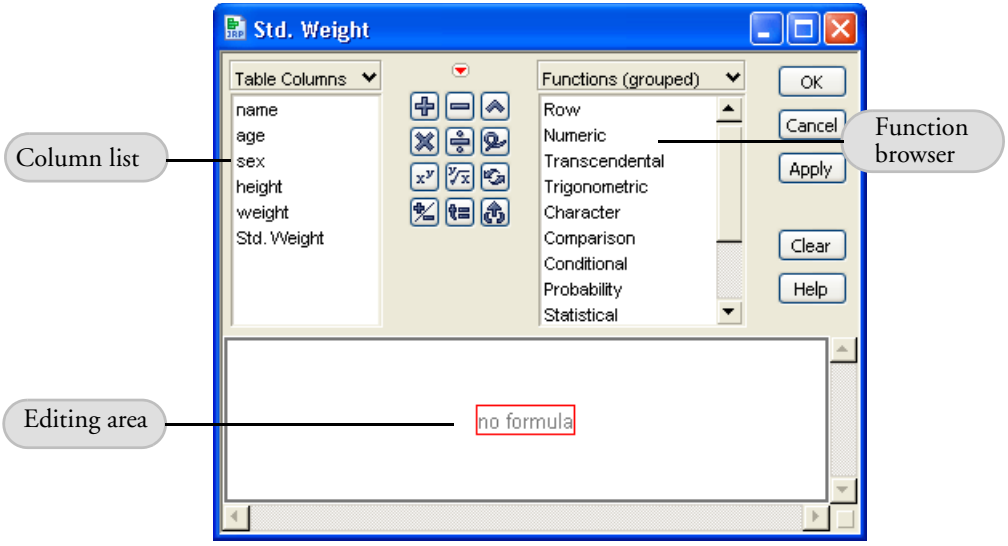
The Formula Editor Window

The JMP Formula Editor is where you create or modify a formula. You can open the Formula Editor for a column in three ways:

- Select **Formula** from the **Cols** menu for one or more selected columns.
- Select **Formula** from the **New Property** popup menu in a New Column dialog and click the **Edit Formula** button that appears.
- Right-click (Windows and Linux) or Control-click (Macintosh) in the heading of a column and select **Formula** from the context menu that appears. This opens the Formula Editor window without first opening the Column Info dialog.

The Formula Editor window is divided into two areas: the *control panel,* consisting of the column list and function browser, and the *formula display*, an area for editing formulas. **Figure 4.1** shows the parts of the Formula Editor. The Formula Editor control panel is composed of buttons (**OK**, **Apply**, **Help**), selection lists for variables and functions, and a keypad. The formula display is an editing area you use to construct and modify formulas.

Figure 4.1 The Formula Editor Window



The sections that follow show you a simple example, define Formula Editor terminology, and give the details for using the control panel and the formula display.

A Quick Example

The following example gives you a quick look at the basic features of the Formula Editor. Suppose you want to compute a *standardized* value. That is, for a set of numeric variables x_i you want to compute

$$\frac{x_i - \bar{x}}{s_x}$$

where $\bar{x}$ is the mean of $x_1, x_2, x_3, \ldots, x_i$

s_x is the standard deviation of $x_1, x_2, x_3, \ldots, x_i$

for each row in a data table.

- For this example, open the Students.jmp data table.

It has a column called weight, and you want a new column that uses the above formula to generate standardized weight values.

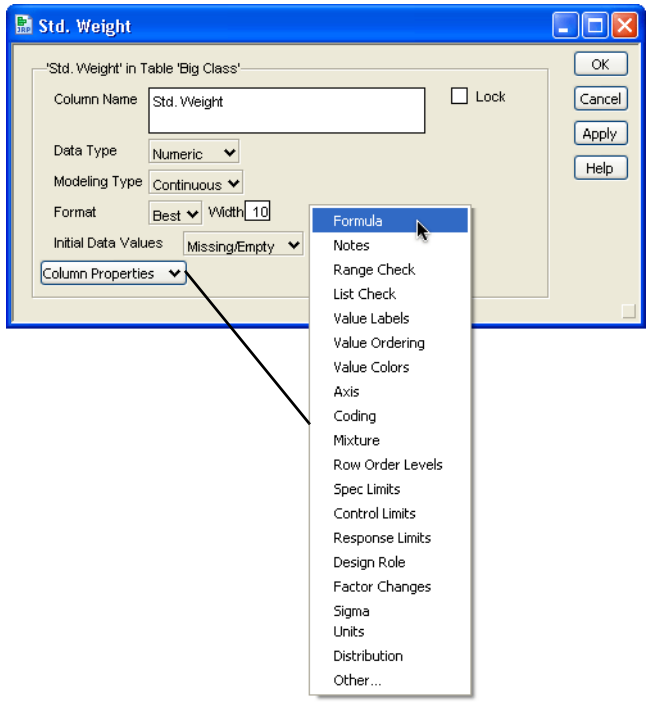
- Begin by choosing **Cols** > **New Column**, which displays a New Column dialog like the one shown in **Figure 4.2**.

The New Column dialog lets you set the new column's characteristics.

- Type the new name, Std. Weight, in the **Column Name** area.

The other default column characteristics define a numeric continuous variable and are correct for this example.

Figure 4.2 The New Column Dialog

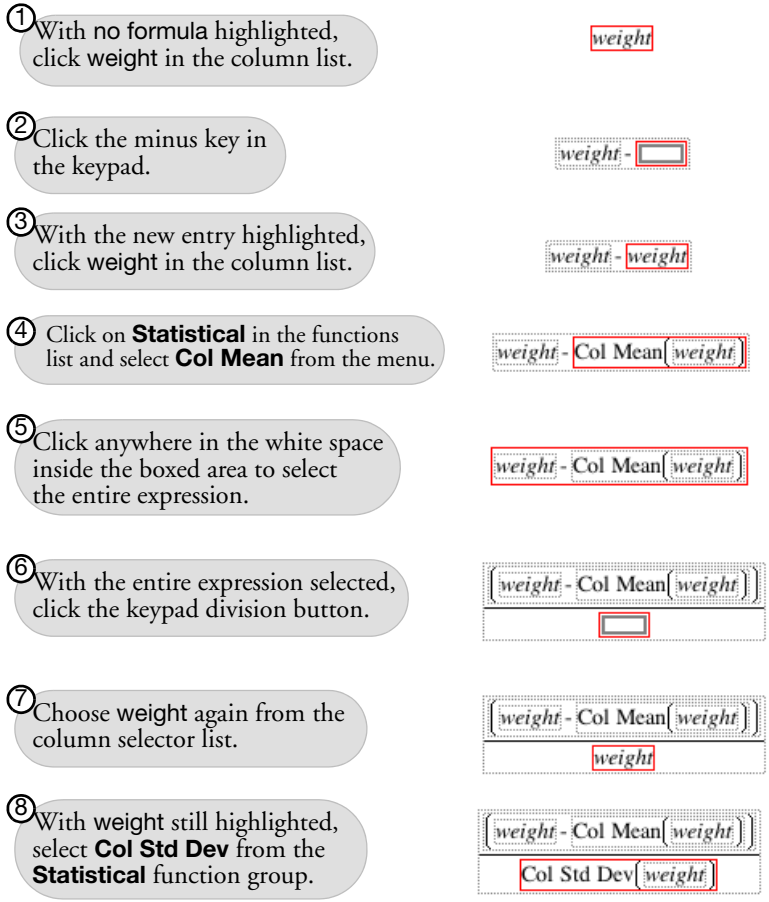


Select **Formula** from the **Column Properties** popup menu.

This opens the Formula Editor window shown in **Figure 4.1**.

Next, enter the formula that standardizes the weight values by following the steps in **Figure 4.3**.

Figure 4.3 Entering a Formula



You have now entered your first formula.

- Close the Formula Editor window by clicking the **OK** button.

The new column fills with values. If you change any of the weight values, the calculated Std. Weight values automatically recompute.

If you make a mistake entering a formula, choose **Undo** from the **Edit** menu. There are other editing commands to help you modify formulas, including **Cut**, **Copy**, and **Paste**. The Delete key removes selected expressions. If you need to rearrange terms or expressions, you can select and drag to move formula pieces.

This example may be all you need to proceed. However, the rest of the chapter covers details about the Formula Editor, and gives a variety of other examples. Complete documentation of the Formula Editor is found in the *JMP User Guide*.

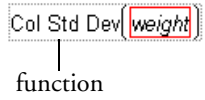
Formula Editor: Pieces and Parts

This section begins with Formula Editor terminology. The Formula Editor has distinct areas, so we also describe its geography. Lastly, this section gives a brief description of all the function categories. Later sections give examples of specific functions.

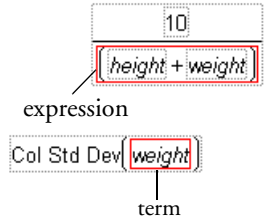
Terminology

The following list is a glossary of terminology used when discussing the Formula Editor:

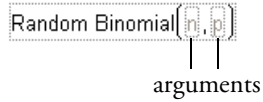
- A *function* is a mathematical or logical operation that performs a specific action on one or more arguments. Functions include most items in the function browser and all keypad operators.



- An *expression* is a formula (or any part of it) that can be highlighted as a single unit, including terms, empty terms, and functions grouped with their arguments.
- A *term* is an indivisible part of an expression. Constants and variables are terms.

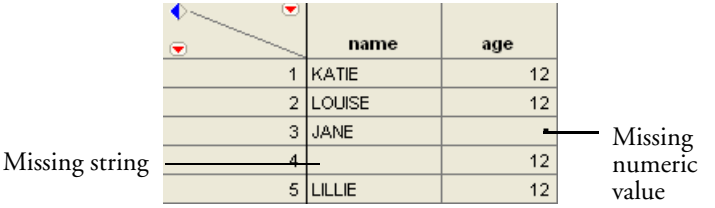


- An *argument* is a constant, a column, or expression (including mathematical operands) that is operated on by the function.



- An *empty term* is a placeholder for an expression, represented by a small empty box.
- A *missing value* in a table cell shows as a missing value mark (a large dot) for numeric data, or a null character string for character data.

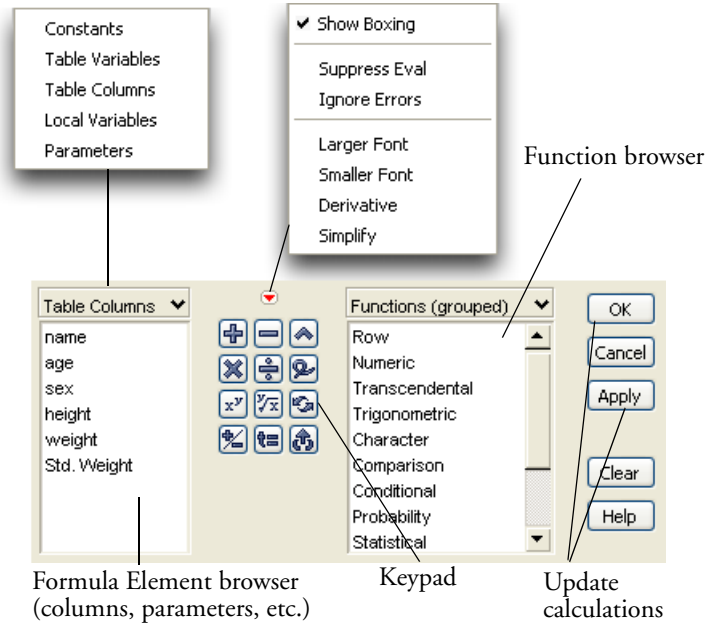
Figure 4.4 Missing Values



The Formula Editor Control Panel

The top part of the Formula Editor is called the *control panel*. It is composed of buttons and selection lists as illustrated in **Figure 4.5**.

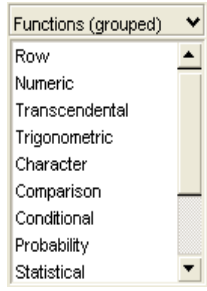
Figure 4.5 The Formula Editor Control Panel



Some of the Formula Editor features, like those in the keypad, behave like those on a hand-held calculator. Other features are unique to the JMP Formula Editor. The following descriptions refer to **Figure 4.5**.

The *Formula Element browser* displays selection lists of table columns, constants, variables or parameters. By default, the list of table columns is visible. You can change the kind of elements listed by choosing from the popup menu at the top of the formula element browser. To choose a formula element, select an expression in the formula editing area, then select an element in the function list (see **Figure 4.5**).

The *keypad* is a set of buttons used to build formulas. Some of the buttons, such as the arithmetic operators, are familiar. Others have special functions, described in the next section.



The *Function browser* groups collections of functions and features in lists organized by topic. To use a function in a formula, select an expression in the editing area and click any item in one of the function browser topics. You can also see a list of ungrouped functions in alphabetical order by selecting **Functions (All)** from the popup menu above the function list.

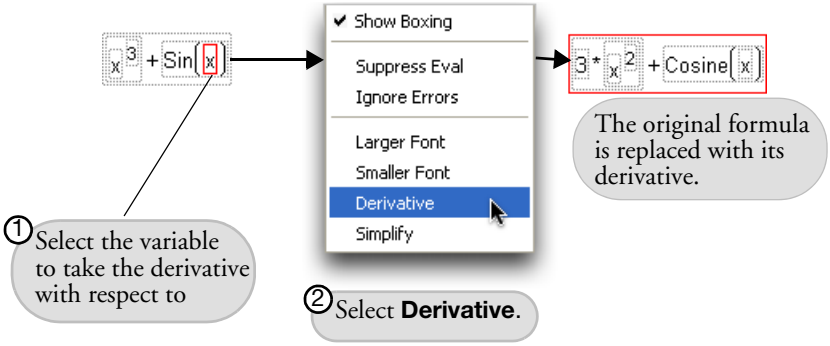
The data table columns automatically fill with calculated values whenever you change a formula and then close the Formula Editor window or make it inactive. Use the **Apply** button to calculate a column's values if you want the Formula Editor window to remain open.

The popup menu above the keypad (shown in **Figure 4.5**) has these commands:

- **Show Boxing** outlines terms within the formula. Boxing is important when you want to select and modify a specific portion of a formula, or need to determine the order of evaluation that takes place.
- **Suppress Eval** suppresses formula evaluation unless you specifically click **Apply** on the Formula Editor. This is a useful development mode when building complex formulas. You can turn off evaluation and build sections of a formula, then evaluate only to test the formula. In particular, you can close the Formula Editor and reopen it at a later time to continue building a formula without evaluating the formula.
- **Ignore Errors** suppresses error messages while a formula is under development. This is useful in situations where you want to see an evaluation for some rows and don't want to see an error message for every row where the formula evaluation finds problems. If you don't select **Ignore Errors**, an error message dialog appears when there is an error and asks if you want to ignore further errors. This has the same effect as the **Ignore Errors** menu selection.
- **Larger Font** increases the font size of the displayed formula.
- **Smaller Font** decreases the font size of the displayed formula.

- **Derivative** takes the first derivative of the entire formula. To use this command, first select a variable for the derivative to be taken with respect to. Then, select the **Derivative** command from the menu. This procedure is illustrated in **Figure 4.6**.

Figure 4.6 Derivatives in the Formula Editor



- **Simplify** algebraically simplifies the formula. For example, the **Simplify** command turns 2*x + 3*x into 5*x.

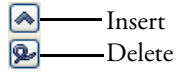
The Keypad Functions

The keypad is composed of common operators (referred to as *functions* here). Enter a keypad function by selecting an expression in the formula display and clicking on the appropriate keypad buttons.

Arithmetic keys

The four arithmetic functions work as they do on a pocket calculator.

Insert and Delete keys



The **Insert** button inserts a new empty formula clause or function argument. To insert a clause into a formula, first select the existing clause or argument you want the new element to follow. When you click the **Insert** button, the new clause appears and is selected. (Hint: You can also insert a new clause or argument by using the **Insert** keyboard shortcut: a comma.) The **Delete** button empties the selected box, or, if it is already empty, removes it as an argument.

Raise to a Power

The general exponential function raises a given value to a specified power. It has an exponent of two by default. Select the exponent and double-click on it to change its value.

Root

The root function calculates the specified root of the radicand. It has an implied index of 2, which is not displayed. To change the index to another value, highlight the argument of the root (the part outside the radical) and double-click to change its value.

Switch Terms

The switch terms keypad function looks at the operator that is central to the selected expression and switches the expressions on either side of that operator. For example, switching *a* + *b* results in *b* + *a.*

Unary Sign Function

The unary sign function inverts the sign of its argument. Apply the function to a selected variable expression or use it to enter negative constants.

Local Variable Assignment Key

This keypad function creates a local variable and assigns it the value of the selected expression. Its value can be as simple as an empty term, or as complicated as a complex formula.

Peel Expression

To use this function, begin by selecting any expression. When you click the **Peel Expression** button, the selected expression is deleted, leaving a selected empty term in its place. This process repeats each time the key is clicked. In this way, you can delete a formula term by term, in the precedence order of the formula, beginning with the first term you select. See the section “Tips on Editing a Formula” on page 90 for a demonstration of peeling expressions.

The Formula Display Area

The formula display area is where you build and view a formula. To compose a formula, select expressions in the formula display area and apply functions and terms from the formula control panel.

Functions always operate upon selected expressions, terms always replace selected expressions, and arguments are always grouped with functions. To find which expressions serve as a

function's arguments, select that function in the formula. When the **Show Boxing** option is in effect, the boxed groupings also show how the order of precedence rules apply and show which arguments are deleted if you delete a function.

Function Browser Definitions

The function browser groups the Formula Editor functions by topic. To enter a function, highlight an expression and click any item in the function browser topics. Examples of some commonly used functions are included later in this chapter.

The function categories are briefly described in the following list. They are presented in the order you find them in the function browser.

- **Row** lists miscellaneous functions such as **Lag**, **Dif**, **Subscript**, **Row** (the current row number), and **NRow** (the total number of rows).
- **Numeric** lists functions such as **Round**, **Floor**, **Ceiling**, **Modulo** and **Absolute Value**.
- **Transcendental** supports logarithmic functions for any base, functions for combinatorical calculations, the beta function, and several gamma functions.
- **Trigonometric** lists the standard trigonometric and hyperbolic functions such as sine, cosine, tangent, inverse functions, and their hyperbolic equivalents.
- **Character** lists functions that operate on character arguments for, among other things, trimming, finding the length of a string, and changing numbers to characters or characters to numbers.
- **Comparison** lists the standard logical comparisons such as less than, less than or equal to, not equal to, and so forth.
- **Conditional** lists the logical functions **Not**, **And**, and **Or**. They also include programming-like functions such as **If/then/else**, **Match**, and **Choose**.
- **Statistical** lists functions that calculate standard statistical quantities such as the mean or standard deviation, both down columns or across rows.
- **Probability** lists functions that compute probabilities and quantiles for the Beta, chi-square, *F*, gamma, Normal, Student's *t,* and a variety of other distributions.
- **Random** is a collection of functions that generate random numbers from a variety of distributions.
- **Date Time** are functions that require arguments with the date data type, which is interpreted as the number of seconds since January 1, 1904. You assign Date as the Data

Type in a New Column or Column Info dialog. Date functions return values such as day, week, or month of the year, compute dates, and can find date intervals.

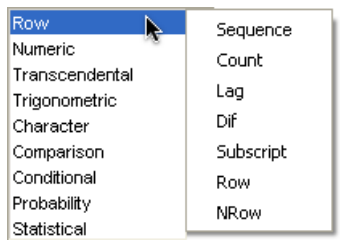
- **Row State** lists functions that assign or detect special row characteristics called *row states*. Row states include color, marker, label, hidden (in plots), excluded (from analyses), and selected or not selected.
- **Assignment** functions work in place. That is, the result returned by the operation (on the right of the operator) is stored in the argument on the left of the operator and replaces its current value. They are named constants that you create and can use in any formula.

Row Function Examples

To do the next examples, create an empty data table and insert some rows and columns.

- Choose **File > New > (New) Data Table**.
- When the new table appears, choose **Rows** > **Add Rows**, and ask for ten rows.
- Choose **Cols > Add Multiple Columns** and ask for nine new columns.

The first category in the function browser is called **Row**. When you click **Row** in the function browser, you see the list of functions shown to the right.



Lag(column,n)

The `Lag` function returns the value of the column in the row defined by the current row less the second argument. That is, it returns the value of column *n* rows back. The default lag is one, which you can change to any number. The value returned for any lag that identifies a row number less than one is missing. Note that `Lag(X,n)` gives the same result as the subscripted notation, $X_{Row()-n}$. But `Lag` is more general, supporting entire expressions as well as simple column names.

Row()

is the current row number when an expression is evaluated for that row. You can incorporate this function in any expression, including those used as column name subscripts (discussed in the next section).

Dif(column,n)

Returns the difference between the value of the column in the current row and the value *n* rows previous.

NRow ()

is the total number of rows in the data table.

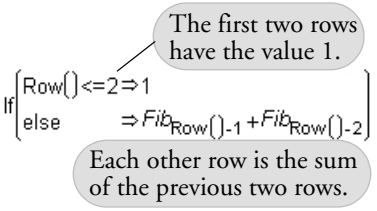
Subscript

enables you to use a column's value from a row other than the current row. Highlight a column name in the formula display and click **Subscript** to display a placeholder for the subscript. The placeholder can be changed to any numeric expression. Subscripts that evaluate to nonexistent row numbers produce missing values or, possibly, error messages. A column name without a subscript refers to the current row. To remove a subscript from a column, select the subscript and delete it. Then delete the empty box that remains. The formula

$$\texttt{Count}_{\text{Row( )}}-\texttt{Count}_{\text{Row( )}-1}$$

uses the subscript of `Row()-1` to calculate the difference of two successive rows in a column named Count. Note that the subscript of Row() alone is not essential, since it refers to the current row, which is the default behavior.

The following formula calculates values for a a column called Fib, which, after the formula is evaluated, contains the terms of the Fibonacci series (each value is the sum of the two preceding values in the calculated column).



It shows the use of subscripts to do recursive calculations. A recursive formula includes the name of the calculated column, subscripted such that it references previously evaluated rows.

Using a Subscript

Use a subscript to refer to a specific row of the subscripted column. A simple use for subscripts is to create a lag variable similar to the Lag function. Follow these steps to create a lag variable:

- Give the first column in the empty data table the name Row ID.
- With the column selected, choose **Cols** > **Column Info** and select Formula from the list of column properties, as shown previously in **Figure 4.2**.

- Click **Row** in the Function browser, select **Row** from its list of functions, and click **OK** to close the Formula Editor window.
- Name the second column Total Rows, open the Formula Editor, select **NRow** from the list of **Row** functions, and click **OK**.
- Name the third column Lag.

Build the lag formula with these steps for the Lag column.

- Open the Formula Editor.
- Click Row ID in the list of columns.
- Select **Subscript** from the list of **Row** functions.
- Click the minus sign on the Formula Editor keypad, or key in a minus from the keyboard.
- Click **Row** from the **Rows** functions for the empty term on the left of the minus sign in the subscript and type "1" in the empty term on the right of the minus sign.

You should now see the lag formula: $Row\ ID_{Row()-1}$

- Click **OK** on the Formula Editor to see the data table results shown in **Figure 4.7**.

Do not be alarmed with the invalid row number error message. Simply click **Continue** to continue with the formula evaluation. The Lag function refers to the row previous to the current row; at row 1, this results in referring to the non-existent row 0, which produces a missing value in the data table.

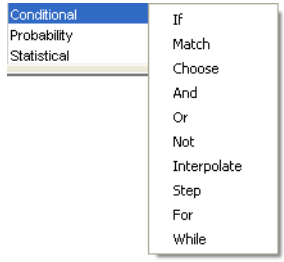
Note that the values in RowID and Lag are functions of individual rows, but the constant value in the TotRows column is a function of the data table.

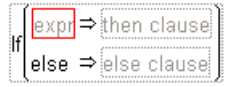
Figure 4.7 Formula Example

	Row ID	Total Rows	Lag
1	1	10	.
2	2	10	1
3	3	10	2
4	4	10	3
5	5	10	4
6	6	10	5
7	7	10	6
8	8	10	7
9	9	10	8
10	10	10	9

Conditional Expressions and Comparison Operators

This category has many familiar programming functions. This section shows examples of conditionals used with comparison operators.





The most basic and general conditional function is the If function. Its arguments are called if, then, and else *clause*s. When you highlight an expression and click **If**, the Formula Editor creates a new conditional expression like the one shown to the right. It has one If argument (a conditional expression denoted *expr*), one then argument, and a corresponding else and else clause. A conditional expression is usually a comparison, like $a < b$. However, any expression that evaluates as a numeric value can be used as a conditional expression. Expressions that evaluate as zero or missing are false. All other numeric expressions are true.

If you need more than one then statement, click the insert icon on the keypad (or type a comma, its keyboard shortcut) to add a new argument. To remove unwanted arguments, use the delete icon on the keypad or hit the Delete key on your keyboard.

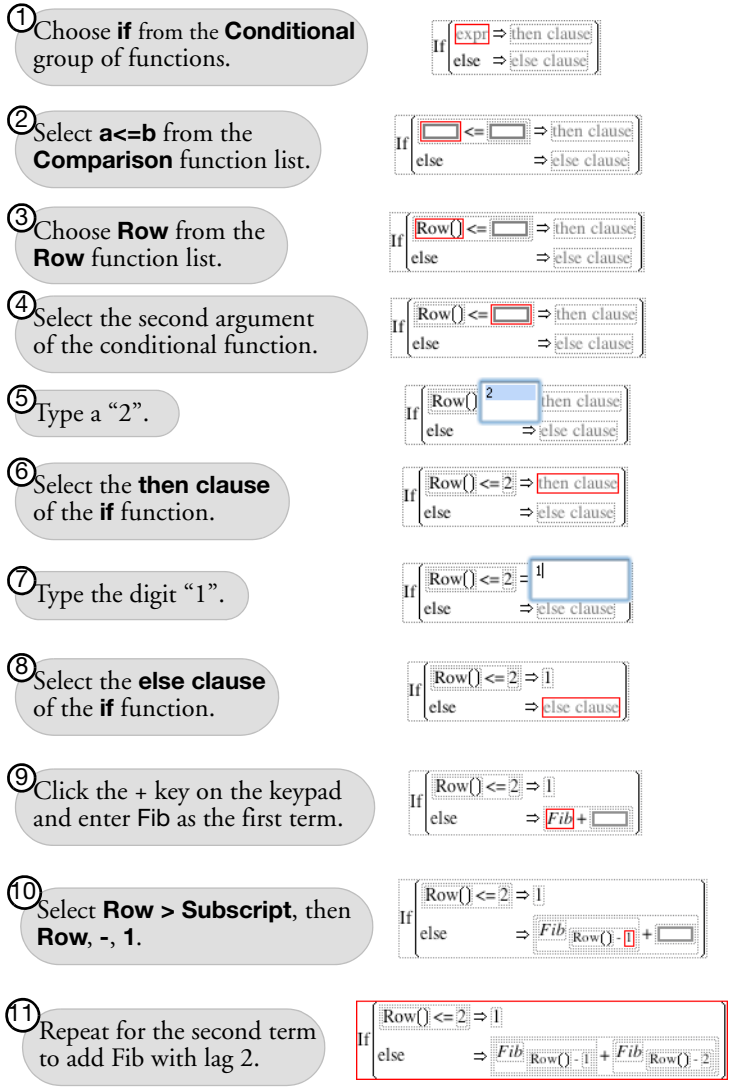
Using the If function

To create your own Fibonacci sequence:

- Name a blank data table column Fib.
- Right-click (Control-click on the Macintosh) and select **Formula** from the resulting menu.

Enter the Fibonacci formula using the steps shown in **Figure 4.8**.

Figure 4.8 Entering the Fibonacci Formula



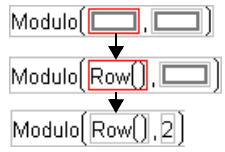
When you close the Formula Editor by clicking **OK**, the formula you entered generates the values shown in **Figure 4.9**.

Figure 4.9 Results of the Formula Example

	Fib	Group
1	1	1
2	1	0
3	2	1
4	3	0
5	5	1
6	8	0
7	13	1
8	21	0
9	34	1
10	55	0

The Fibonacci sequence has many interesting and easy-to-understand properties that are discussed in many number theory textbooks.

- For practice, create the values in the Group column shown in **Figure 4.9**. Use **Modulo** from the **Numeric** functions with RowID as its argument, as shown to the right.



Using the Match Function

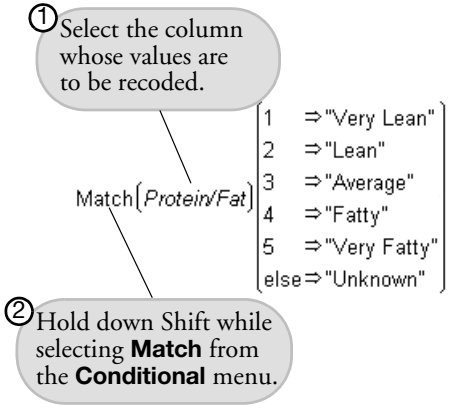
A common use for conditional functions is to re-code variables. Often, a numeric coding variable represents a descriptive character value. The following example uses the `Match` function for re-coding (the `If` function could also be used for re-coding).

Note: Value labels can also be used to add a descriptive label to a column of numeric variables. Details on value labels and their use are in the *JMP User Guide*.

When you select **Match** from the **Conditional** list, the Formula Editor shows a single `Match` condition with an empty expression and an empty `then` term. You add and delete clauses in a `Match` conditional the same way as in the `If` conditional described previously: select a `then` clause and click the add or delete button. The `Match` conditional compares an expression to a list of clauses and returns the value of the result expression for the first matching argument encountered. With `Match`, you provide the matching expression only once and then give a match for each argument.

As an example, open the Hot dogs.jmp data table. Suppose you want to create a value to recode the protein/fat ratio into categories “Very Lean,” “Lean,” “Average,” and so on (see the formula below).

- Create a new character column and add the following formula according to the following picture, which shows a shortcut for using the Match function.

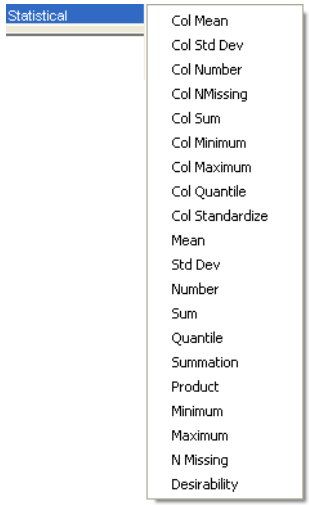


Note: `Match` evaluates faster and uses less memory than an equivalent `If`.

Summarize Down Columns or Across Rows

The Formula Editor evaluates statistical functions differently from other functions. Most functions evaluate data only for the current row. However, all **Statistical** functions require a set of values upon which to operate. Some **Statistical** functions compute statistics for the set of values in a column, and other functions compute statistics for the set of arguments you provide.

Figure 4.10 Statistical Functions



The functions with names prefaced by "Col" (`Col Mean`, `Col Sum`, and so on) always evaluate for all of the rows in a column. Thus, used alone as a column formula, these functions produce the same value for each row. These functions accept only a single argument, which can be a column name, or an expression involving any combination of column names, constants, or other functions.

The other statistical functions (`Mean`, `Std Dev`, and so on) accept multiple arguments that can be variables, constants, and expressions.

The `Sum` and `Product` functions evaluate over an explicitly specified range of values.

The Quantile Function

The `Col Quantile` function computes a quantile for a column of *n* nonmissing values. The `Col Quantile` function's quantile argument (call it *p*) represents the quantile percentage divided by 100.

The following examples are quantile formulas for a column named age:

`Col Quantile (age, 1)` finds the maximum age.

`Col Quantile (age, 0.75)` calculates the upper quartile age.

`Col Quantile (age, 0.5)` calculates the median age.

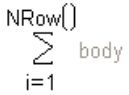
`Col Quantile (age, 0.25)` calculates the lower quartile age.

`Col Quantile (age, 0.0)` calculates the minimum age.

The pth quantile value is calculated using the formula $I = p(N + 1)$ where p is the quantile and N is the total number of nonmissing values. If I is an integer, then the quantile value is $y_p = y_i$. If I is not an integer, then the value is interpolated by assigning the integer part of the result to i, and the fractional part to f, and by applying the formula

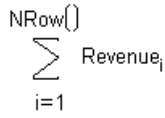
$$q_p = (1 - f)y_i + (f)y_{i+1}$$

Using the Summation Function



The **Summation** (Σ) function uses the summation notation shown to the left. To calculate a sum, select **Summation** from the **Statistical** function list and create its argument. The **Summation** function repeatedly evaluates the expression for the index you apply to the body of the function from the lower summation limit to the upper summation limit, and then adds the nonmissing results together to determine the final result. You can replace the index i, the index constant 1, and the upper limit, `NRow()`, with any expressions appropriate for your formula. Use the **Subscript** function in the **Row** function category to create a subscript for the body of the summation.

For example, the summation shown to the right computes the total of all revenue values for row 1 through the current row number, filling the calculated column with the cumulative totals of the revenue column.



Let's see how to compute a moving average using the summation function.

- Open the XYZ Stock Averages(plots).JMP sample table.
- Delete the existing moving avg. column.
- Create a new column called Moving Average and select **Formula** from the Column **Properties** list in the new column.
- Use the Column Info dialog to change the format from **Best** to **Fixed Dec** to give the numeric representation in the data table two decimal places.

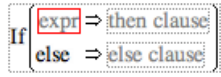
When you specify **Formula** as a new column property and click **Edit Formula**, the new column appears in the table with missing values, and the Formula Editor window opens. You should see a table like the one shown in **Figure 4.11**.

Figure 4.11 Example Table for Building a Moving Average

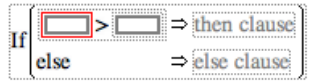
	Date	DJI High	DJI Close	DJI Low	XYZ	Moving Average
1	04/15/1991	2957.18	2933.17	2896.29	62.250	.
2	04/16/1991	2995.79	2986.88	2912.13	64.250	.
3	04/17/1991	3030.45	3004.46	2963.12	63.250	.
4	04/18/1991	3027.72	2999.26	2976.24	61.000	.
5	04/19/1991	3000.25	2965.59	2943.56	59.625	.

A *moving average* is the average of a fixed number of consecutive values in a column, updated for each row. The following example shows you how to compute a 10-day moving average for the XYZ stock. This means that for each row the Formula Editor computes the sum of the current XYZ value with the nine preceding values, then divides that sum by 10.

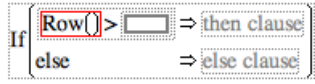
- Because you only want to compute the moving average starting with the 10th row, begin by selecting the conditional If function.



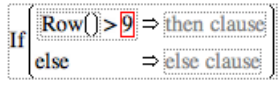
- With the If expression highlighted, select **a>b** from the **Comparison** function category, which will be used to determine the row number.



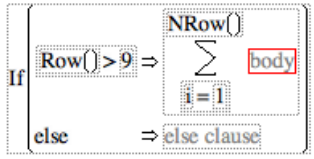
- For the left side of the comparison, select **Row** from the **Row** functions.



- Highlight the right side of the comparison and type in the number nine. The If expression should now appear as Row()>9.

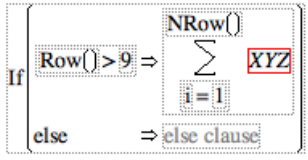


- Now highlight the then clause and begin the formula to compute the ten-day moving average by selecting the **Summation** function from the **Statistical** function category. Highlight the body of the summation and click XYZ in the column selector list.

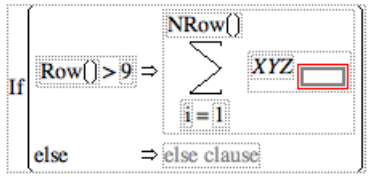


Now tailor the summation indices to sum just the 10 values you want:

- Highlight the summation body, XYZ.

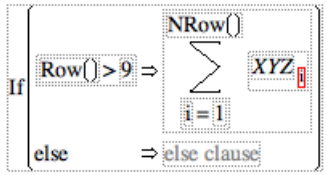


- Select **Subscript** from the **Row** function category.

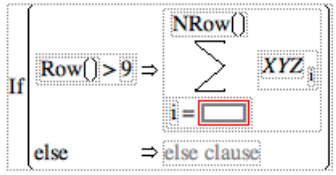


An empty subscript now appears with the summation body.

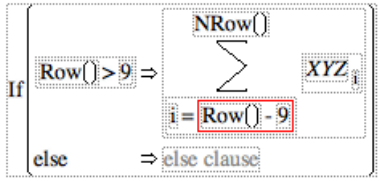
- To assign the subscript, either type the letter “i”, or drag the “i” from the lower limit of the summation into the empty subscript.



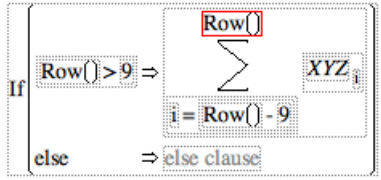
- Highlight the 1 in the lower summation limit and press the Delete key to change it to an empty term.



- Enter the expression Row()–9 inside the parentheses, using the **Row** selection in the **Row** function category.

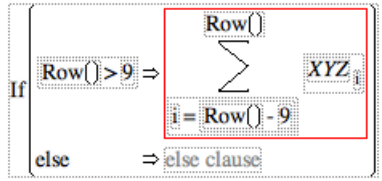


- Click the upper index to highlight it and select **Row** from the **Row** functions.

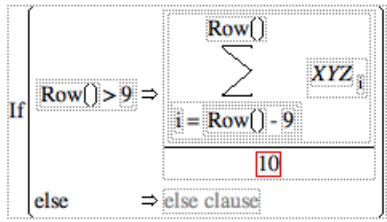


To finish the moving average formula, you want to divide the sum by 10, but not start the averaging process until you actually have 10 values to work with.

- Click in the summation to highlight the whole summation expression.



- Click the divide operator on the control panel, and then enter the constant 10 into the highlighted denominator that appears.



All that is left to do is use a conditional so that you don't compute anything for the first nine values in the table.

- When you click **Apply** or close the Formula Editor, the Moving Avg column fills with values.

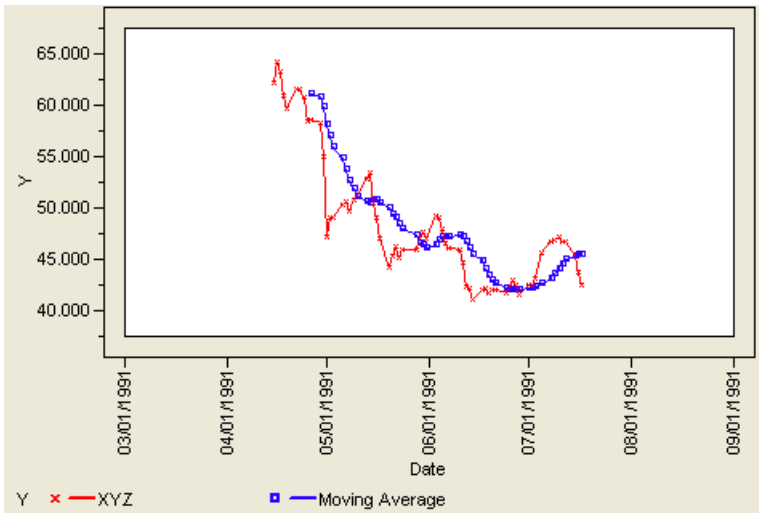
Now generate a plot to see the result of your efforts.

- Choose **Graph > Overlay Plot**, and select Date as **X** and both XYZ and Moving Avg as **Y**.

- When you click **OK**, select **Y Options > Connect Points** from the platform drop-down menu.

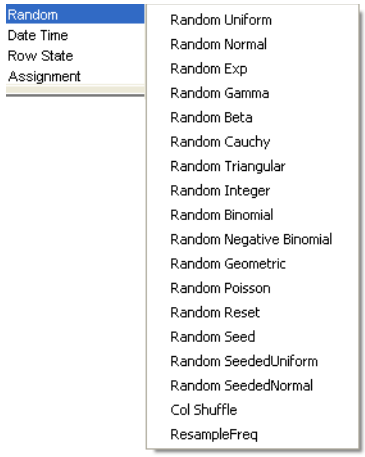
You then see the plot in **Figure 4.12**, which compares the XYZ stock price with its ten-day moving average.

Figure 4.12 Plot of Stock Prices and Their Moving Average



Random Number Functions

Random number functions generate real numbers by essentially "rolling the dice" within the constraints of the specified distribution. You can use the random number functions with a default 'seed' that provides a pseudo-random starting point for the random series, or use the `Random Reset` function and give a specific starting seed.



Random numbers in JMP are calculated using the Mersenne-Twister technique, far more efficient than in older versions of JMP. However, if you want to use the older functions, they are provided as `RandomSeededNormal` and `RandomSeededUniform`. Their seed is set using the `Random Seed` function.

Each time you click **Apply** in the Formula Editor window, random number functions produce a new set of numbers. This section shows examples of two commonly used random functions, `Uniform` and `Normal`.

The Uniform Distribution

The `Uniform` function generates random numbers uniformly distributed between 0 and 1. This means that any number between 0 and 1 is as likely to occur as any other. You can use the `Uniform` function to generate any set of numbers by modifying the function with the appropriate constants.

You can see simulated distributions using the `Uniform` function and the Distribution platform.

- Choose **File > New** to create a new data table.
- Right-click Column 1 and choose **Formula**.

- When the Formula Editor window opens, select **Random Uniform** from the **Random** function list in the function browser, and then close the Formula Editor.
- Choose **Cols > New Column** to create a second column.

Follow the same steps as before, except modify the **Uniform** function to generate the integers from 1 to 10 as follows.

- Click **Random** in the function browser and select **Random Uniform** from its list.
- Click the multiply sign on the Formula Editor key pad and enter 10 as the multiplier.
- Select the entire formula and click the addition sign on the Formula Editor keypad.
- Enter 1 in the empty argument term of the plus operator.

Note: JMP has a `Random Integer(n)` function that selects integers from a uniform distribution from 1 to `n`. It could be used here for the same effect. We're using the `Random Uniform` function to illustrate how to manipulate a random number by multiplying and adding constants. You can see an example of the `Random Integer` function in "Rolling Dice" on page 108.

The next steps are the key to generating a uniform distribution of integers (as opposed to real numbers as in Column 1):

- Click to select the entire formula.
- Select the **Floor** function from the **Numeric** function list.

The final formula is

```
Floor(Random Uniform()*10+1)
```

- Close the Formula Editor.

You now have a table template for creating two uniform distributions. Add as many columns as you want.

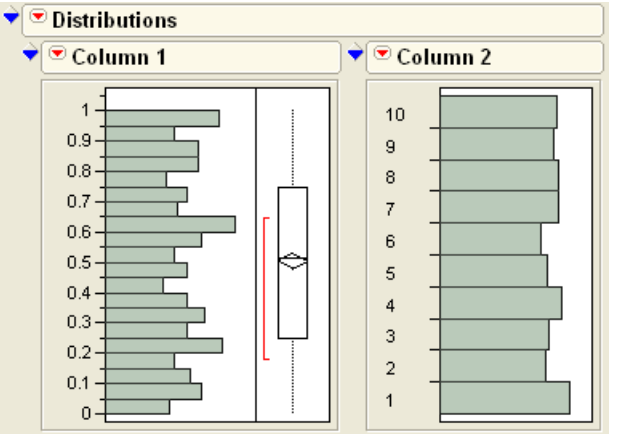
- Choose **Rows > Add Rows** and add 500 rows.

The table fills with values.

- Change the modeling type of the integer column to nominal so JMP treats it as a discrete distribution.
- Choose **Analyze > Distribution**, use both columns as Y variables, and then click **OK**.

You see two histograms similar to those shown in **Figure 4.13**. The histogram on the left represents simple uniform random numbers and the histogram on the right shows random integers from 1 to 10.

Figure 4.13 Example of Two Uniform Distribution Simulations



The Normal Distribution

Random Normal generates random numbers that approximate a Normal distribution with a mean of 0 and variance of 1. The Normal distribution is bell-shaped and symmetrical. You can modify the **Random Normal** function with arguments that specify a Normal distribution with a different mean and standard deviation.

As an exercise, follow the same instructions described previously for the Uniform random number function.

- Create a table with a column for a standard Normal distribution using the `Random Normal()` function.

The Random Normal function takes two optional arguments. The first specifies the mean of the distribution; the second specifies the standard deviation of the distribution.

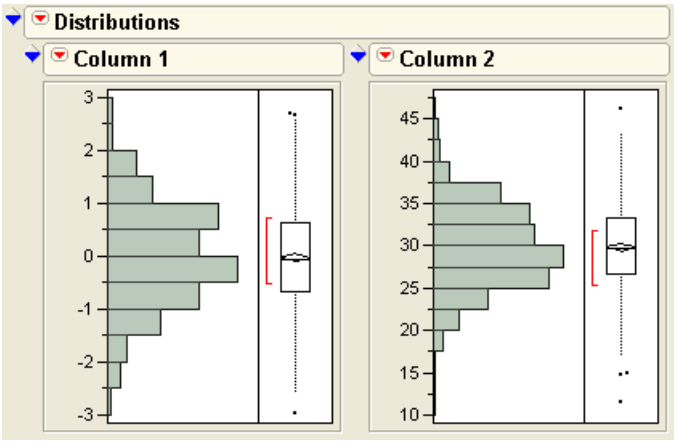
- Create a second column for a random Normal distribution with mean 30 and standard deviation 5.

The modified Normal formula is

```
Random Normal(30, 5)
```

Figure 4.14 shows the Distribution platform results for these Normal simulations.

Figure 4.14 Illustration of Normal Distributions



The Col Shuffle Command

Col Shuffle selects a row number at random from the current data table. Each row number is selected only once. When **Col Shuffle** is used as a subscript, it returns a value selected at random from the column that serves as its argument.

For example, to identify a 50% random sample without replacement, use the following formula:

$$\text{If}\left(\text{Row}() < \frac{\text{NRow}()}{2} \Rightarrow \textit{Column 1}_{\text{Col Shuffle}()}\right)$$

This formula chooses half the values ($n/2$) from Column 1 and assigns them to the first half of the rows in the computed column. The remaining rows of the computed column remain missing.

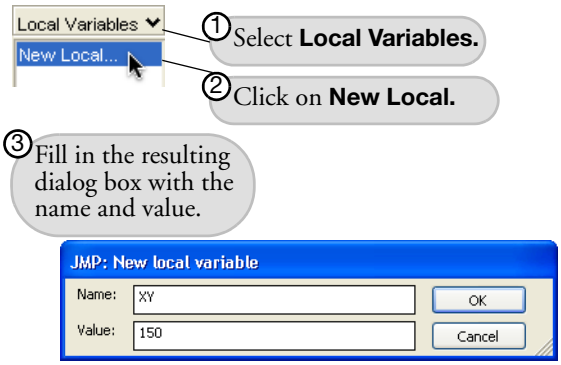
Local Variables and Table Variables

Local variables let you define temporary numeric variables to use in expressions. Local variables exist only for the column in which they are defined.

To create a new local variable, use the button on the formula editor keypad. This button adds a temporary local variable to the formula editing area, which appears as a command ending in a semicolon. Alternatively, you can select **Local Variables** from the Formula Elements popup menu, select **New Local**, and complete the dialog that appears.

By default, local variables have the names t1, t2, and so on, and initially have missing values. Local variables appear in a formula as bold italic terms.

Optionally, you can create a local variable, change its name and assign a starting value in the Local Variable dialog. To use the Local Variable dialog, select **Local Variables** from the Formula Elements popup menu, then select **New Local** and complete the dialog, as illustrated here.



As an example, suppose you have variables *x* and *y* and you want to compute the slope in a simple linear regression of *y* on *x* using the standard formula shown here.

$$\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

One way to do this is to create two local variables, called XY and Xsqrd, as described in the numerator and denominator in the equation above. Then assign them to the numerator and the denominator calculations of the slope formula. The slope computation is simplified to XY divided by Xsqrd.

$$XY = \sum_{i=1}^{NRow()} \left[\left(X_i\text{-Col Mean}(X)\right)*\left(Y_i\text{-Col Mean}(Y)\right)\right];$$

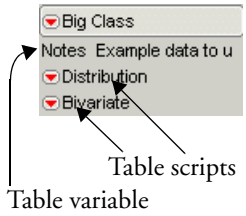
$$Xsqrd = \sum_{i=1}^{NRow()} \left[X_i\text{-Col Mean}(X)\right]^2;$$

$$\frac{XY}{Xsqrd}$$

The **Local Variables** command in the Formula Editor popup menu lists all the local variables that have been created.

Table variables are available to the entire table. Table variable names are displayed in the Tables panel at the left of the data grid. The Formula Editor can refer to a table variable in a formula.

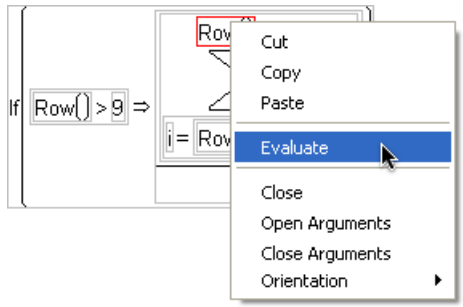
Many of the sample data files have a table variable called Notes. The **Table Variables** command in the Formula Elements popup menu lists all the Table variables that exist for a table. You can create additional Table variables with the **New Table Variable** command in the Tables panel of the data table, or edit the values of existing table variables.



Tips on Building Formulas

Examining Expression Values

Once JMP has evaluated a formula, you can select an expression to see its value. This is true for both parameters and expressions that evaluate to a constant value. To do this, select the expression you want to know about and right-click (PC) or control-click (Mac) on it. This displays a popup menu as shown here. When you select **Evaluate**, the current value of the selected expression shows until you move the cursor.



Cutting, Dragging, and Pasting Formulas

You can cut or copy a formula or an expression, and paste it into another formula display. Or you can drag any selected part of a formula to another location within the same formula. When you place the arrow cursor inside an expression and click, the expression is highlighted. When the cursor is over a selected area, it changes to a hand cursor, indicating that you can drag the highlighted formula under the cursor. As you drag across the formula, destination expressions are highlighted. When you release the drag, the selected expression is copied to the new location where it replaces the existing expression.

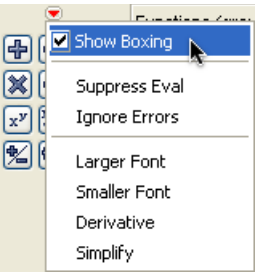
When you copy (or drag) an expression from one data table to another, JMP expects to find matching column names. If a formula column name does not appear in the destination table, an error alerts you when the formula attempts to evaluate.

Selecting Expressions

You can click on any single term in an expression to select it for editing. You can use the keyboard arrow keys to select expressions for editing or to view the grouping of terms within a formula when parentheses are not present or the boxing option is not in effect.

Once an operand is selected, the left and right arrow keys move the selection across other associative operands within the expression. The left arrow highlights the next formula element to the left of the currently highlighted term, or extends the selection to include an additional term that is part of a group.

Tip: Turn the boxing option on to see how the elements and terms are grouped in a formula you create. It is often easier to leave the boxing option on while creating a formula.



The up arrow extends the current selection by adding the next operand and operator of the formula term to the selection. The down arrow reduces the current selection by removing an operand and operator from the selection.

Tips on Editing a Formula

If you need to change a formula, highlight the column and select **Formula** from the **Cols** menu. Alternatively, right-click at the top of the column or on the column name in the Columns panel and select **Formula** from the context menu that appears.

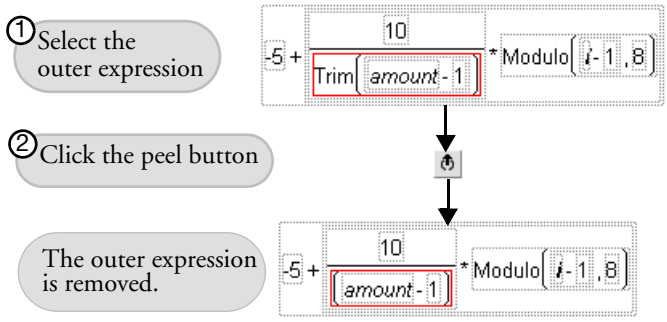
Deleting a function also deletes its arguments. Deleting a required argument or empty term from a function sometimes deletes the function as well. You can save complicated expressions that serve as arguments and paste them where needed by using the **Copy** command to copy the arguments to the clipboard.

Another useful editing technique is to peel a function from its argument as shown in **Figure 4.15**. To peel a function from a single argument, first click to select the function. Then, choose the **peel** button from the keypad, as shown in **Figure 4.15**.

After you complete formula changes, the new values fill the column automatically when you click **Apply** or close the Formula Editor window.

Once you have created a formula, you can change values in columns that are referenced by your formula. JMP automatically recalculates all affected values in the formula's column.

Figure 4.15 Peeling Expressions or Arguments



Exercises

1. The file Pendulum.jmp contains the results of an experiment in a physics class comparing the length of a pendulum to its period (the time it takes the pendulum to make a complete swing). Calculations were made for a range of pendulums from short (2 cm) to long (20 m). We will use the calculator to determine a model to predict the period of a pendulum from its length.

   (a) Produce a scatterplot of the data by selecting **Analyze > Fit Y By X** and selecting Period as the Y variable and Length as the X variable.

   (b) Create a new column named Transformed Period that contains a formula to take the square root of the Period column. Produce a scatterplot of Transformed Period vs. Length. Is this the graph of a linear function?

   (c) Try other transformations (for example, natural log of Period, reciprocal of Period, or square of Period) until the scatterplot looks linear.

   (d) Find the line of best fit for the linear transformed data by selecting **Fit Line** from the popup menu beside the title of the scatterplot.

   (e) This line is not the fit of the original data, but of the transformed data. Substitute a term representing the transformation you did to linearize the data (for example, if the square root transformation made the data linear, substitute $\sqrt{\text{Period}}$ into the regression equation) and solve the equation for Period.

   A Physics textbook reveals the relationship between period and length to be

   $$\text{Period} = \frac{2\pi}{\sqrt{g}}\sqrt{\text{Length}} \quad \text{where } g = 9.8\frac{\text{m}}{\text{s}^2}.$$

(f) Create a new column to use this formula to calculate the theoretical values. Next, construct another column to calculate the difference between the observed values of the students and the theoretical values.

(g) Examine a histogram of these differences. Does it appear that there was a trend in the observations of the students?

2. Is there a correlation among the mean, minimum, maximum, and standard deviation of a set of data? To investigate, create a new data table in JMP with these characteristics:

   (a) Create ten columns of data named Data 1 through Data 10, each with the formula Random Uniform(). Add 500 rows to this data table.

   (b) Create four columns to hold the four summary statistics of interest, one column each for the mean, minimum, maximum, and standard deviation. Create a formula in each column to calculate the appropriate statistic of the ten data rows.

   (c) Select the Multivariate platform in the **Analyze** menu and include the four summary statistics as the Y's in the resulting dialog box. Pressing **OK** should produce 16 scatterplots. Which statistics seem to show a correlation?

   (d) As an extension, select two of the statistics that seem to show a correlation. Produce a single scatterplot of these two statistics using the Fit Y By X platform in the **Analyze** menu. From the red arrow drop-down menu beside the title of the plot, select **Nonpar Density**. Then choose **Save Density Grid** from the popup menu beside the density legend.

   (e) Finally, select **Graph** > **Spinning Plot** and include the first three columns of the saved density grid. You should now see a 3D scatterplot of the correlation, with a peak where the data points overlapped each other. Use the hand tool to move the spinning plot around.

3. Make a data table consisting of 20 rows and a single column that holds the Fibonacci sequence (whose formula is shown on page 73). Label this column Fib.

   (a) Add a new column called Ratio, and give it the following formula to take the ratio of adjacent rows. **Note:** This will produce an error alert when evaluated for the first row. Click **OK** on the alert dialog to resolve this issue.

$$\frac{Fib}{Fib_{\text{Row}()-1}}$$

What value does this ratio converge to?

(b) A more generalized Fibonacci sequence uses values aside from 1 as the first two elements. A Lucas sequence uses the same recursive rule as the Fibonacci, but uses different starting values. Create a column to hold a Lucas sequence beginning with 2 and 5 called Lucas with the formula

$$\text{If}\begin{cases}\text{Row}()==1 \Rightarrow 2\\ \text{Row}()==2 \Rightarrow 5\\ \text{else} \quad \Rightarrow Lucas_{\text{Row}()-1}+Lucas_{\text{Row}()-2}\end{cases}$$

(c) Create a column to calculate the ratio of two successive terms of the Lucas sequence in part (b). Is it the same as the number in part (a)?

(d) Create a Lucas sequence starting with the values 1, 2. Calculate the ratio of successive terms and compare it to the answer in part (c).

(e) There are innumerable other properties of the Fibonacci sequence. For example, add a column that contains the following formula, and comment on the result.

$$\sum_{i=1}^{\text{Row}()} \text{ArcTangent}\left(\frac{1}{Fib_{2*i+1}}\right)*4$$

5 What Are Statistics?

Overview

Statistics are numbers, but the practice of statistics is the craft of measuring imperfect knowledge. That's one definition, and there are many more.

This chapter is a collection of short essays to get you started on the many ways of statistical thinking and to get you used to the terminology of the field.

Ponderings

The Business of Statistics

The discipline of statistics provides the framework of balance sheets and income statements for scientific knowledge. Statistics is an accounting discipline, but instead of accounting for money, it is accounting for scientific credibility. It is the craft of weighing and balancing observational evidence. Scientific conclusions are based on experimental data in the presence of uncertainty, and statistics is the mechanism to judge the merit of those conclusions. The statistical tests are like credibility audits. Of course, you can juggle the books and sometimes make poor science look better than it is. However, there are important phenomena that you just can't uncover without statistics.

A special joy in statistics is when it is used as a discovery tool to find out new phenomena. There are many views of your data—the more perspectives you have on your data, the more likely you are to find out something new. Statistics as a discovery tool is the auditing process that unveils phenomena that are not anticipated by a scientific model and are unseen with a straightforward analysis. These anomalies lead to better scientific models.

Statistics fits models, weighs evidence, helps identify patterns in data, and then helps find data points that don't fit the patterns. Statistics is induction from experience; it is there to keep score on the evidence that supports scientific models.

Statistics is the science of uncertainty, credibility accounting, measurement science, truth-craft, the stain you apply to your data to reveal the hidden structure, the sleuthing tool of a scientific detective.

Statistics is a necessary bureaucracy of science.

The Yin and Yang of Statistics

There are two sides to statistics.

First, there is the Yang of statistics, a shining sun. The Yang is always illuminating, condensing, evaporating uncertainty, and leaving behind the precipitate of knowledge. It pushes phenomena into forms. The Yang is out to prove things in the presence of uncertainty and ultimately compel the data to confess its form, conquering ignorance headlong. The Yang demolishes hypotheses by ridiculing their improbability. The Yang mechanically cranks through the ore of knowledge and distills it to the answer.

On the other side, we find the contrapositive Yin, the moon, reflecting the light. The Yin catches the shadow of the world, feeling the shape of truth under the umbra. The Yin is

forever looking and listening for clues, nurturing seeds of pattern and anomaly into maturing discoveries. The Yin whispers its secrets to our left hemisphere. It unlocks doors for us in the middle of the night, planting dream seeds, making connections. The Yin draws out the puzzle pieces to tantalize our curiosity. It teases our awareness and tickles our sense of mystery until the climax of revelation—Eureka!

The Yin and Yang are forever interacting, catalyzed by Random, the agent of uncertainty. As we see the world reflected in the pool of experience, the waters are stirred, and in the agitated surface we can't see exactly how things are. Emerging from this, we find that the world is knowable only by degree, that we have knowledge in measure, not in absolute.

The Faces of Statistics

Everyone has a different view of statistics.

Match the definition on this side....	**with someone likely to have said it on this side**
1. The literature of numerical facts.	a. Engineer
2. An applied branch of mathematics.	b. Original meaning
3. The science of evidence in the face of uncertainty.	c. Social scientist
4. A digestive process that condenses a mass of raw data into a few high-nutrient pellets of knowledge.	d. Philosopher
5. A cooking discipline with data as the ingredients and methods as the recipes.	e. Economist
6. The calculus of empiricism.	f. Computer scientist
7. The lubricant for models of the world.	g. Mathematician
8. A calibration aid.	h. Physicist
9. Adjustment for imperfect measurement.	i. Baseball fan
10. An application of information theory.	j. Lawyer
11. Involves a measurable space, a sigma algebra, and Lebesgue integration.	k. Joe College
12. The nation's state.	l. Politician
13. The proof of the pudding.	m. Businessman
14. The craft of separating signal from noise.	n. Statistician
15. A way to predict the future.	

An interesting way to think of statistics is as a toy for grown-ups. Remember that toys are proxies that children use to model the world. Children use toys to learn behaviors and develop explanations and strategies, as aids for internalizing the external. This is the case with statistical models. You model the world with a mathematical equation, and then see how the model stacks up to the observed world.

Statistics lives in the interface of the real world data and mathematical models, between induction and deduction, empiricism and idealism, thought and experience. It seeks to balance real data and a mathematical model. The model addresses the data and stretches to fit. The model changes and the change of fit is measured. When the model doesn't fit, the data suspends from the model and leaves clues. You see patterns in the data that don't fit, which leads to a better model, and points that don't fit into patterns can lead to important discoveries.

Don't Panic

Some university students have a panic reaction to the subject of statistics. Yet most science, engineering, business, and social science majors usually have to take at least one statistics course. What are some of the sources of our phobias about statistics?

Abstract Mathematics

Though statistics can become quite mathematical to those so inclined, applied statistics can be used effectively with only basic mathematics. You can talk about statistical properties and procedures without plunging into abstract mathematical depths. In this book, we are interested in looking at applied statistics.

Lingo

Statisticians often don't bother to translate terms like 'heteroschedasticity' into 'varying variances' or 'multicollinearity' into 'closely related variables.' Or, for that matter, further translate 'varying variances' into 'difference in the spread of values between samples,' and 'loosely related variables' into 'variables that give similar information.' We tame some of the common statistical terms in the discussions that follow.

Awkward Phrasing

There is a lot of subtlety in statistical statements that can sound awkward, but the phrasing is very precise and means exactly what it says. Sometimes statistical statements include multiple negatives. For example, "The statistical test failed to reject the null hypothesis of no effect at the specified alpha level." That is a quadruple negative statement. Count the negatives: 'fail,' 'reject,' 'null,' and 'no effect.' You can reduce the number of negatives by saying "the statistical results are not significant" as long as you are careful not to confuse that with the statement "there is no effect." Failing to prove something does not prove the opposite!

A Bad Reputation

The tendency to assume the proof of an effect because you cannot statistically prove the absence of the effect is the origin of the saying, "Statistics can prove anything." This is what happens when you twist a term like 'nonsignificant' into 'no effect.' This idea is common in a courtroom; you can't twist the phrase "there is not enough evidence to prove beyond reasonable doubt that the accused committed the crime" with "the accused is innocent." What nonsignificant really means is that there is not enough data to show a significant effect—it does not mean that there is no effect at all.

Uncertainty

Although we are comfortable with uncertainty in ordinary daily life, we are not used to embracing it in our knowledge of the world. We think of knowledge in terms of hard facts and solid logic, though much of our most valuable real knowledge is far from solid. We can say when we know something for sure (yesterday it rained), and we can say when we don't know (don't know whether it will rain tomorrow). But when we describe knowing something with incomplete certainty, it sounds apologetic or uncommitted. For example, it sounds like a form of equivocation to say that there is a 90% chance that it will rain tomorrow. Yet much of what we think we know contains just that kind of uncertainty.

Preparations

A few fundamental concepts will prepare you for absorbing details in upcoming chapters.

Three Levels of Uncertainty

Statistics is about uncertainty, but there are several different levels of uncertainty that you have to keep in separate accounts.

Random Events

Even if you know everything possible about the current world, unpredictable events still exist. You can see an obvious example of this in any gambling casino. You can be an expert at playing blackjack, but the randomness of the card deck renders the outcome of any game indeterminate. We make models with random error terms to account for uncertainty due to randomness. Some of the error term may be due to ignoring details; some may be measurement error; but much of it is attributed to inherent randomness.

Unknown Parameters

Not only are you uncertain how an event is going to turn out, you often don't even know what the numbers (parameters) are in the model that generates the events. You have to estimate the parameters and test if hypothesized values of them are plausible, given the data. This is the chief responsibility of the field of statistics.

Unknown Models

Sometimes you not only don't know how an event is going to turn out, and you don't know what the numbers are in the model, but you don't even know if the form of the model is right.

Statistics is very limited in its help for certifying that a model is correct. Most statistical conclusions assume that the hypothesized model is correct. The correctness of the model is the responsibility of the subject-matter science. Statistics might give you clues if the model is not carrying the data very well. Statistical analyses can give diagnostic plots to help you see patterns that could lead to new insights, to better models.

Probability and Randomness

In the old days, statistics texts all began with chapters on probability. Today, many popular statistics books discuss probability in later chapters. We mostly omit the topic in this book, though probability is the essence of our subject.

Randomness makes the world interesting and probability is needed as the measuring stick. Probability is the aspect of uncertainty that allows the information content of events to be measured. If the world were deterministic, then the information value of an event would be zero because it would already be known to occur; the probability of the event occurring would be 1. The sun rising tomorrow is a nearly deterministic event and doesn't make the front page of the newspaper when it happens. The event that happens but has been attributed to having probability near zero would be big news. For example, the event of extraterrestrial intelligent life forms landing on earth would make the headlines.

Statistical language uses the term probability on several levels:

- When we make observations or collect measurements, our responses are said to have a *probability distribution*. For whatever reason, we assume that something in the world adds randomness to our observed responses, which makes for all the fun in analyzing data that has uncertainty in it.
- We calculate statistics using probability distributions, seeking the safe position of maximum likelihood, which is the position of least improbability.
- The significance of an event is reported in terms of probability. We demolish statistical null hypotheses by making their consequences look incredibly improbable.

Assumptions

Statisticians are naturally conservative professionals. Like the boilerplate of official financial audits, statisticians' opinions are full of provisos such as "assuming that the model is correct, and assuming that the error is Normally distributed, and assuming that the observations are

independent and identically distributed, and assuming that there is no measurement error, and assuming...." Even then the conclusions are hypothetical, with phrases like "if you say the hypothesis is false, then the probability of being wrong is less than 0.05."

Statisticians are just being precise, though they sound like they are combining the skills of equivocation like a politician, techno-babble like a technocrat, and trick-prediction like the Oracle at Delphi.

Ceteris Paribus

A crucial assumption is the *ceteris paribus* clause, which is Latin for other things being equal. This means we assume that the response we observed was really only affected by the model's factors and random error; all other factors that might affect the response were maintained at the same controlled value across all observations or experimental units. This is, of course, often not the case, especially in observational studies, and the researcher must try to make whatever adjustments, appeals, or apologies to atone for this. When statistical evidence is admitted in court cases, there are endless ways to challenge it, based on the assumptions that may have been violated.

Is the Model Correct?

The most important assumption is that your model is right. There are no easy tests for this assumption. Statistics almost always measure one model against a submodel, and these have no validity if neither model is appropriate in the first place.

Is the Sample Valid?

The other supremely important issue is that the data relate to your model; that is, that you have collected your data in a way that is fair to the questions that you ask it. If your sample is ill-chosen, or if you have skewed your data by rejecting data in a process that relates to its applicability to the questions, then your judgments will be flawed. If you have not taken careful consideration of the direction of causation, you may be in trouble. If taking a response affects the value of another response, then the responses are not independent of each other, which can affect the study conclusions.

In brief, are your samples fairly taken and are your experimental units independent?

Data Mining?

One issue that most researchers are guilty of to a certain extent is stringing together a whole series of conclusions and assuming that the joint conclusion has the same confidence as the individual ones. An example of this is data mining, in which hundreds of models are tried until one is found with the hoped-for results. Just think about the fact that if you collect purely random data, you will find a given test significant at the 0.05 level about 5% of the

time. So you could just repeat the experiment until you get what you want, discarding the rest. That's obviously bad science, but something similar often happens in published studies. This multiple-testing problem remains largely unaddressed by statistical literature and software except for some special cases such as means comparisons, a few general methods that may be inefficient (Bonferroni's adjustment), and expensive, brute-force approaches (resampling methods).

Another problem with this issue is that the same kind of bias is present across unrelated researchers because nonsignificant results are often not published. Suppose that 20 unrelated researchers do the same experiment, and by random chance one researcher got a 0.05-level significant result. That's the result that gets published.

In light of all the assumptions and pitfalls, it is appropriate that statisticians are cautious in the way they phrase results. Our trust in our results has limits.

Statistical Terms

Statisticians are often unaware that they use certain words in a completely different way than other professionals. In the following list, some definitions are the same as you are used to, and some are the opposite:

Model

A statistical model is a mathematical equation that predicts the response variable as a function of other variables, together with some distributional statements about the random terms that allow it to not fit exactly. Sometimes this model is taken casually in order to look at trends and tease out phenomena, and sometimes the model is taken seriously.

Parameters

To a statistician, parameters are the unknown coefficients in a model, to be estimated and to test hypotheses about. They are the indices to distributions; the mean and standard deviation are the location and scale parameters in the Normal distribution family.

Unfortunately, engineers use the same word (parameters) to describe the factors themselves.

Statisticians usually name parameters after Greek letters, like mu(μ), sigma(σ), beta(β), and theta(θ). You can tell where statisticians went to school by which Greek and Roman letters they use in various situations. For example, in multivariate models, the L-Beta-M fraternity is distinguished from C-Eta-M.

Hypotheses

In science, the hypothesis is the bright idea that you want to confirm. In statistics, this is turned upside down because it uses logic analogous to a proof-by-contradiction. The so-called

null hypothesis is usually the statement that you want to demolish. The usual null hypothesis is that some factor has no effect on the response. You are of course trying to support the opposite, which is called the *alternative hypothesis*. You support the alternative hypothesis by statistically rejecting the null hypothesis.

Two-Sided versus One-Sided, Two-Tailed versus One-Tailed

Most often, the null hypothesis can be stated as some parameter in a model being zero. The alternative is that it is not zero, which is called a two-sided alternative. In some cases, you may be willing to state the hypothesis with a one-sided alternative, for example that the parameter is greater than zero. The one-sided test has greater power at the cost of less generality. These terms have only this narrow technical meaning, and it has nothing to do with common English phrases like presenting a one-sided argument (prejudiced, biased in the everyday sense) or being two-faced (hypocrisy or equivocation). You can also substitute the word "tailed" for "sided." The idea is to get a big statistic that is way out in the tails of the distribution where it is highly improbable. You measure how improbable by calculating the area of one of the tails, or the other, or both.

Statistical Significance

Statistical significance is a precise statistical term that has no relation to whether an effect is of practical significance in the real world. Statistical significance usually means that the data gives you the evidence to believe that some parameter is not the value specified in the null hypothesis. If you have a ton of data, you can get a statistically significant test when the values of the estimates are practically zero. If you have very little data, you may get an estimate of an effect that would indicate enormous practical significance, but it is supported by so little data that it is not statistically significant. A nonsignificant result is one that might be the result of random variation rather than a real effect.

Significance Level, *p*-value, α-level

To reject a null hypothesis, you want small *p*-values. The *p-value* is the probability of being wrong if you declare an effect to be non-null; that is, the probability of rejecting a 'true' null hypothesis. The *p*-value is sometimes labeled the *significance probability*, or sometimes labeled more precisely in terms of the distribution that is doing the measuring. The *p*-value labeled "Prob>|t|" is read as "the probability of getting a greater t (in absolute value)." The α-level is your standard of the *p*-value that you claim, so that *p*-values below this reject the null hypothesis (that is, they show that there is an effect).

Power, β-level

Power is how likely you are to detect an effect if it is there. The more data you have, the more statistical power. The greater the real effect, the more power. The less random variation in your world, the more power. The more sensitive your statistical method, the more power. If

you had a method that always declared an effect significant, regardless of the data, it would have a perfect power of 1 (but it would have an α-level of 1, too, the probability of declaring significance when there was no effect). The goal in experimental design is usually to get the most power you can afford, given a certain α-level. It is not a mistake to connect the statistical term power with the common sense of power as persuasive ability. It has nothing to do with work or energy, though.

Confidence Intervals

A confidence interval is an interval around a parameter estimate that has a given probability of containing the true value of a parameter. Most often the probability is 95%. Confidence intervals are now considered one of the best ways to report results. It is expressed as a percentage of $1 - \alpha$, so an 0.05 alpha level for a two-tailed *t*-quantile can be used for a 95% confidence interval. (For linear estimates, it is constructed by multiplying the standard error by a *t*-statistic and adding and subtracting that to the estimate. If the model involves nonlinearities, then the linear estimates are just approximations, and there are better confidence intervals called *profile likelihood confidence intervals*. If you want to form confidence regions involving several parameters, it is not valid to just combine confidence limits on individual parameters.) You can learn more about confidence intervals in Chapter 6.

Biased, Unbiased

An *unbiased estimator* is one where the expected value of an estimator is the parameter being estimated. It is considered a desirable trait, but not an overwhelming one. There are cases when statisticians recommend biased estimators. For example, the maximum likelihood estimator of the variance has a small (but nonzero) bias.

Sample Mean versus True Mean

You calculate a sample man from your data—the sum divided by the number. It is a statistic, that is, a function of your data. The *true mean* is the expected value of the probability distribution that generated your data. You usually don't know the true mean, which is why you collect data, so you can estimate the true mean with the sample mean.

Variance and Standard Deviation, Standard Error

Variance is the expected squared deviation of a random variable from its expected value. It is estimated by the sample variance. *Standard deviation* is the square root of the variance, and we prefer to report it because it is in the same units as the original random variable (or response values). The sample standard deviation is the square root of the sample variance. The term standard error describes an estimate of the standard deviation of another (unbiased) estimate.

Degrees of Freedom

Degrees of freedom (df) are the specific name for a value that indexes some popular distributions of test statistics. It is called degrees of freedom because it relates to differences in

numbers of parameters that are or could be in the model. The more parameters a model has, the more freedom it has to fit the data better. The DF (degrees of freedom) for a test statistic is usually the difference in the number of parameters between two models.

6 Simulations

Overview

A good way to learn how statistics measure and model a process is to first build an imaginary process, then see how well the statistics see it. Simulation is the word for building an imaginary process; Monte Carlo simulations are simulations done with a random number generator.

Simulations do not have to be complex programs or scripts. As you will see, they can be simple data tables that accrue information repeatedly.

Rolling Dice

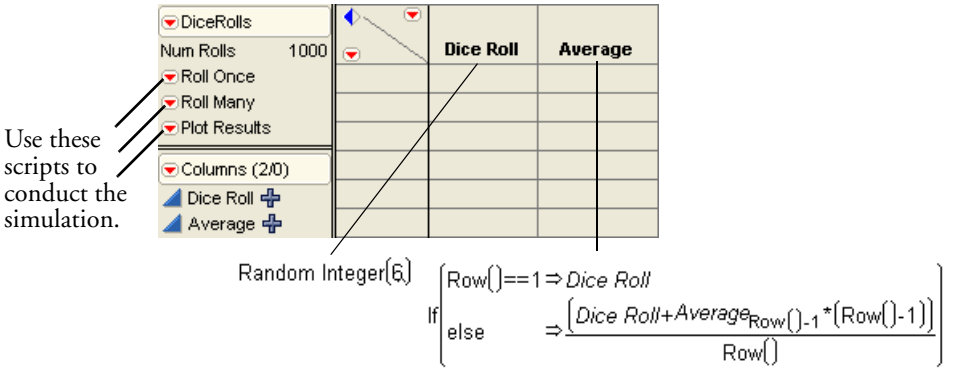
A simple example of a Monte Carlo simulation from elementary probability is rolling a six-sided die and recording the results over a long period of time. Of course, it is impractical to physically roll a die repeatedly, so JMP is used to simulate the rolling of the die.

The assumption that each face has an equal probability of appearing means that we should simulate the rolls using a function that draws from a uniform distribution. We could use the `Random Uniform()` function, which pulls random real numbers from the (0,1) interval. However, JMP has a special version of this function for cases where we want random integers (in this case, we want random integers from 1 to 6).

- Open the Dice Rolls.jmp data table.

The table consists of a column named Dice Roll to hold the random integers. Each row of the data table represents a single roll of the die. A second column is used to keep a running average of all the rolls up to that point.

Figure 6.1 Dice Rolls.jmp Data Table



The law of large numbers states that as we increase the number of observations, the average should approach the true theoretical average of the process. In this case, we expect the average to approach $\frac{1+2+3+4+5+6}{6}$, or 3.5.

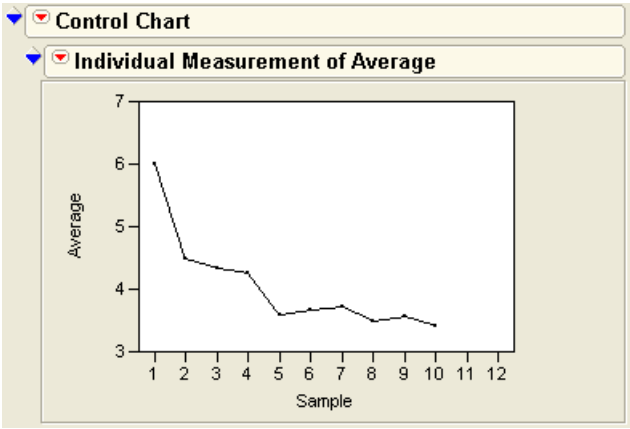
- Click on the red triangle beside the **Roll Once** script in the side panel of the data table and select **Run Script**.

This adds a single roll to the data table. Note that this is equivalent to adding rows through the **Rows > Add Rows** command. It is included as a script simply to reduce the number of mouse clicks needed to perform the function.

- Repeat this process several times to add ten rows to the data table.
- After ten rows have been added, run the **Plot Results** script in the side panel of the data table.

This produces a control chart of the results. Note that the results fluctuate fairly widely at this point.

Figure 6.2 Plot of Results After Ten Rolls



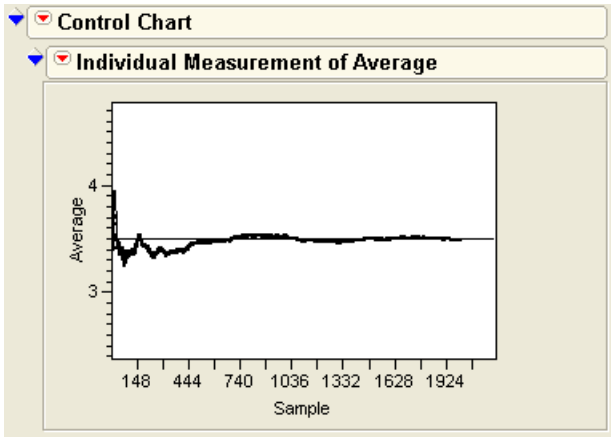
- Run the **Roll Many** script in the side panel of the data table.

This adds many rolls at once. In fact, it adds the number of rows specified in the table variable **Num Rolls** each time it is clicked. To add more or fewer rolls at one time, adjust the value of the **Num Rolls** variable.

Also note that the control chart has automatically updated itself. The chart reflects the new observations just added.

- Continue adding points until there are 2000 points in the data table.

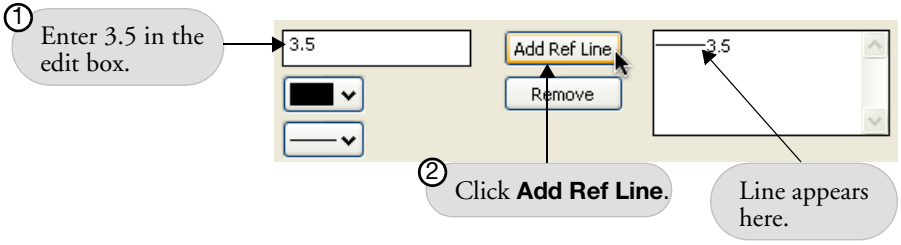
Figure 6.3 Observed Mean Approaches Theoretical Mean



The control chart shows that the mean is leveling off, just as the law of large numbers predicts, at the value 3.5. In fact, you can add a horizontal line to the plot to emphasize this point.

- Double-click the *y*-axis to open the axis specification dialog.
- Enter values into the dialog box as shown in **Figure 6.4**.

Figure 6.4 Adding a Reference Line to a Plot



Although this is not a complicated example, it shows how easy it is to produce a simulation based on random events. In addition, this data table could be used as a basis for other simulations. like the following.

Rolling Several Dice

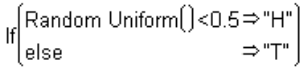
If you need to roll more than one die at a time, simply copy and paste the formula from the existing Die Roll column into other columns.

Flipping Coins, Sampling Candy, or Drawing Marbles

The techniques for rolling dice can easily be extended to other situations. Instead of displaying an actual number, use JMP to re-code the random number into something else.

For example, suppose you want to simulate coin flips. There are two outcomes that (in a fair coin) occur with equal probability. One way to simulate this is to draw random numbers from a uniform distribution, where all numbers between 0 and 1 occur with equal probability. If the selected number is below 0.5, declare that the coin landed heads up. Otherwise, declare that the coin landed tails up.

- Create a new data table with two columns.
- In the first column, enter the following formula:



- Add rows to the data table to see the column fill with coin flips.

Extending this to sampling candies of different colors is easy. Suppose you have a bag of multi-colored candies with the following distribution.

Color	Percentage
Blue	10%
Brown	10%
Green	10%
Orange	10%
Red	20%
Yellow	20%
Purple	20%

Also, suppose you had a column named t that held random numbers from a uniform distribution. Then an appropriate JMP formula could be

```
If(t<0.1 ⇒ "Blue",
   t<0.2 ⇒ "Brown",
   t<0.3 ⇒ "Green",
   t<0.4 ⇒ "Orange",
   t<0.6 ⇒ "Red",
   t<0.8 ⇒ "Yellow",
   else  ⇒ "Purple")
```

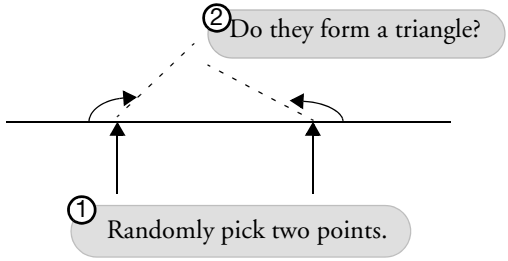
Note that JMP assigns the value associated with the first condition that is true. So, if t=0.18, “Brown” is assigned, and no further formula evaluation is done.

Or, you could use a slightly more complicated formula using a local variable to combine the random number and candy selection into one formula. Note the semicolon separating the two statements.

```
t=Random Uniform();
If(t<0.1 ⇒ "Blue",
   t<0.2 ⇒ "Brown",
   t<0.3 ⇒ "Green",
   t<0.4 ⇒ "Orange",
   t<0.6 ⇒ "Red",
   t<0.8 ⇒ "Yellow",
   else  ⇒ "Purple");
```

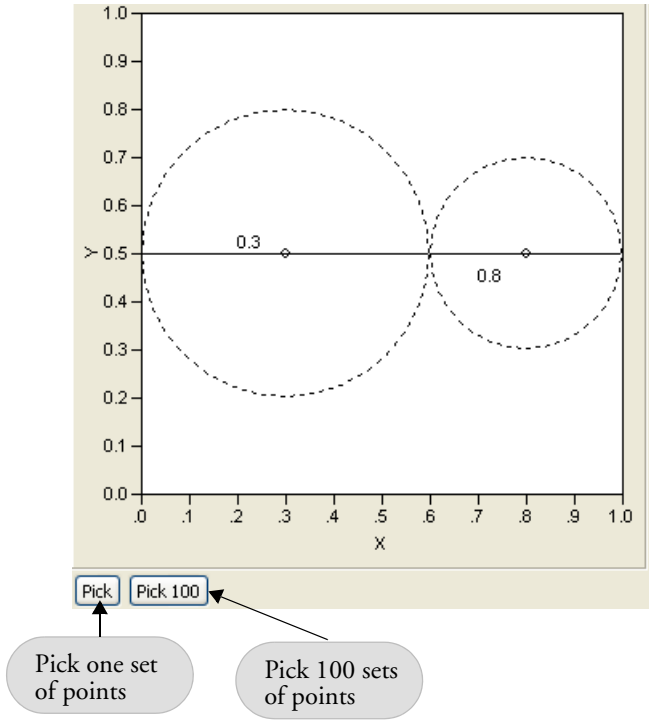
Probability of Making a Triangle

Suppose you randomly pick two points along a line segment. Then, break the line segment at these two points, forming three line segments. What is the probability that a triangle can be formed from these three segments? (Isaac, 1995)



This situation is simulated in the Triangle Probability.jsl script, found in the Sample Scripts folder. Upon running the script, a data table is created to hold the simulation results. The initial window is shown in **Figure 6.5**. For each of the two selected points, a dotted circle indicates the possible positions of the "broken" line segment that they determine.

Figure 6.5 Initial Triangle Probability Window

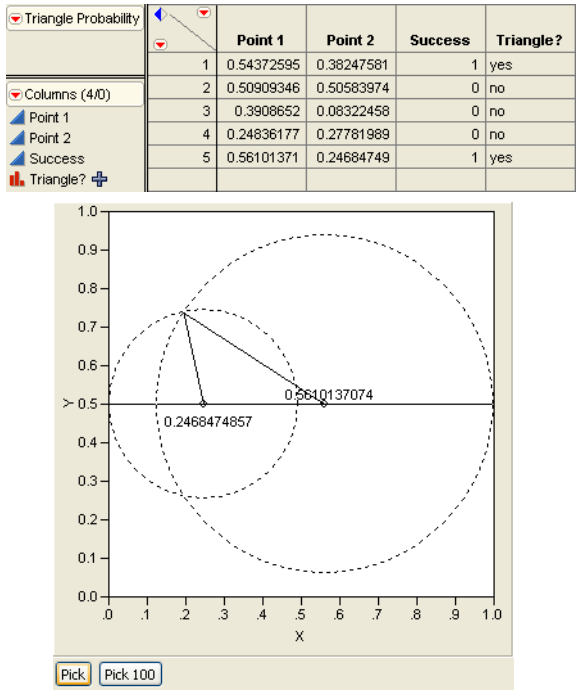


To use this simulation,

- Click the **Pick** button to pick a single pair of points.

Two points are selected and their information is added to a data table. The results after five simulations are shown in **Figure 6.6**.

Figure 6.6 Triangle Simulation after Five Iterations



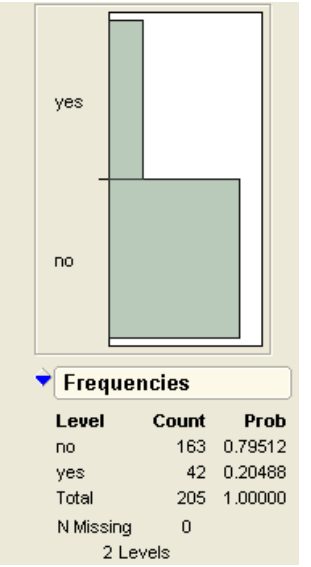
To get an idea of the theoretical probability, you need many rows in the data table.

- Click the **Pick 100** button a couple of times to generate a large number of samples.

When finished,

- Select **Analyze > Distribution** and select Triangle? as the **Y, Columns** variable.
- Click **OK** to see the distribution report in **Figure 6.7**.

Figure 6.7 Triangle Probability Distribution Report



It appears (in this case) that about 27% of the samples result in triangles. To investigate whether there is a relationship between the two selected points and their formation of a triangle,

- Select **Rows > Color or Mark by Column**.

When the selection dialog appears,

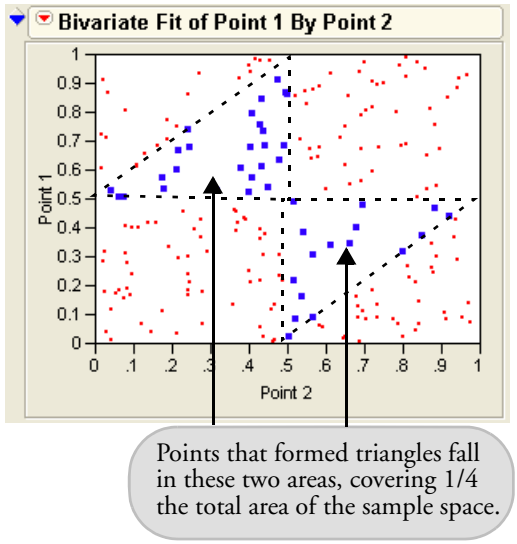
- Select the Triangle? column, making sure the **Set Color By Value** checkbox is checked.

This puts a different color on each row depending on whether it formed a triangle or not.

- Select **Analyze > Fit Y By X**, assigning Point1 to **Y** and Point2 to **X**.

This reveals a scatterplot that clearly shows a pattern.

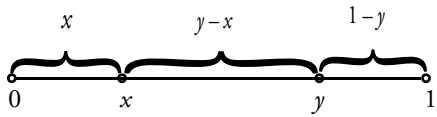
Figure 6.8 Scatterplot of Point1 by Point2



The entire sample space is in a unit square, and the points that formed triangles occupy one fourth of that area. This means that there is a 25% probability that two randomly selected points form a triangle.

Analytically, this makes sense. If the two randomly selected points are x and y, letting x represent the smaller of the two, then we know $0 < x < y < 1$, and the three segments have length x, $y - x$, and $1 - y$ (see **Figure 6.9**).

Figure 6.9 Illustration of Points



To make a triangle, the sum of the lengths of any two segments must be larger than the third, giving us the following conditions on the three points:

$$x + (y - x) > 1 - y$$
$$(y - x) + (1 - y) > x$$
$$(1 - y) + x > y - x$$

Elementary algebra simplifies these inequalities to

$$x < 0.5$$
$$y > 0.5$$
$$y - x < 0.5$$

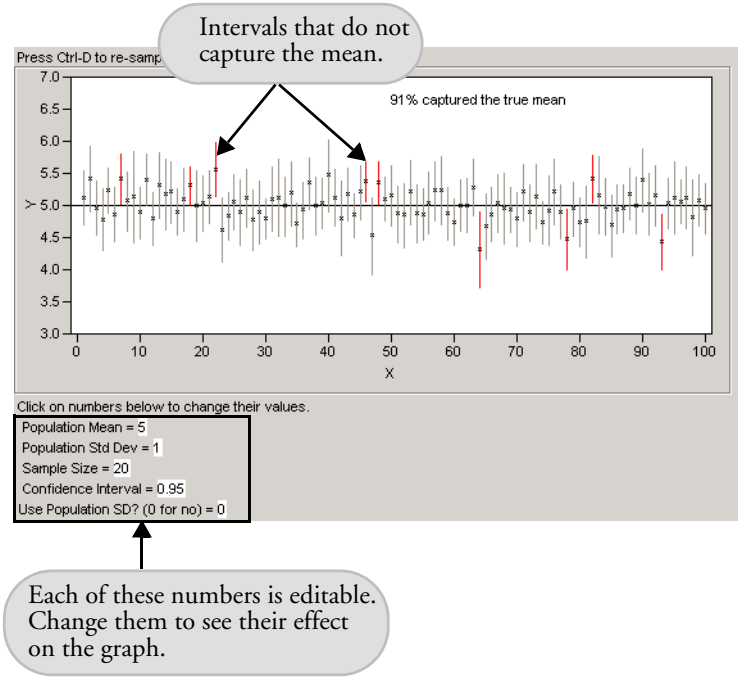
which explain the upper triangle in **Figure 6.8**. Repeating the same argument with y as the smaller of the two variables explains the lower triangle.

Confidence Intervals

Introductory students of statistics are often confused by the exact meaning of confidence intervals. For example, they frequently think that a 95% confidence interval contains 95% of the data. They do not realize that the confidence measurement is on the test methodology itself.

To demonstrate the concept, use the Confidence.jsl script. Its output is shown in **Figure 6.10**.

Figure 6.10 Confidence Interval Script



The script draws 100 samples of sample size 20 from a Normal distribution. For each sample, the mean is computed with a 95% confidence interval. Each interval is graphed, in gray if the interval captures the overall mean and in red if it doesn't.

Press Ctrl+D (⌘+D on the Macintosh) to generate another series of 100 samples. Each time, note the number of times the interval captures the theoretical mean. The ones that don't capture the mean are due only to chance, since we are randomly drawing the samples. For a 95% confidence interval, we expect that around five will not capture the mean, so seeing a few is not remarkable.

This script can also be used to illustrate the effect of changing the confidence interval.

- Change the confidence interval to 0.5.

This shrinks the size of the confidence intervals on the graph.

The **Use Population SD** interval allows you to use the population standard deviation in the computation of the confidence intervals (rather than the one from the sample). When enabled, all the confidence intervals are of the same size.

7 Univariate Distributions: One Variable, One Sample

Overview

This chapter introduces statistics in the simplest possible setting—the distribution of values in one variable. The **Distribution** command in the **Analyze** menu launches the JMP *Distribution platform*. This platform is used to describe the distribution of a single column of values from a table, using graphs and summary statistics.

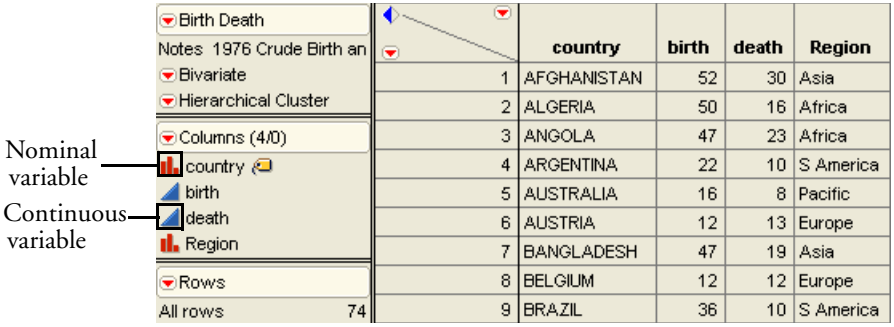
This chapter also introduces the concept of the distribution of a statistic, and how confidence intervals and hypothesis tests can be obtained.

Looking at Distributions

Let's take a look at some actual data and start noticing aspects of its distribution.

- Begin by opening the data table called Birth Death.jmp, which contains the 1976 birth and death rates of 74 nations **(Figure 7.1)**.
- Choose **Analyze > Distribution**.
- Select birth, death, and Region columns as the Y variables and click **OK**.

Figure 7.1 Partial Listing of the Birth Death.jmp Data Table

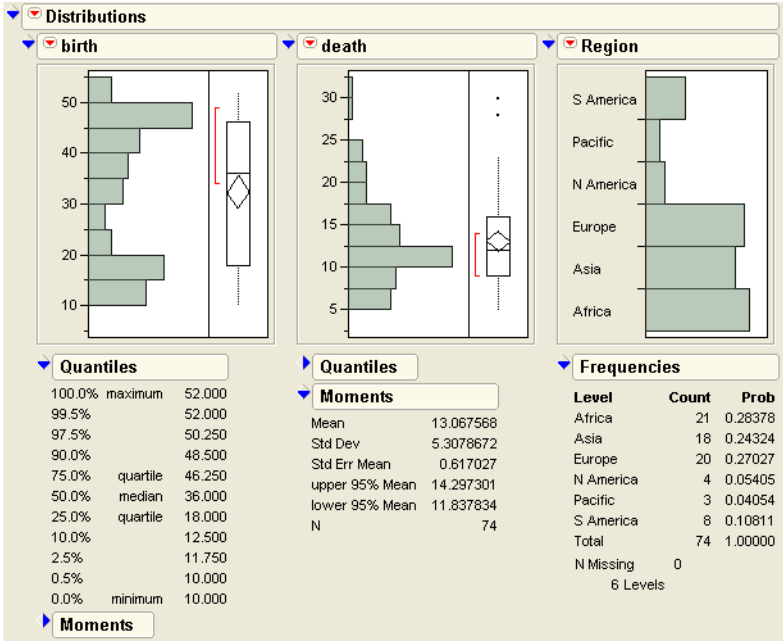


	country	birth	death	Region
1	AFGHANISTAN	52	30	Asia
2	ALGERIA	50	16	Africa
3	ANGOLA	47	23	Africa
4	ARGENTINA	22	10	S America
5	AUSTRALIA	16	8	Pacific
6	AUSTRIA	12	13	Europe
7	BANGLADESH	47	19	Asia
8	BELGIUM	12	12	Europe
9	BRAZIL	36	10	S America

When you see the report (**Figure 7.2**), be adventuresome: scroll around and click in various places on the surface of the report. Notice that a histogram or statistical table can be opened or closed by clicking the disclosure button on its title bar.

- Open and close tables, and click on bars until you have the configuration shown in **Figure 7.2**.

Figure 7.2 Histograms, Quantiles, Moments, and Frequencies



Note that there are two kinds of analyses:

- The analyses for birth and death are for continuous distributions. These are the kind of reports you get when the column in the data table has the continuous modeling type. The next to the column name in the columns panel indicates that this variable is continuous.
- The analysis for Region is for a categorical distribution. These are the kind of graphs and reports you get when the column in the data table has the modeling type of nominal or ordinal, or next to the column name in the Columns panel.

You can change the modeling type of any variable in the Columns panel to control which kind of report you get.

For continuous distributions, the graphs give a general idea of the shape of the distribution. The death data cluster together with most values near the center. Distributions like this one, with one peak, are called *unimodal*. However, the birth data have a different distribution. There are many countries with high birth rates, many with low birth rates, but few with middle rates. Therefore, the birth data has two peaks, and is referred to as *bimodal*.

The text reports for birth and death show a number of measurements concerning the distributions. There are two broad families of measures:

- *Quantiles* are the points at which various percentages of the total sample are above or below.
- *Moments* combine the individual data points to form descriptions of the entire data set. These combinations are usually simple arithmetic operations that involve sums of values raised to a power. Two common moments are the *mean* and *standard deviation*.

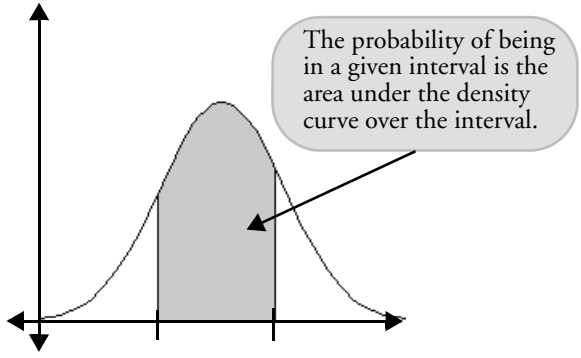
The report for the categorical distribution focuses on frequency counts. This chapter concentrates on continuous distributions and postpones the discussion of categorical distributions until Chapter 9, "Categorical Distributions."

Before going into the details of the analysis, let's review the distinctions between the properties of a distribution and the estimates that can be obtained from a distribution.

Probability Distributions

A *probability distribution* is the mathematical description of how a random process distributes its values. Continuous distributions are described by a *density function*. In statistics, we are often interested in the probability of a random value falling between two values described by this density function (for example, "What's the probability that I will gain between 100 and 300 points if I take the SAT a second time?"). The probability that a random value falls in a particular interval is represented by the area under the density curve in this interval, as illustrated in **Figure 7.3**.

Figure 7.3 Continuous Distribution



The density function describes all possible values of the random variable, so the area under the whole density curve must be 1, representing 100% probability. In fact, this is a defining

characteristic of all density functions. In order for a function to be a density function, it must be non-negative and the area underneath the curve must be 1.

These mathematical probability distributions are useful because they can model distributions of values in the real world. This book avoids distributional functions, but you should learn their names and their uses.

True Distribution Function or Real-World Sample Distribution

Sometimes it is hard to keep straight when you are referring to the real data sample and when you are referring to its abstract mathematical distribution.

This distinction of the *property* from its *estimate* is crucial in avoiding misunderstanding. Consider the following problem:

How is it that statisticians talk about the variability of a mean, that is, the variability of a single number? When you talk about variability in a sample of values, you can see the variability because you have many different values. However, when computing a mean, the entire list of numbers has been condensed to a single number. How does this mean—a single number—have variability?

To get the idea of variance, you have to separate the abstract quality from its estimate. When you do statistics, you are assuming that the data come from a process that has a random element to it. Even if you have a single response value (like a mean), there is variability associated with it—a magnitude whose value is possibly unknown.

For instance, suppose you are interested in finding the average height of males in the United States. You decide to compute the mean of a sample of 100 people. If you replicate this experiment several times gathering different samples each time, do you expect to get the same mean for every sample you pick? Of course not. There is variability in the sample means. It is this variability that statistics tries to capture—even if you don't replicate the experiment. Statistics can estimate the variability in the mean, even if it has only a single experiment to examine. The variability in the mean is called the *standard error* of the mean.

If you take a collection of values from a random process, sum them, and divide by the number of them, you have calculated a mean. You can then calculate the variance associated with this single number. There is a simple algebraic relationship between the variability of the responses (the standard deviation of the original data) and the variability of the sum of the responses

divided by n (the standard error of the mean). Complete details follow in the section "Standard Error of the Mean" on page 135.

Table 7.1. Properties of Distribution Functions and Samples

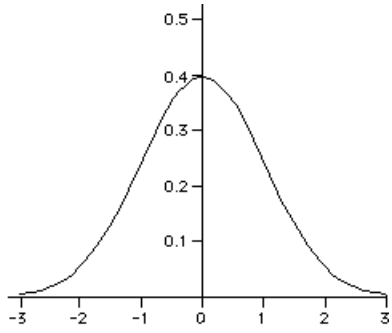
Concept	Abstract mathematical form, probability distribution	Numbers from the real world, data, sample
Mean	Expected value or true mean, the point that balances each side of the density	Sample mean, the sum of values divided by the number of values
Median	Median, the mid-value of the density area, where 50% of the density is on either side	Sample median, the middle value where 50% of the data are on either side
Quantile	The value where some percent of the density is below it	Sample quantile, the value for which some percent of the data are below it. For example, the 90th percentile represents a point where 90 percent of the variables are below it.
Spread	Variance, the expected squared deviation from the expected value	Sample variance, the sum of squared deviations from the sample mean divided by $n - 1$
General Properties	Any function of the distribution: parameter, property	Any function of the data: estimate, statistic

The statistic from the real world estimates the parameter from the distribution.

The Normal Distribution

The most notable continuous probability distribution is the *Normal distribution*, also known as the *Gaussian distribution*, or the *bell curve*, like the one shown in **Figure 7.4**.

Figure 7.4 Standard Normal Density Curve



It is an amazing distribution.

Mathematically, its greatest distinction is that it is the most random distribution for a given variance. (It is "most random" in a very precise sense, having maximum expected unexpectedness or entropy.) Its values are as if they had been realized by adding up billions of little random events.

It is also amazing because so much of real world data are Normally distributed. The Normal distribution is so basic that it is the benchmark used as a comparison with the shape of other distributions. Statisticians describe sample distributions by saying how they differ from the Normal. Many of the methods in JMP serve mainly to highlight how a distribution of values differs from a Normal distribution. However, the usefulness of the Normal distribution doesn't end there. The Normal distribution is also the standard used to derive the distribution of estimates and test statistics.

The famous *Central Limit Theorem* says that under various fairly general conditions, the sum of a large number of independent and identically distributed random variables is approximately Normally distributed. Because most statistics can be written as these sums, they are Normally distributed if you have enough data. Many other useful distributions can be derived as simple functions of random Normal distributions.

Later in this chapter, you meet the distribution of the mean and learn how to test hypotheses about it. The next sections introduce the four most useful distributions of test statistics: the Normal, Student's *t*, chi-square, and *F* distributions.

Describing Distributions of Values

The following sections take you on a tour of the graphs and statistics in the JMP Distribution platform. These statistics try to show the properties of the distribution of a sample, especially these four focus areas:

- *Location* refers to the center of the distribution.
- *Spread* describes how concentrated or "spread out" the distribution is.
- *Shape* refers to symmetry, whether the distribution is unimodal, and especially how it compares to a Normal distribution.
- *Extremes* are outlying values far away from the rest of the distribution.

Generating Random Data

Before getting into more real data, let's make some random data with familiar distributions, and then see what an analysis reveals. This is an important exercise because there is no other way to get experience on the distinction between the true distribution of a random process and the distribution of the values you get in a sample.

In Plato's mode of thinking, the "true" world is some ideal form, and what you perceive as real data is only a shadow that gives hints at what the true world is like. Most of the time the true state is unknown, so an experience where the true state is known is valuable.

In the following example, the true world is a distribution, and you use the random number generator in JMP to obtain realizations of the random process to make a sample of values. Then you will see that the sample mean of those values is not exactly the same as the true mean of the original distribution. This distinction is fundamental to what statistics is all about.

To create your own random data,

- Open Randdist.Jmp.

This data table has four columns, but no rows. The columns contain formulas used to generate random data having the distributions Uniform, Normal, Exponential, and Dbl Exponential (double exponential).

- **Choose Rows** > **Add Rows** and enter 1000 to see a table like the one in **Figure 7.5**.

Adding rows generates the random data using the column formulas. Note that your random results will be a little different from those shown in **Figure 7.5** because the random number generator produces a different set of numbers each time a table is created.

Figure 7.5 Partial Listing of the Randdist Data Table

Randdist

Columns (4/0)
- Uniform
- Normal
- Exponential
- Dbl Expon

Rows
All rows 1000

	Uniform	Normal	Exponential	Dbl Expon
1	0.883446	-0.60082	0.381821	0.195129
2	0.337475	0.51715	1.143111	0.895243
3	0.979203	0.363712	3.215993	-1.70657
4	0.386247	-0.68428	0.450302	0.232997
5	0.776403	-0.74783	0.255144	-0.04085
6	0.002862	-0.21798	0.895585	0.927462
7	0.709449	-0.20849	0.582944	1.063657
8	0.489618	-0.19025	0.086532	0.258683
9	0.363204	0.181757	0.304229	0.102432

- To look at the distributions of the columns in the Randdist.jmp table, choose **Analyze** > **Distribution**, and select the four columns as Y variables.
- Click **OK**.

The resulting analysis automatically shows a number of graphs and statistical reports. Further graphs and reports (**Figure 7.6**, for example) can be clicking on the red triangle menu in the report title bar of each analysis. The following sections examine these graphs and the text reports available in the Distribution platform.

Histograms

A *histogram* defines a set of intervals and shows how many values in a sample fall into each interval. It shows the shape of the density of a batch of values.

Try out the following histogram features:

- Click in a histogram bar.

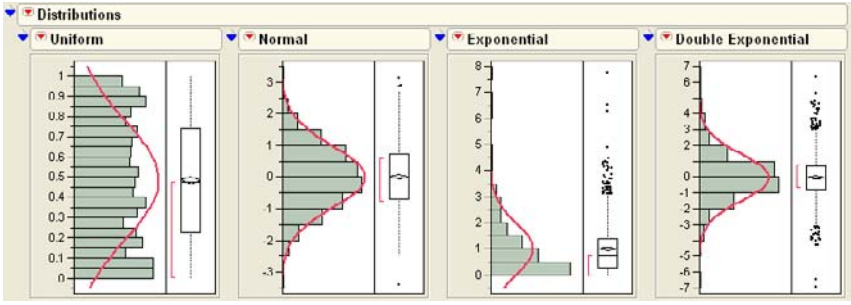
When the bar highlights, the corresponding portions of bars in other histograms also highlight, as do the corresponding data table rows. When you do this, you are seeing *conditional distributions*—the distributions of other variables corresponding to a subset of the selected variable's distribution.

- Double-click on a histogram bar to produce a new JMP table that is a subset corresponding to that bar.

- On the original Distribution plot that you just clicked, choose the **Normal** option from the **Fit Distribution** command on the menu at the left of the report titles.

This superimposes over the histogram the Normal density corresponding to the mean and standard deviation in your sample. **Figure 7.6** shows the histograms from the analysis with Normal curves superimposed on them.

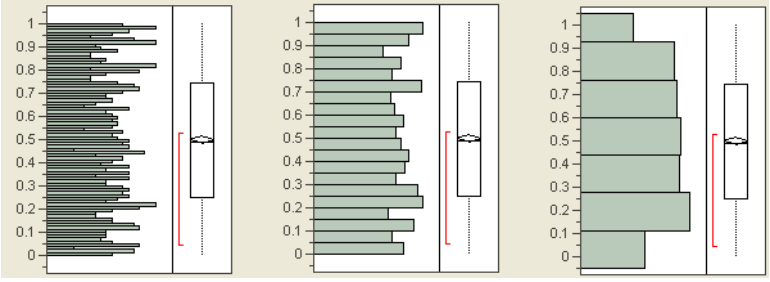
Figure 7.6 Histograms of Various Continuous Distributions



- Get the hand tool from the **Tools** menu or toolbar.
- Click on the Uniform histogram and drag to the right, then back to the left.

The histogram bars get narrower and wider (see **Figure 7.7**).

Figure 7.7 The Hand Tool Adjusts Histogram Bar Widths



- Make them wider and then drag up and down to change the position of the bars.

Stem-and-Leaf Plots

A *stem-and-leaf plot* is a variation on the histogram. It was developed for tallying data in the days when computers were rare and histograms took a lot of time to make. Each line of the plot has a stem value that is the leading digits of a range of column values. The leaf values are

made from other digits of the values. As a result, the stem-and-leaf plot has a shape that looks similar to a histogram, but also shows the data points themselves.

To see two examples, open the Big Class.jmp and the Automess.jmp tables.

- For each table choose **Analyze > Distribution**. The Y variables are weight from the Big Class table and Auto theft from the Automess table.
- When the histograms appear, select **Stem and Leaf** from the options popup menu next to the histogram name.

This option appends stem-and-leaf plots to the end of the text reports.

Figure 7.8 shows the plot for weight on the left and the plot for Auto theft on the right. The values in the stem column of the plot are chosen as a function of the range of values to be plotted.

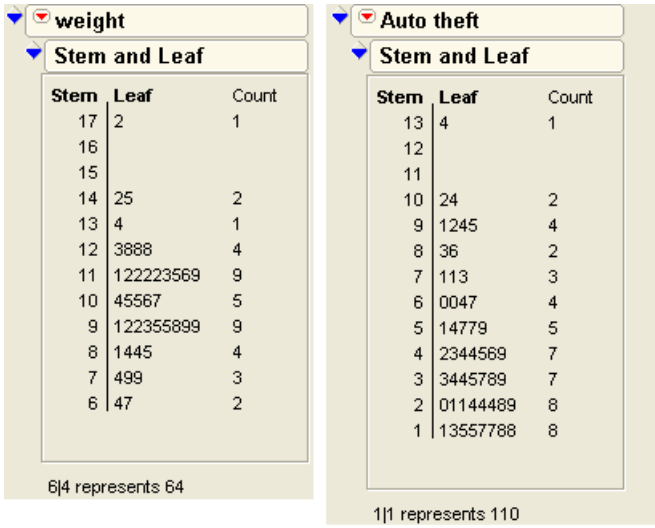
You can reconstruct the data values by joining the stem and leaf as indicated by the legend on the bottom of the plot. For example, on the bottom line of the weight plot, corresponding to data values 64 and 67 (6 from the stem, 4 and 7 from the leaf). At the top, the weight is 172 (17 from the stem, 2 from the leaf).

The leaves respond to mouse clicks.

- Click on the two 5s on the bottom stem of the Auto theft plot.

This highlights the corresponding rows in the data table, which are “California” with the value 154 and the “District of Columbia” with value of 149.

Figure 7.8 Examples of Stem-and-Leaf Plots



weight

Stem and Leaf

Stem	Leaf	Count
17	2	1
16		
15		
14	25	2
13	4	1
12	3888	4
11	122223569	9
10	45567	5
9	122355899	9
8	1445	4
7	499	3
6	47	2

6|4 represents 64

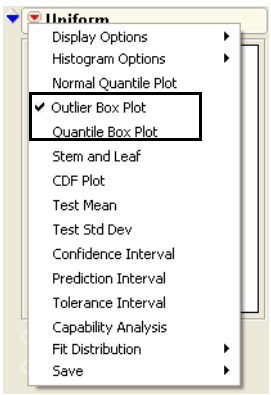
Auto theft

Stem and Leaf

Stem	Leaf	Count
13	4	1
12		
11		
10	24	2
9	1245	4
8	36	2
7	113	3
6	0047	4
5	14779	5
4	2344569	7
3	3445789	7
2	01144489	8
1	13557788	8

1|1 represents 110

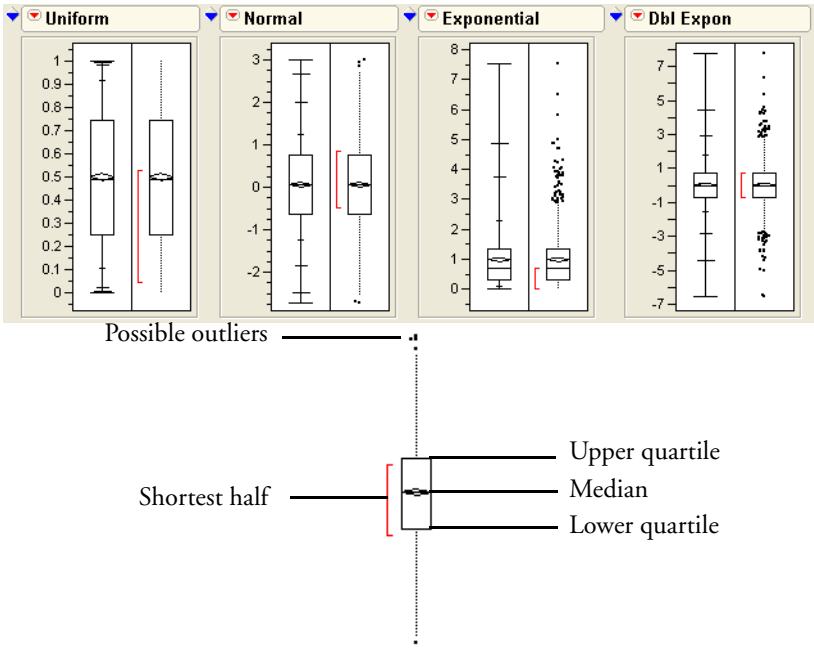
Outlier and Quantile Box Plots



Box plots are schematics that also show how data are distributed. The Distribution platform offers two varieties of box plots that you can turn on or off with options accessed by the red triangle menu on the report title bar, as shown here.These are outlier and quantile box plots.

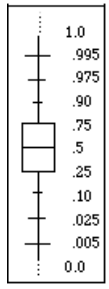
Figure 7.9 shows these box plots for the simulated distributions. The box part within each plot surrounds the middle half of the data. The lower edge of the rectangle represents the lower quartile, the higher edge represents the upper quartile, and the line in the middle of the rectangle is the median. The distance between the two edges of the rectangle is called the *interquartile range*. The lines extending from the box show the tails of the distribution, points that the data occupy outside the quartiles. These lines are sometimes called *whiskers*.

Figure 7.9 Quantile and Outlier Box Plots



In the outlier box plots, shown on the right of each panel in **Figure 7.9**, the tail extends to the farthest point that is still within 1.5 interquartile ranges from the quartiles. Points farther away are possible outliers and are shown individually as points.

In the quantile box plots (shown on the left in each panel) the tails are marked at certain quantiles. The quantiles are chosen so that if the distribution is Normal, the marks appear approximately equidistant, like the figure on the right. The spacing of the marks in these box plots give you a clue about the Normality of the underlying distribution.



Look again at the boxes in the four distributions in **Figure 7.9**, and examine the middle half of the data in each graph. The middle half of the data is wide in the uniform, thin in the double exponential, and very one-sided in the exponential distribution.

In the outlier box plot, the shortest half (the shortest interval containing 50% of the data) is shown by a red bracket on the side of the box plot. The shortest half is at the center for the symmetric distributions, but off-center for non-symmetric ones. Look at the exponential distribution to see an example of a non-symmetric distribution.

In the quantile box plot, the mean and its 95% confidence interval are shown by a diamond. Since this experiment was created with 1000 observations, the mean is estimated with great precision, giving a very short confidence interval, and thus a thin diamond. Confidence intervals are discussed in the following sections.

Mean and Standard Deviation

The *mean* of a collection of values is its average value, computed as the sum of the values divided by the number of values in the sum. Expressed mathematically,

$$\bar{x} = \frac{x_1 + x_1 + \ldots + x_n}{n} = \sum \frac{x_i}{n}$$

The sample mean has these properties:

- It is the balance point. The sum of deviations of each sample value from the sample mean is zero.
- It is the least squares estimate. The sum of squared deviations of the values from the mean is minimized. That sum is less than would be computed from any estimate other than the sample mean.
- It is the maximum likelihood estimator of the true mean when the distribution is Normal. It is the estimate that makes the data you collected more likely than any other estimate of the true mean would.

The sample *variance* (denoted s^2) is the average squared deviation from the sample mean, which is shown as the expression

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$$

The sample *standard deviation* is the square root of the sample variance.

$$s = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}}$$

The standard deviation is preferred in reports because (among other reasons) it is in the same units as the original data (rather than squares of units).

If you assume a distribution is Normal, you can completely characterize its distribution by its mean and standard deviation.

When you say "mean" and "standard deviation," you are allowed to be ambiguous as to whether you are referring to the true (and usually unknown) parameters of the distribution, or the sample statistics you use to estimate the parameters.

Median and Other Quantiles

Half the data are above and half are below the sample *median*. It estimates the 50th quantile of the distribution. A sample quantile can be defined for any percentage between 0% and 100%; the 100% quantile is the maximum value, where 100% of the data values are at or below. The 75% quantile is the upper quartile, the value for which 75% of the data values are at or below.

There is an interesting indeterminacy about how to report the median and other quantiles. If you have an even number of observations, there may be several values where half the data are above, half below. There are about a dozen different ways for reporting medians in the statistical literature, many of which are only different if you have tied points on either or both sides of the middle. You can take one side, the other, the midpoint, or a weighted average of the middle values, with a number of weighting options. For example, if the sample values are {1, 2, 3, 4, 4, 5, 5, 5, 7, 8}, the median can be defined anywhere between 4 and 5, including one side or the other, or half way, or two-thirds of the way into the interval. The halfway point is the most common value chosen.

Another interesting property of the median is that it is the least-absolute-values estimator. That is, it is the number that minimizes the sum of the absolute differences between itself and each value in the sample. Least-absolute-values estimators are also called *L1 estimators*, or *Minimum Absolute Deviation (MAD) estimators*.

Mean versus Median

If the distribution is symmetric, the mean and median are estimates of both the expected value of the underlying distribution and its 50% quantile. If the distribution is Normal, the mean is a "better" estimate (in terms of variance) than the median, by a ratio of 2 to 3.1416 (2: π). In other words, the mean has only 63% of the variance of the median.

If an outlier contaminates the data, the median is not greatly affected, but the mean could be greatly influenced, especially if the outlier is extreme. The median is said to be *outlier-resistant*, or *robust*.

Suppose you have a skewed distribution, like household income in the United States. This set of data has lots of extreme points on the high end, but is limited to zero on the low end. If you

want to know the income of a typical person, it makes more sense to report the median than the mean. However, if you want to track per-capita income as an aggregating measure, then the mean income might be better to report.

Higher Moments: Skewness and Kurtosis

Moment statistics are those that are formed from sums of powers of the data's values. The first four moments are defined as follows:

- The first moment is the mean, which is calculated from a sum of values to the power 1. The mean measures the center of the distribution.
- The second moment is the variance (and, consequently, the standard deviation), which is calculated from sums of the values to the second power. Variance measures the spread of the distribution.
- The third moment is *skewness*, which is calculated from sums of values to the third power. Skewness measures the asymmetry of the distribution.
- The fourth moment is *kurtosis*, which is calculated from sums of the values to the fourth power. Kurtosis measures the relative shape of the middle and tails of the distribution.

One use of skewness and kurtosis is to help determine if a distribution is Normal and, if not, what the distribution might be. A problem with the higher order moments is that the statistics have higher variance and are more sensitive to outliers.

To get the skewness and kurtosis, use the red popup menu beside the title of the histogram and select **Display Options > More Moments** from the drop-down list next to the histogram's title.

Extremes, Tail Detail

The extremes (the minimum and maximum) are the 0% and 100% quantiles.

At first glance, the most interesting aspect of a distribution appears to be where its center lies. However, statisticians often look first at the outlying points—they can carry useful information. That's where the unusual values are, the possible contaminants, the rogues, and the potential discoveries.

In the Normal distribution (with infinite tails), the extremes tend to extend farther as you collect more data. However, this is not necessarily the case with other distributions. For data that are uniformly distributed across an interval, the extremes change less and less as more data are collected. Sometimes this is not helpful, since the extremes are often the most informative statistics on the distribution.

Statistical Inference on the Mean

The previous sections talked about descriptive graphs and statistics. This section moves on to the real business of statistics: inference. We want to form confidence intervals for a mean and test hypotheses about it.

Standard Error of the Mean

Suppose there exists some true (but unknown) population mean that you estimate with the sample mean. The sample mean comes from a random process, so there is variability associated with it.

The mean is the arithmetic average—the sum of n values divided by n. The variance of the mean has $1/n$ of the variance of the original data. Since the standard deviation is the square root of the variance, the standard deviation of the sample mean is $1/\sqrt{n}$ of the standard deviation of the original data.

Substituting in the estimate of the standard deviation of the data, we now define the *standard error of the mean*, which estimates the standard deviation of the sample mean. It is the standard deviation of the data divided by the square root of n.

Symbolically, this is written

$$s_y = \frac{s_y}{\sqrt{n}}$$

where s_y is the sample standard deviation.

The mean and its standard error are the key quantities involved in statistical inference concerning the mean.

Confidence Intervals for the Mean

The sample mean is sometimes called a *point estimate*, because it's only a single number. The true mean is not this point, but rather this point is an estimate of the true mean.

Instead of this single number, it would be more useful to have an interval that you are pretty sure contains the true mean (say, 95% sure). This interval is called a *95% confidence interval* for the true mean.

To construct a confidence interval, first make some assumptions. Assume:

- The data are Normal, and

- The true standard deviation is the sample standard deviation. (This assumption will be revised later.)

Then, the exact distribution of the mean estimate is known, except for its location (because you don't know the true mean).

If you knew the true mean and had to forecast a sample mean, you could construct an interval around the true mean that would contain the sample mean with probability 0.95. To do this, first obtain the quantiles of the standard Normal distribution that have 5% of the area in their tails. These quantiles are –1.96 and +1.96.

Then, scale this interval by the standard deviation and add in the true mean. Symbolically, compute

$$\mu \pm 1.96 s_{\bar{y}}$$

However, our present example is the reverse of this situation. Instead of a forecast, you already have the sample mean; instead of an interval for the sample mean, you need an interval to capture the true mean. If the sample mean is 95% likely to be within this distance of the true mean, then the true mean is 95% likely to be within this distance of the sample mean. Therefore, the interval is centered at the sample mean. The formula for the approximate 95% confidence interval is

$$\text{95\% C.I. for the mean} = \bar{x} \pm 1.96 s_{\bar{y}}$$

Figure 7.10 illustrates the construction of confidence intervals. This is not exactly the confidence interval that JMP calculates. Instead of using the quantile of 1.96 (from the Normal distribution), it uses a quantile from Student's *t* distribution, discussed later. It is necessary to use this slightly modified version of the Normal distribution because of the extra uncertainty that results from estimating the standard error of the mean (which, in this example, we are assuming is known). So the formula for the confidence interval is

$$(1-\alpha)\text{ C.I. for the mean} = \bar{x} \pm \left(t_{1-\frac{\alpha}{2}} \cdot s_{\bar{y}} \right)$$

The alpha (α) in the formula is the probability that the interval does not capture the true mean. That probability is 0.05 for a 95% interval. The confidence interval is reported on the Distribution platform in the Moments report as the Upper 95% Mean and Lower 95% Mean. It is represented in the quantile box plot by the ends of a diamond (see **Figure 7.11**).

Figure 7.10 Illustration of Confidence Interval

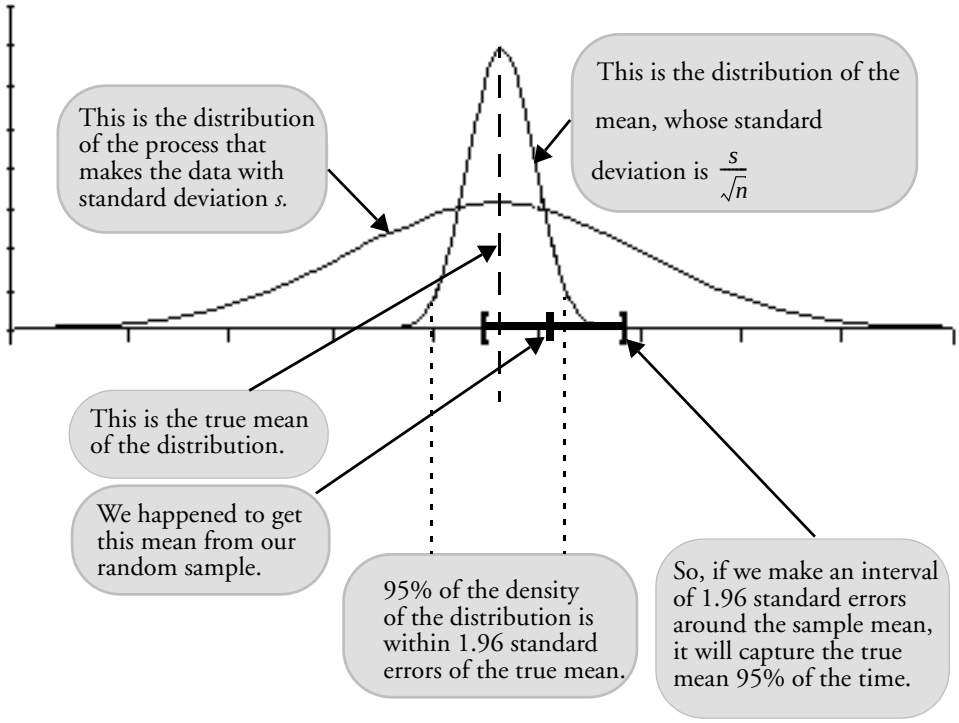
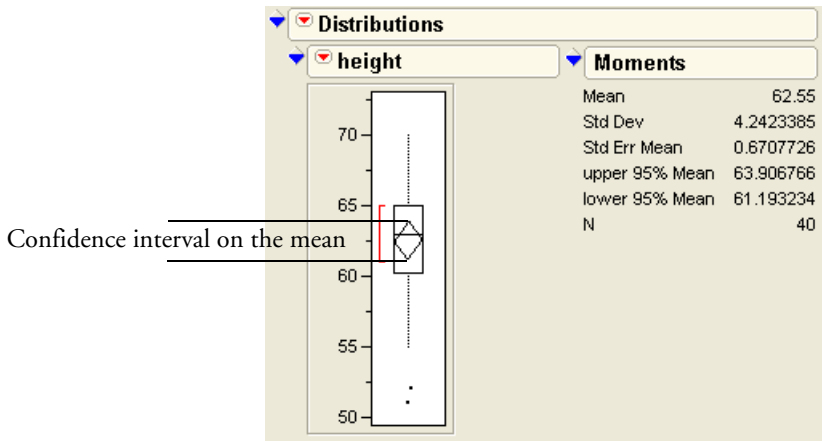


Figure 7.11 Moments Report and Quantile Box Plot



If you have not done so, you should read the section "Confidence Intervals" on page 117 in the Simulations chapter and run the associated script.

Testing Hypotheses: Terminology

Suppose you want to test whether the mean of a collection of sample values is significantly different from a hypothesized value. The strategy is to calculate a statistic so that if the true mean were the hypothesized value, getting such a large computed statistic value would be an extremely unlikely event. You would rather believe the hypothesis to be false than to believe this rare coincidence happened. This is a probabilistic version of *proof by contradiction*.

The way you see an event as rare is to see that its probability is past a point in the tail of the probability distribution of the hypothesis. Often, researchers use 0.05 as a significance indicator, which means you believe that the mean is different from the hypothesized value if the chance of being wrong is only 5% (one in twenty).

Statisticians have a precise and formal terminology for hypothesis testing:

- The possibility of the true mean being the hypothesized value is called the *null hypothesis*. This is frequently denoted H_0, and is the hypothesis you want to reject. Said another way, the null hypothesis is the possibility that the hypothesized value is not different from the true mean. The *alternative hypothesis,* denoted H_A, is that the mean is different from the hypothesized value. This can be phrased as greater than, less than, or unequal. The latter is called a *two-sided alternative*.
- The situation where you reject the null hypothesis when it happens to be true is called a *Type I error*. This declares that some difference is nonzero when it is really zero. The opposite mistake (not detecting a difference when there is a difference) is called a *Type II error.*
- The probability of getting a Type I error in a test is called the *alpha-level* (α-level) of the test. This is the probability that you are wrong if you say that there is a difference. The *beta-level* (β-level) or *power* of the test is the probability of being right when you say that there is a difference. $1 - \beta$ is the probability of a Type II error.
- Statistics and tests are constructed so that the power is maximized subject to the α-level being maintained.

In the past, people obtained critical values for α-levels and ended with an accept/reject decision based on whether the statistic was bigger or smaller than the critical value. For example, a researcher would declare that his experiment was significant if his test statistic fell in the region of the distribution corresponding to an α-level of 0.05. This α-level was specified in advance, before the study was conducted.

Computers have changed this strategy. Now, the α-level isn't pre-determined, but rather is produced by the computer after the analysis is complete. In this context, it is called a *p-value* or *significance level*. The definition of a *p*-value can be phrased in many ways:

- The *p*-value is the α-level at which the statistic would be significant.
- The *p*-value is how unlikely getting so large a statistic would be if the true mean were the hypothesized value.
- The *p*-value is the probability of being wrong if you rejected the null hypothesis. It is the probability of a Type I error.
- The *p*-value is the area in the tail of the distribution of the test statistic under the null hypothesis.

The *p*-value is the number you want to be very small, certainly below 0.05, so that you can say that the mean is significantly different from the hypothesized value. The *p*-values in JMP are labeled according to the test statistic's distribution. *p*-values below 0.05 are marked with an asterisk in many JMP reports. The label "Prob >|t|" is read as the "probability of getting an even greater absolute *t* statistic, given that the null hypothesis is true."

The Normal *z*-Test for the Mean

The Central Limit Theorem tells us that if the original response data are Normally distributed, then when many samples are drawn, the means of the samples are Normally distributed. More surprisingly, it says that even if the original response data are not Normally distributed, the sample mean still has an approximate Normal distribution if the sample size is large enough. So the Normal distribution provides a reference to use to compare a sample mean to an hypothesized value.

The standard Normal distribution has a mean of zero and a standard deviation of one. You can center any variable to mean zero by subtracting the mean (even the hypothesized mean). You can standardize any variable to have standard deviation 1 ("unit standard deviation") by dividing by the true standard deviation, assuming for now that you know what it is. This process is called *centering and scaling*. If the hypothesis were true, the test statistic you construct should have this standard distribution. Tests using the Normal distribution constructed like this (hypothesized mean but known standard deviation) are called *z-test*s. The formula for a *z*-statistic is

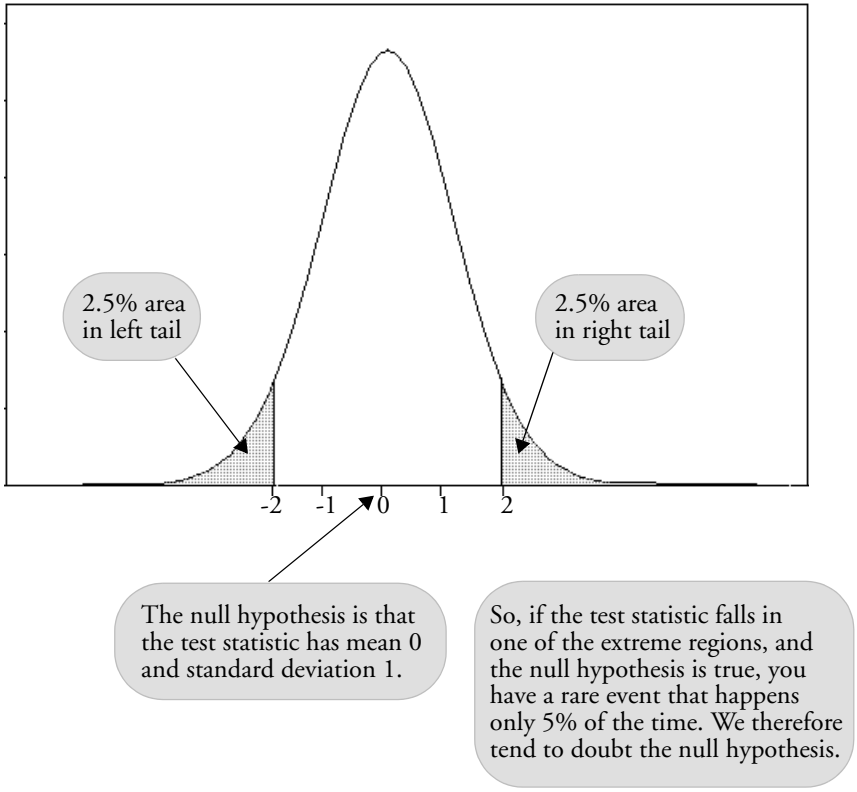
$$z - \text{statistic} = \frac{\text{estimate - hypothesized value}}{\text{standard deviation}}$$

You want to find out how unusual your computed *z*-value is from the point of view of believing the hypothesis. If the value is too improbable, then you doubt the null hypothesis.

To get a significance probability, you take the computed *z*-value and find the probability of getting an even greater absolute value. This involves finding the areas in the tails of the

Normal distribution that are greater than absolute *z* and less than negative absolute *z*. **Figure 7.12** illustrates a two-tailed *z*-test for $\alpha = 0.05$.

Figure 7.12 Illustration of Two-Tailed *z*-test



Case Study: The Earth's Ecliptic

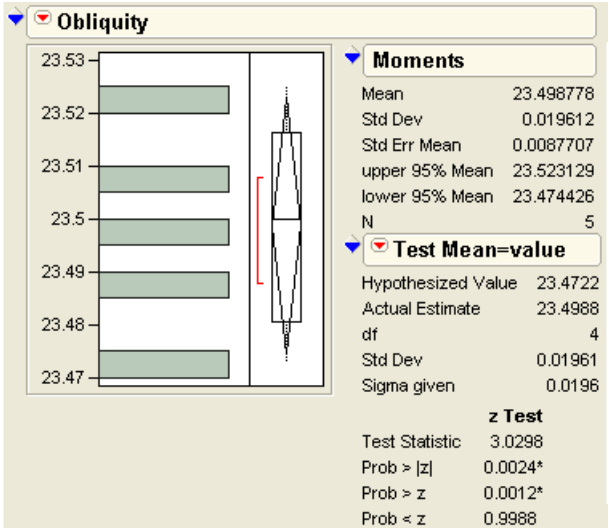
In 1738, the Paris observatory determined with high accuracy that the angle of the earth's spin was 23.472 degrees. However, someone suggested that the angle changes over time. Examining historical documents found five measurements dating from 1460 to 1570. These measurements were somewhat different than the Paris measurement, and they were done using much less precise methods. The question is whether the differences in the measurements can be attributed to the errors in measurement of the earlier observations, or whether the angle of the earth's rotation actually changed. We need to test the hypothesis that the earth's angle has actually changed.

Open Cassub.jmp (Stigler, 1986).

- Choose **Analyze > Distribution** and select Obliquity as the Y variable.
- Click **OK**.

The Distribution report shows a histogram of the five values.

Figure 7.13 Report of Observed Ecliptic Values



We now want to test that the mean of these values is different than the value from the Paris observatory. Our null hypothesis is that the mean is not different.

- Click on the red triangle menu on the report title and select **Test Mean**.
- In the dialog that appears, enter the hypothesized value of 23.47222 (the value measured by the Paris observatory), and enter the standard deviation of 0.0196 found in the Moments table (see **Figure 7.13**).
- Click **OK**.

The *z*-test statistic has the value 3.0298. The area under the Normal curve to the right of this value is reported as Prob > *z*, which is the probability (*p*-value) of getting an even greater *z*-value if there was no difference. In this case, the *p*-value is 0.001. This is an extremely small p-value. If our null hypothesis were true (for example, the measurements were the same), our measurement would be a highly unlikely observation. Rather than believe the unlikely result, we reject H_0 and claim the measurements are different.

Notice that we are only interested in whether the mean is greater than the hypothesized value. We therefore look at the value of Prob > z, a one-sided test. To test that the mean is different in either direction, the area in both tails is needed. This statistic is two-sided and listed as Prob >|z|, in this case 0.002. The one-sided test Prob < z has a p-value of 0.999, indicating that you are not going to prove that the mean is less than the hypothesized value. The two-sided p-value is always twice the smaller of the one-sided p-values.

Student's *t*-Test

The z-test has a restrictive requirement. It requires the value of the true standard deviation of the response, and thus the standard deviation of the mean estimate, be known. Usually this true standard deviation value is unknown and you have use an estimate of the standard deviation.

Using the estimate in the denominator of the statistical test computation requires an adjustment to the distribution that was used for the test. Instead of using a Normal distribution, statisticians use a *Student's t-distribution*. The statistic is called the *Student's t-statistic* and is computed by the formula shown to the right, where x_0 is the hypothesized mean and s is the sample standard deviation of the sample data. In words, you can say

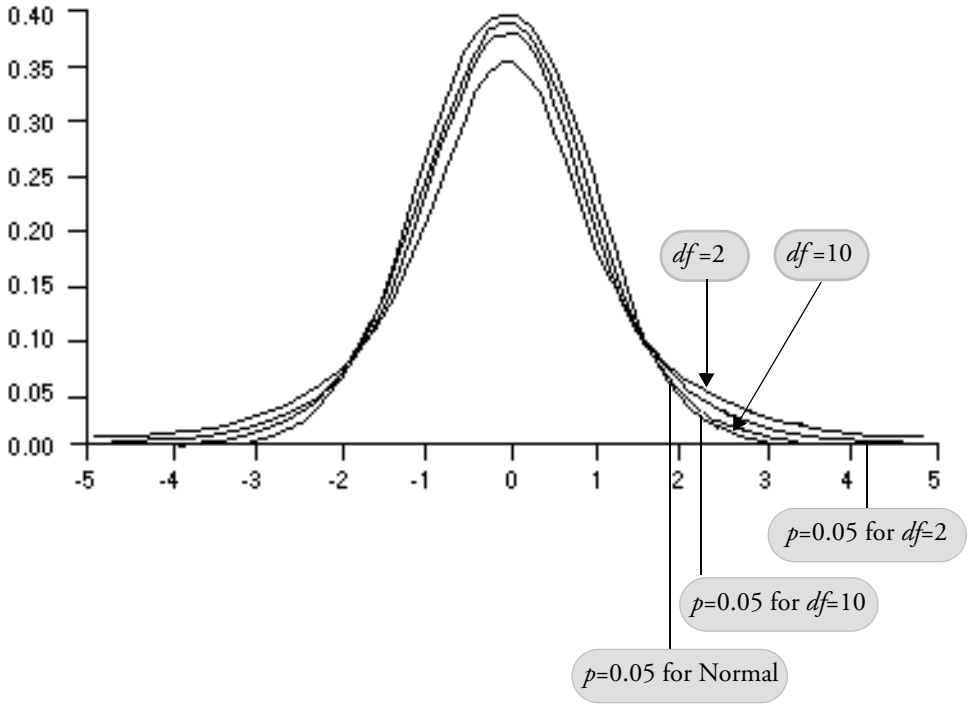
$$t = \frac{\bar{x} - x_0}{\frac{s}{\sqrt{n}}}$$

$$t\text{-statistic} = \frac{\text{sample mean} - \text{hypothesized value}}{\text{standard error of the mean}}$$

A large sample estimates the standard deviation very well, and the Student's t-distribution is remarkably similar to the Normal distribution, as illustrated in **Figure 7.14**. However, in this example there were only five observations.

There is a different t-distribution for each number of observations, indexed by a value called degrees of freedom, which is the number of observations minus the number of parameters estimated in fitting the model. In this case, five observations minus one parameter (the mean) yields 5 – 1 = 4 degrees of freedom. As you can see in **Figure 7.14**, the quantiles for the t-distribution spread out farther than the Normal when there are few degrees of freedom.

Figure 7.14 Comparison of Normal and Student's *t* Distributions



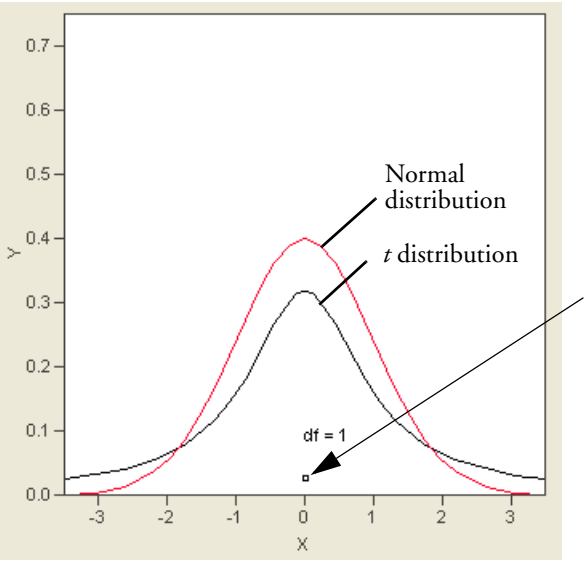
Comparing the Normal and Student's *t* Distributions

JMP can produce an animation to show you the relationships in **Figure 7.14**. This demonstration uses the Normal vs. t.JSL script. For more information on opening and running scripts, see "Working with Scripts" on page 57.

- Open the Normal vs t.JSL script.
- Choose **Edit > Run Script**.

You should see the window shown in **Figure 7.14**.

Figure 7.15 Normal vs *t* Comparison



The small square located just above 0 is called a *handle*. It is draggable, and adjusts the degrees of freedom associated with the black *t*-distribution as it moves. The Normal distribution is drawn in red.

- Click and drag the handle up and down to adjust the degrees of freedom of the *t*-distribution.

Notice both the height and the tails of the *t*-distribution. At what number of degrees of freedom do you feel that the two distributions are close to identical?

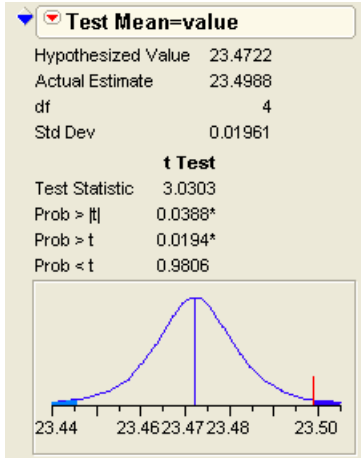
Testing the Mean

We now reconsider the ecliptic case study, so return to the **Cassub-Distribution of Obliquity** window. It turns out that for a 5% two-tailed test, which uses a *p*-value of a 0.975, the *t*-quantile for 4 degrees of freedom is 2.7764, which is far greater than the corresponding *z*-quantile of 1.96. That is, the bar for rejecting H_0 is higher, due to the fact that we don't know the standard deviation. Let's do the same test again, using this different value. Our null hypothesis is still that there is no change in the values.

- Select **Test Mean** and again enter 23.47222 for the hypothesized mean value. This time, do not fill in the standard deviation.

- Click **OK**.

The Test Mean table (shown here) now displays a *t*-test instead of a *z*-test (as in the Obliquity report in **Figure 7.13** on page 141).



When you don't specify a standard deviation, JMP uses the sample estimate of the standard deviation. The significance is smaller, but the *p*-value of 0.039 still looks convincing, so you can reject H_0 and conclude that the angle has changed.When you have a significant result, the idea is that under the null hypothesis, the expected value of the *t*-statistic is zero. It is highly unlikely (probability less than α) for the *t*-statistic to be so far out in the tails. Therefore, you don't put much belief in the null hypothesis.

Note: You may have noticed that the test dialog offers the options of a Wilcoxon signed-rank nonparametric test. Some statisticians favor nonparametric tests because the results don't depend on the response having a Normal distribution. Nonparametric tests are covered in more detail in the chapter "Comparing Many Means: One-Way Analysis of Variance" on page 209.

The *p*-Value Animation

Figure 7.12 on page 140 illustrates the relationship between the two-tailed test and the Normal distribution. Some questions may arise after looking at this picture.

- How would the *p*-value change if the difference between the truth and my observation were different?
- How would the *p*-value change if my test were one-sided instead of two sided?
- How would the *p*-value change if my sample size were different?

To answer these questions, JMP provides an animated demonstration, written in JMP scripting language. Often, these scripts are stored as separate files (samples are included in the Sample Scripts folder). However, some scripts are built into JMP. This *p*-value animation is an example of a built-in script.

Select **PValue Animation** from the red triangle menu on the **Test Mean** report title, which produces the window in **Figure 7.16**.

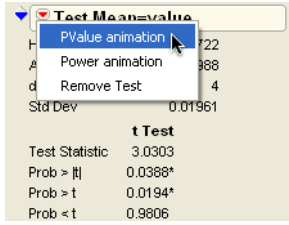
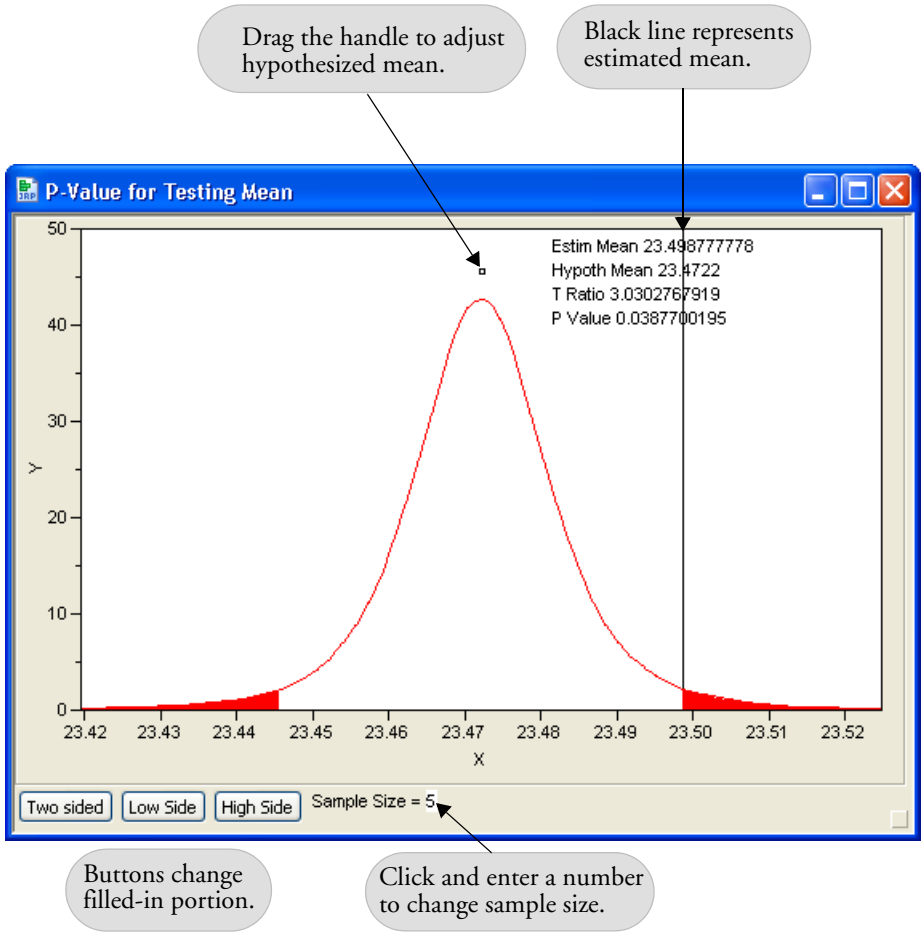


Figure 7.16 *p*-Value Animation Window for the Ecliptic Case Study



The black vertical line represents the mean estimated by the historical measurements. The handle can be dragged around the window with the mouse. In this case, the handle represents the true mean under the null hypothesis. To reject this true mean, there must be a significant difference between it and the mean estimated by the data.

The *p*-value calculated by JMP is affected by the difference between this true mean and the estimated mean, and you can see the effect of a different true mean by dragging the handle.

- Use the mouse to drag the handle left and right. Observe the changes in the *p*-value as the true mean changes.

As expected, the *p*-value decreases as the difference between the true and hypothesized mean increases.

The effect of changing this mean is also illustrated geometrically. As illustrated previously in **Figure 7.12**, the shaded area represents the region where the null hypothesis is rejected. As the area of this region increases, the *p*-value of the test also increases. This demonstrates that the closer your estimated mean is to the true mean under the null hypothesis, the less likely you are to reject the null hypothesis.

This demonstration can also be used to extract other information about the data. For example, you can determine the smallest difference that your data would be able to detect for specific *p*-values. To determine this difference for $p = 0.10$:

- Drag the handle until the *p*-value is as close to 0.10 as possible.

You can then read the estimated mean and hypothesized mean from the text display. The difference between these two numbers is the smallest difference that would be significant at the 0.10 level. Anything smaller would not be significant.

To see the difference between *p*-values for two and one sided tests, use the buttons at the bottom of the window.

- Press the **High Side** button to change the test to a one-sided *t*-test.

The *p*-value decreases because the region where the null hypothesis is rejected has become larger—it is all piled up on one side of the distribution, so smaller differences between the true mean and the estimated mean become significant.

- Repeatedly press the **Two Sided** and **High Side** buttons.

What is the relationship between the *p*-values when the test is one- and two-sided? To edit and see the effect of different sample sizes:

Click on the values for sample size beneath the plot.

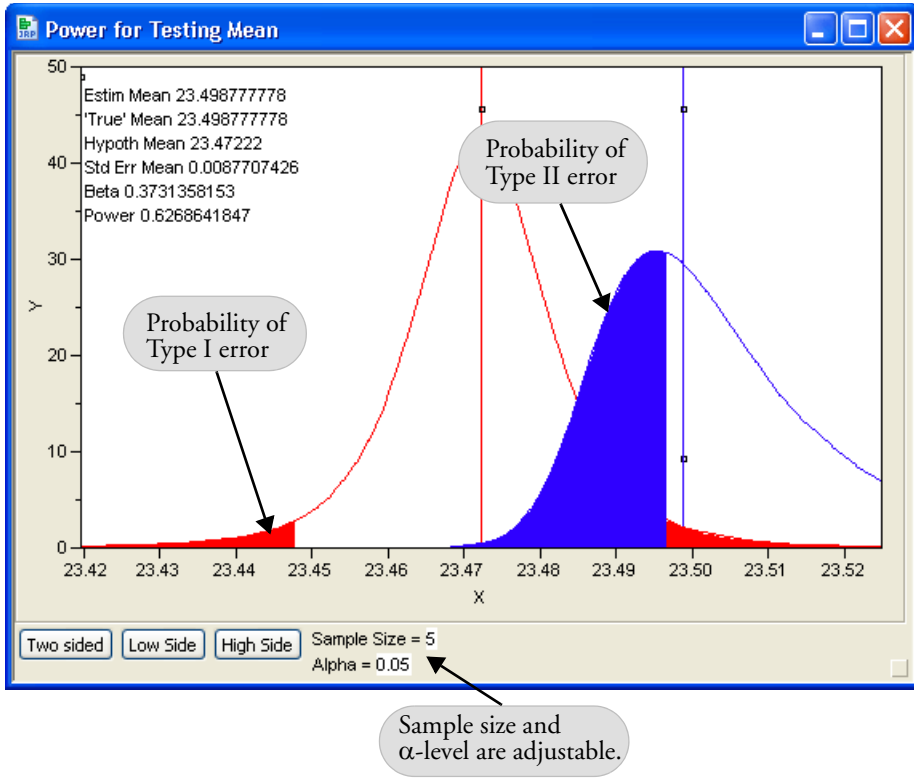
What effect would a larger sample size have on the *p*-value?

Power of the *t*-Test

As discussed in the section "Testing Hypotheses: Terminology" on page 138, there are two types of error that a statistician is concerned with when conducting a statistical test—Type I and Type II. JMP contains a built-in script to graphically demonstrate the quantities involved in computing the power of a *t*-test.

In the same menu where you found the **Pvalue animation**, select **Power animation** to display the window shown in **Figure 7.17**.

Figure 7.17 Power Animation Window



The probability of committing a Type I error (reject the null hypothesis when it is true), often represented by α, is shaded in red. The probability of committing a Type II error (not

detecting a difference when there is a difference), often represented as β, is shaded in blue. Power is 1 – β, which is the probability of detecting a difference. The case where the difference is zero is examined below.

There are three handles in this window, one each for the estimated mean (calculated from the data), the true mean (an unknowable quantity that the data estimates), and the hypothesized mean (the mean assumed under the null hypothesis). You can drag these handles to see how their positions affect power.

Note: Click on the values for sample size and alpha beneath the plot to edit them.

- Drag the 'True' mean until it coincides with the hypothesized mean.

This simulates the situation where the true mean is the hypothesized mean in a test where α=0.05. What is the power of the test?

- Continue dragging the 'True' mean around the graph.

Can you make the probability of committing a Type II error smaller than the case above, where the two means coincide?

- Drag the 'True' mean so that it is far away from the hypothesized mean.

Notice the shape of the blue distribution (around the 'True' mean) is no longer symmetrical. This is an example of a *non-central t-distribution*.

Finally, as with the *p*-value animation, these same situations can be further explored for one-sided tests using the buttons along the bottom of the window.

- Explore different values for sample size and alpha.

Practical Significance vs. Statistical Significance

This section demonstrates that a *statistically* significant difference can be quite different than a *practically* significant difference. Dr. Quick and Dr. Quack are both in the business of selling diets, and they have claims that appear contradictory. Dr. Quack studied 500 dieters and claims,

"A statistical analysis of my dieters shows a statistically significant weight loss for my Quack diet."

The Quick diet, by contrast, shows no significant weight loss by its dieters. Dr. Quick followed the progress of 20 dieters and claims,

"A statistical study shows that on average my dieters lose over three times as much weight on the Quick diet as on the Quack diet."

So which claim is right?

- To compare the Quick and Quack diets, open the Diet.jmp sample data table.

Figure 7.18 shows a partial listing of the Diet data table.

- Choose **Analyze > Distribution** and assign both variables to Y.
- Click **OK**.
- Select **Test Mean** from the popup menu at the top of each plot to compare the mean weight loss to zero.

You should use the one-sided *t*-test because you are only interested in significant weight loss (not gain).

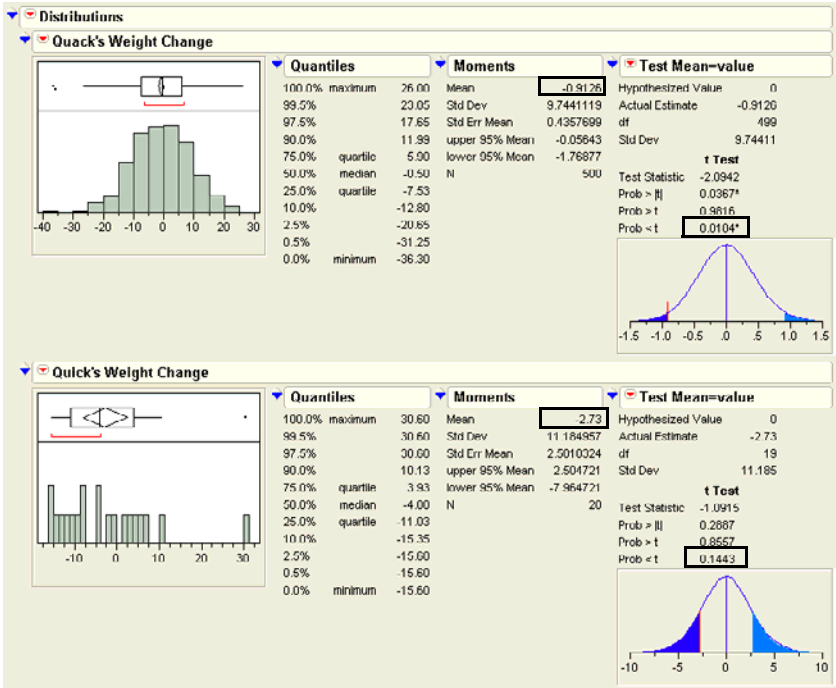
Figure 7.18 Diet Data

Diet
Notes Data demonstr
Distribution
Columns (2/0)
Quack's Weight Ch
Quick's Weight Ch
Rows
All rows 500

	Quack's Weight Change	Quick's Weight Change
1	1.8	-9.4
2	6.8	-4.2
3	1.1	10.5
4	2.6	-15.6
5	11.9	3.4
6	-4.7	-0.6
7	-11.8	-11.3
8	-0.4	30.6

If you look closely at the *t*-test results in **Figure 7.19**, you can verify both claims!

Figure 7.19 Reports of the Quick and Quack Example



Quick's average weight loss of 2.73 is over three times the 0.91 weight loss reported by Quack. However, Quick's larger mean weight loss was not significantly different from zero, and Quack's small weight loss was significantly different from zero. Quack might not have a better diet, but he has more evidence—500 cases compared with 20 cases. So even though the diet produced a weight loss of less than a pound, it is statistically significant. Significance is about evidence, and having a large sample size can make up for having a small effect.

Dr. Quick needs to collect more cases, and then he can easily dominate the Quack diet (though it seems like even a 2.7-pound loss may not be enough of a practical difference to a customer).

If you have a large enough sample size, even a very small difference can be significant. If your sample size is small, even a large difference may not be significant.

Looking closer at the claims, note that Quick reports on the estimated difference between the two diets, whereas Quack reports on the significance probabilities. Both are somewhat empty statements. It is not enough to report an estimate without a measure of variability. It is not enough to report a significance without an estimate of the difference.

The best report is a confidence interval for the estimate, which shows both the statistical and practical significance. The next chapter presents the tools to do a more complete analysis on data like the Quick and Quack diet data.

Examining for Normality

Sometimes you may want to test whether a set of values has a particular distribution. Perhaps you are verifying assumptions and want to test that the values have a Normal distribution.

Normal Quantile Plots

Normal quantile plots show all the values of the data as points in a plot. If the data are Normal, the points tend to follow a straight line.

- Return to the Randdist.jmp histograms.
- From the red triangle menu on the report title bar, select **Normal Quantile Plot** for each of the four distributions.

The histograms and Normal quantile plots for the four simulated distributions are shown in **Figure 7.21** to **Figure 7.24**.

The *y* (vertical) coordinate is the actual value of each data point. The *x* (horizontal) coordinate is the Normal quantile associated with the rank of the value after sorting the data.

If you are interested in the details, the precise formula used for the Normal quantile values is

$$\Phi^{-1}\left(\frac{r_i}{N+1}\right)$$

where r_i is the rank of the observation being scored, N is the number of observations, and Φ^{-1} is the function that returns the Normal quantile associated with the probability argument

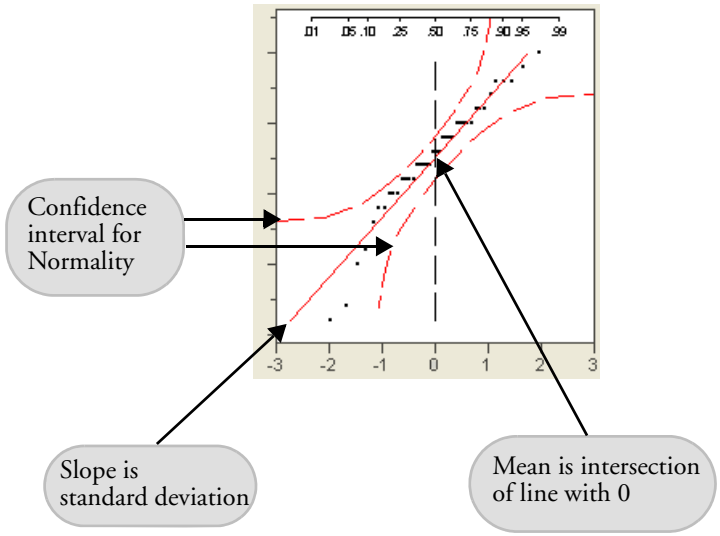
$$\frac{r_i}{N+1}$$

The Normal quantile is the value on the *x*-axis of the Normal density that has the portion *p* of the area below it, where *p* is the probability argument. For example, the quantile for 0.5 (the probability of being less than the median) is zero, because half (50%) of the density of the standard Normal is below zero. The technical name for the quantiles JMP uses is the *van der*

Waerden Normal scores; they are computationally cheap (but good) approximations to the more expensive, exact expected Normal order statistics.

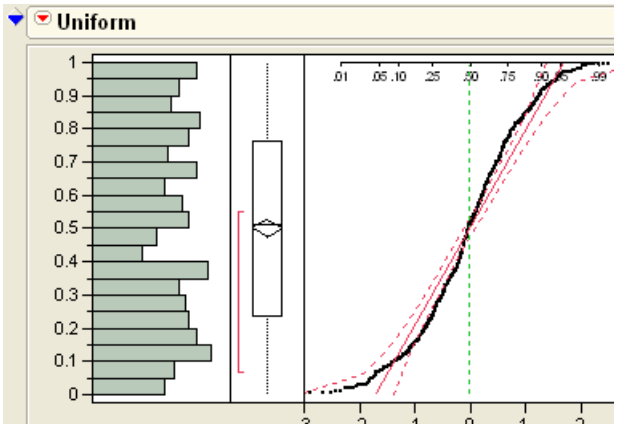
- A red straight line, with confidence limits, shows where the points would tend to lie if the data were Normal. This line is purely a function of the mean and standard deviation of the sample. The line crosses the mean of the data at the Normal quantile of 0. The slope of the line is the standard deviation of the data.
- Dashed lines surrounding the straight line form a confidence interval for the Normal distribution. If the points fall outside these dashed lines, you are seeing a significant departure from Normality.
- If the slope of the points is small (relative to the Normal) then you are crossing a lot of (ranked) data with very little variation in the real values, and therefore encounter a dense cluster. If the slope of the points is large, then you are crossing a lot of space that has only a few (ranked) points. Dense clusters make flat sections, and thinly populated regions make steep sections (see upcoming figures for examples of this).The overall slope is the standard deviation.

Figure 7.20 Normal Quantile Plot Explanation



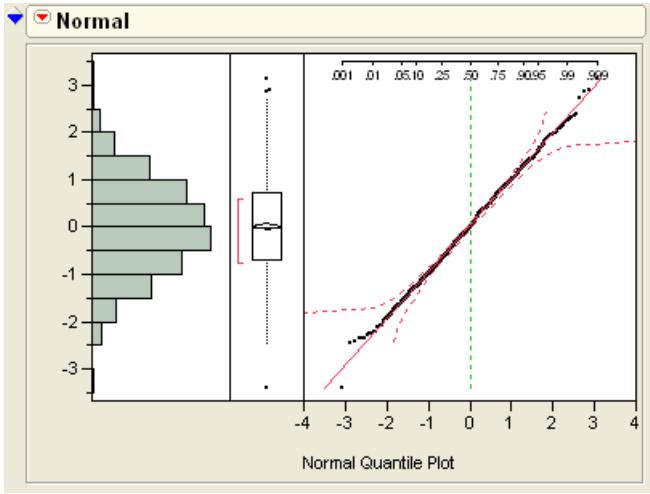
The middle portion of the uniform distribution (**Figure 7.21**) is steeper (less dense) than the Normal. In the tails, the uniform is flatter (more dense) than the Normal. In fact, the tails are truncated at the end of the range, where the Normal tails extend infinitely.

Figure 7.21 Uniform Distribution



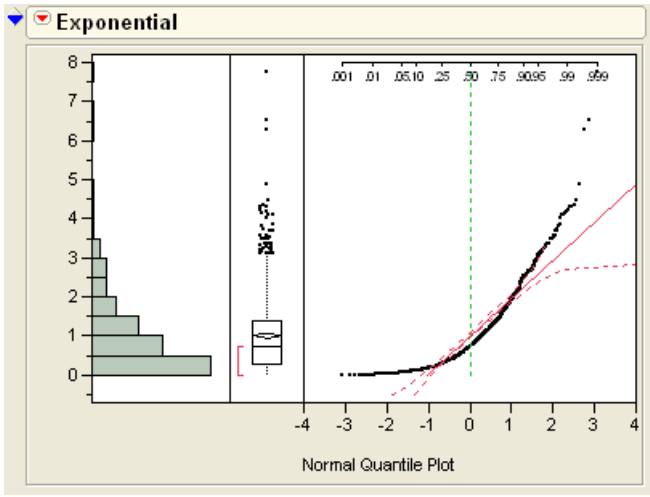
The Normal distribution (**Figure 7.22**) has a Normal quantile plot that follows a straight line. Points at the tails usually have the highest variance and are most likely to fall farther from the line. Because of this, the confidence limits flair near the ends.

Figure 7.22 Normal Distribution



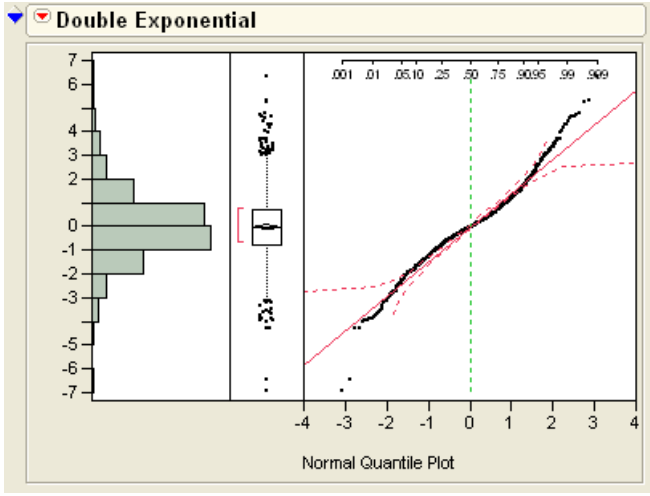
The exponential distribution (**Figure 7.23**) is skewed—that is, one-sided. The top tail runs steeply past the Normal line; it is spread out more than the Normal. The bottom tail is shallow and much denser than the Normal.

Figure 7.23 Exponential Distribution



The middle portion of the double exponential (**Figure 7.24**) is denser (more shallow) than the Normal. In the tails, the double exponential spreads out more (is steeper) than the Normal.

Figure 7.24 Double Exponential Distribution



Statistical Tests for Normality

A widely used test that the data are from a specific distribution is the *Kolmogorov test* (also called the *Kolmogorov-Smirnov test*). The test statistic is the greatest absolute difference

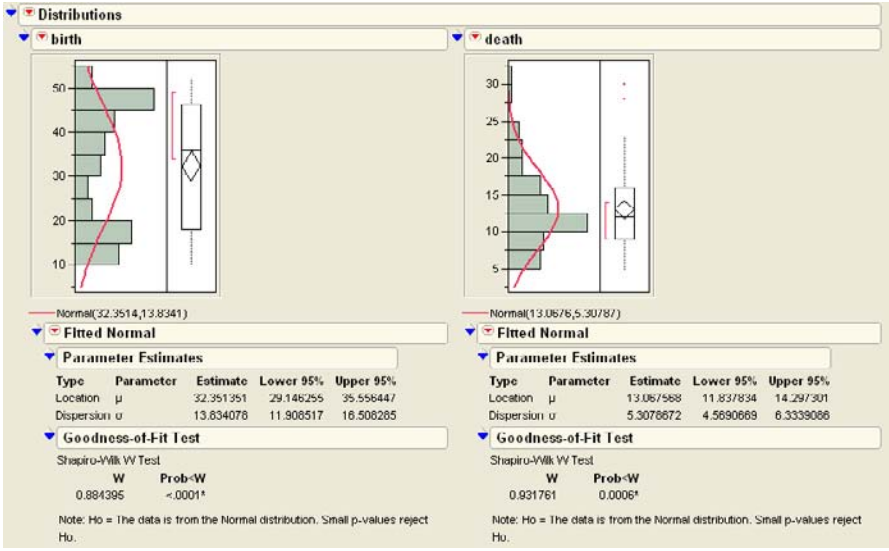
between the hypothesized distribution function and the empirical distribution function of the data. The empirical distribution function goes from 0 to 1 in steps of $1/n$ as it crosses data values. When the Kolmogorov test is applied to the Normal distribution and adapted to use estimates for the mean and standard deviation, it is called the *Lilliefor's test* or the *KSL test*. In JMP, the Lilliefor's quantiles on the cumulative distribution function (cdf) are translated into confidence limits in the Normal quantile plot, so that you can see where the distribution departs from Normality by where it crosses the confidence curves.

Another test of Normality produced by JMP is the *Shapiro-Wilk test* (or the *W-statistic*), which is implemented for samples as large as 2000. The null hypothesis for this test is that the data are normal. Rejecting this hypothesis would imply the distribution is non-normal.

- Look at the Birth Death.jmp data table again or re-open it if it is closed.
- Choose **Analyze > Distribution** for the variables birth and death. Click **OK**.
- Select **Fit Distribution > Normal** from the red triangle menu on the birth report title bar.
- Select **Goodness of Fit** from the popup menu next to the Fitted Normal report.
- Repeat for the death distribution.

Its results are shown in **Figure 7.25**.

Figure 7.25 Test Distributions for Normality



The conclusion is that neither distribution is Normal, although the second distribution is much closer than the first.

This is an example of an unusual situation where you hope the test fails to be significant, because the null hypothesis is that the data are Normal.

If you have a large number of observations, you may want to reconsider this tactic. The Normality tests are sensitive to small departures from Normality, and small departures do not jeopardize other analyses because of the Central Limit Theorem, especially because they will also probably be highly significant. All the distributional tests assume that the data are independent and identically distributed.

Some researchers test the Normality of residuals from model fits, because the other tests assume a Normal distribution. We strongly recommend that you do not conduct these tests, but instead rely on normal quantile plots to look for patterns and outliers.

So far we have been doing correct statistics, but a few remarks are in order.

- In most tests, the null hypothesis is something you want to disprove. It is disproven by the contradiction of getting a statistic that would be unlikely if the hypothesis were true. But in Normality tests, you want the null hypothesis to be true. Most testing for Normality is to verify assumptions for other statistical tests.
- The mechanics for any test where the null hypothesis is desirable are backwards. You can get an undesirable result, but the failure to get it does not prove the opposite—it only says that you have insufficient evidence to prove it is true. “Special Topic: Practical Difference” on page 158 gives more details on this issue.
- When testing for Normality, it is more likely to get a desirable (inconclusive) result if you have very little data. Conversely, if you have thousands of observations, almost any set of data from the real world appears significantly non-Normal.
- If you have a large sample, the estimate of the mean will be distributed Normally even if the original data is not. This result, from the Central Limit Theorem, is demonstrated in a later section beginning on page 160.
- The test statistic itself doesn't tell you about the nature of the difference from Normality. The Normal quantile plot is better for this. Residuals from regressions (“Examine Residuals” on page 246) can have both these problems.

Special Topic: Practical Difference

Suppose you really want to show that the mean of a process is a certain value. Standard statistical tests are of no help, because the failure of a test to show that a mean is *different* from the hypothetical value does not show that it *is* that value. It only says that there is not enough evidence to confirm that it isn't that value. In other words, saying "I can't say the result is different from 5" is not the same as saying "The result must be 5."

You can never show that a mean is exactly some hypothesized value, because the mean could be different from that hypothesized value by an infinitesimal amount. No matter what sample size you have, there is a value that is different from the hypothesized mean by an amount that is so small that it is quite unlikely to get a significant difference even if the true difference is zero.

So instead of trying to show that the mean is exactly equal to an hypothesized value, you need to choose an interval around that hypothesized value and try to show that the mean is not outside that interval. This can be done.

There are many situations where you want to control a mean within some specification interval. For example, suppose that you make 20 amp electrical circuit breakers. You need to demonstrate that the mean breaking current for the population of breakers is between 19.9 and 20.1 amps. (Actually, you probably also require that most individual units be in some specification interval, but for now we just focus on the mean.) You'll never be able to prove that the mean of the population of breakers is exactly 20 amps. You can, however, show that the mean is close—within 0.1 of 20.

The standard way to do this is *TOST method*, an acronym for Two One-Sided Tests [Westlake(1981), Schuirmann(1981), Berger and Hsu (1996)]:

1. First you do a one-sided *t*-test that the mean is the low value of the interval, with an upper tail alternative.
2. Then you do a one-sided *t*-test that the mean is the high value of the interval, with a lower tail alternative.
3. If both tests are significant at some level α, then you can conclude that the mean is outside the interval with probability less than or equal to α, the significance level. In other words, the mean is not significantly practically different from the hypothesized value, or, in still other words, the mean is practically equivalent to the hypothesized value.

Note: Technically, the test works by a union intersection rule, whose description is beyond the scope of this book.

For example,

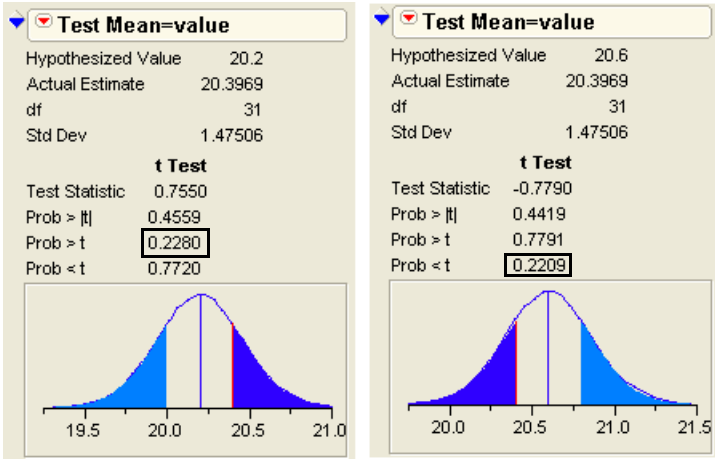
- Open the Coating.jmp sample data table, found in the Quality Control subfolder.
- Select **Analyze > Distribution** and assign weight to the **Y, Columns** role.
- Click **OK**.

When the report appears,

- Select **Test Mean** from the platform drop-down menu and enter 20.2 as the hypothesized value.
- Click **OK**.
- Select **Test Mean** again and enter 20.6 as the hypothesized value.
- Click **OK**.

This tests the null hypothesis that the mean weight is between 20.2 and 20.6 (that is, 20.4±0.2) with a protection level (α) of 0.05.

Figure 7.26 Compare Test for Mean at Two Values



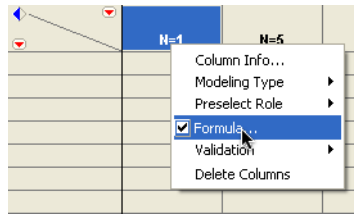
The *p*-value for the hypothesis from below is 0.22, and the *p*-value for the hypothesis from above is also 0.22. Since both of these values are far above the α of 0.05 that we were looking for, we declare it not significant. We cannot reject the null hypothesis. The conclusion is that we have not shown that the mean is practically equivalent to 20.4 ± 0.2 at the 0.05 significance level. We need more data.

Special Topic: Simulating the Central Limit Theorem

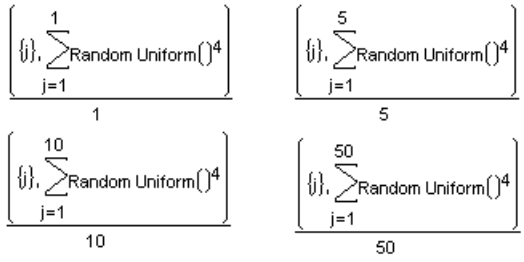
The Central Limit Theorem says that for a very large sample size, the sample mean is very close to Normally distributed, regardless of the shape of the underlying distribution. That is, if you compute means from many samples of a given size, the distribution of those means approaches Normality, even if the underlying population from which the samples were drawn is not.

You can see the Central Limit Theorem in action using the template called Cntrlmt.JMP.

- Open Central Limit Theorem.JMP.
- Right-click (Control-click on the Macintosh) in the column heading for the N=1 column and select **Formula** from the menu that appears.
- Do the same thing for the rest of the columns, called N=5, N=10, and so on, to look at their formulas.



Looking at the formulas (shown to the right) may help you understand what's going on. The expression raising the uniform random number values to the 4th power creates a highly skewed distribution. For each row, the first column, N=1, generates a single uniform random number to the fourth power. For each row in the second column, N=5, the formula generates a sample of five uniform numbers, takes each to the fourth power, and computes the mean. The next column does the same for a sample size of 10, and the remaining columns generate means for sample sizes of 50 and 100.



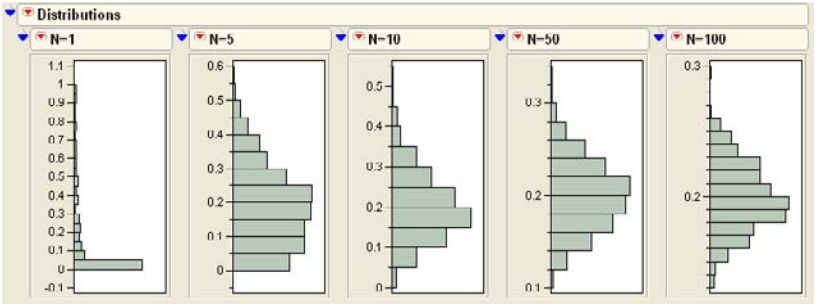
- Add 500 rows to the data table using **Rows > Add Rows**.

When the computations are complete:

- Choose **Analyze > Distribution**. Select all the variables, assign them as Y variables and click **OK**.

You should see the results in **Figure 7.27**. When the sample size is only 1, the skewed distribution is apparent. As the sample size increases, you can clearly see the distributions becoming more and more Normal.

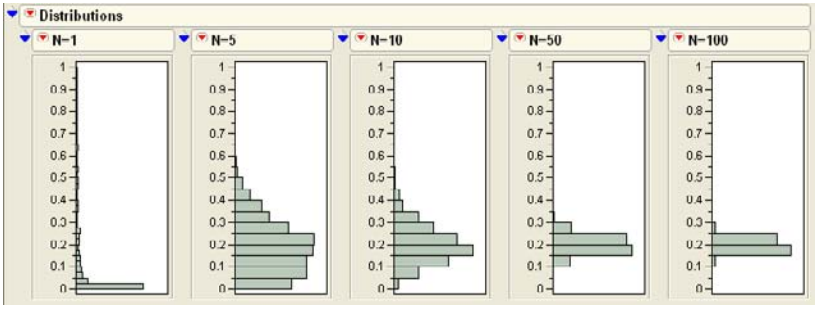
Figure 7.27 Example of the Central Limit Theorem in Action



The distributions also become less spread out, since the standard deviation (s) of a mean of n items is $\frac{s}{\sqrt{n}}$.

To see this, select the **Uniform Scaling** option from the red triangle menu on the Distribution title bar.

Figure 7.28 Distributions with Uniform Scales



Seeing Kernel Density Estimates

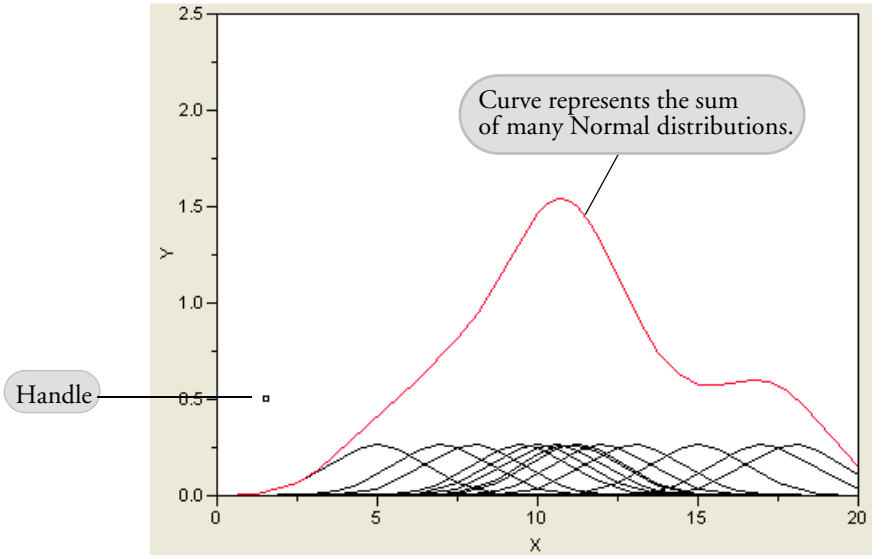
The idea behind kernel density estimators is not difficult. In essence, a Normal distribution is placed over each data point with a specified standard deviation. Each of these Normal distributions is then summed to produce the overall curve.

JMP can animate this process for a simple set of data. For details on using scripts, see “Working with Scripts” on page 57.

- Open the demoKernel.jsl script, and make sure the check box for **Run this script after opening** is checked.

You should see a window like the one in **Figure 7.29**.

Figure 7.29 Kernel Addition Demonstration



The handle on the left side of the graph can be dragged with the mouse.

- Move the handle to adjust the spread of the individual Normal distributions associated with each data point.

The larger red curve is the smoothing spline generated by the sum of the Normal distributions. As you can see, merely adjusting the spread of the small Normal distributions dictates the smoothness of the spline fit.

Exercises

1. The file movies.jmp contains a list of the top grossing movies of all time (as of June 2003). It contains data showing the name of a movie, the amount of money it made in the United States (Domestic) and in foreign markets (in millions of dollars), its year of release, and the type of movie.

(a) Create a histogram of the types of movies in the data. What are the levels of this variable? How many of each level are in the data set?

(b) Create a histogram of the domestic gross for each movie. What is the range of values for this variable? What is the average domestic gross of these movies?

(c) Consider the histogram you created in part (b) for the domestic gross of these movies. You should notice several outliers in the outlier box plot. Based on your experience, guess what these movies are. Then, move the pointer over each outlier to reveal its name. Were you correct?

(d) Create a subset of the data consisting of only drama movies. Create a histogram and find the average domestic and foreign grosses for your subset. Are there outliers in either variable?

2. The file Analgesics.jmp contains pain ratings from patients after treatments from three different pain relievers. The patients are labeled only by gender in this study. The study was meant to determine if the three pain relievers were different in the amount of pain relief the patients experienced.

(a) Create a histogram of the variables gender, drug, and pain. Click on the histogram bars to determine if the distribution of gender is roughly equal among the three analgesics.

(b) Create a separate histogram for the variable pain for each of the three different analgesics (Hint: Use the **By** button). Does the mean pain response seem the same for each of the three analgesics?

3. The file Scores.jmp contains data for the United States from the Third International Mathematics and Science Study, conducted in 1995. The variables came from testing over 5000 students for their abilities in Calculus and Physics, and are separated into four regions of the United States. Note that some students took the Calculus test, some took the Physics test, and some took both. Assume that the scores represent a random sample for each of the four regions of the United States.

(a) Produce a histogram and find the mean scores for the United States on both tests. By clicking on the bars of the histogram, can you determine whether a high calculus score correlates highly with a high Physics score?

(b) Find the mean scores for the Calculus test for the four regions of the country. Do they appear to be roughly equal?

(c) Find the mean scores for the Physics tests for the four regions of the country. Do they appear to be roughly equal?

(d) Suppose that from an equivalent former test, the mean score of United States Calculus students was 450. Does this study show evidence that the score has increased since the last test?

(e) Construct a 95% confidence interval for the mean calculus score.

(f) Suppose that Physics teachers say that the overall United States score on the Physics test should be higher than 420. Do the data support their claim?

(g) Construct a 95% confidence interval for the mean Physics score.

4. The file Cereal.jmp contains nutritional information for 76 kinds of cereal.

   (a) Find the mean number of fat grams for the cereals in this data set. List any unusual observations.

   (b) Use the Distribution platform to find the two kinds of cereal with unusually high fiber content.

   (c) The hot/cold variable is used to specify whether the cereal was meant to be eaten hot or cold. Find the mean amount of sugars contained in the hot cereals and the cold cereals. Construct a 95% confidence interval for each.

5. Various crime statistics for each of the 50 states in the United States are stored in the file Crime.jmp.

   (a) Examine the distributions of each statistic. Which (if any) do not appear to follow a Normal distribution?

   (b) Which two states are outliers with respect to the robbery variable?

6. Data for the Brigham Young football team are stored in the Football.jmp data file.

   (a) Find the average height and weight of the players on the team.

   (b) The Position variable identifies the primary position of each player. Which position has the smallest average weight? Which has the highest?

   (c) What position has the largest average neck measurements? What position (on average) can bench press the most weight?

7. The Hot Dogs.jmp data file came from an investigation of the taste and nutritional content of hot dogs.

   (a) Construct a histogram of the type of hot dogs (beef, meat, and poultry). Are there an equal number of each type considered?

(b) The $/oz variable represents the cost per ounce of hot dog. Construct an outlier box plot of this variable and find any outliers.

(c) Construct a 95% confidence interval for the caloric content of the three types of hot dogs. Which type gives (on average) the lowest calories?

(d) Test the conjecture that the mean sodium content of all hot dogs is 410 grams.

8. Three brands of typewriters were tested for typing speed by having expert typists type identical passages of text. The results are stored in Typing Data.jmp.

(a) Are the data for typing speeds Normally distributed?

(b) What is the mean typing speed for all typewriters?

(c) Find a 95% confidence interval on typing speed for each of the three typewriter types.

8 The Difference between Two Means

Overview

Are the mean responses from two groups different? What evidence would it take to convince you? This question opens the door to many of the issues that pervade statistical inference, and this chapter explores these issues. Comparing group means also introduces an important statistical distinction regarding how the measurement or sampling process affects the way the resulting data are analyzed. This chapter also talks about validating statistical assumptions.

When two groups are considered, there are two distinct situations that lead to two different analyses:

Independent Groups—the responses from the two groups are unrelated and statistically independent. For example, the two groups might be two classrooms with two sets of students in them. The responses come from different experimental units or subjects. The responses are uncorrelated and the means from the two groups are uncorrelated.

Matched Pairs—the two responses form a pair of measurements coming from the same experimental unit or subject. For example, a matched pair might be a before-and-after blood pressure measurement from the same subject. These responses are correlated, and the statistical method must take that into account.

Two Independent Groups

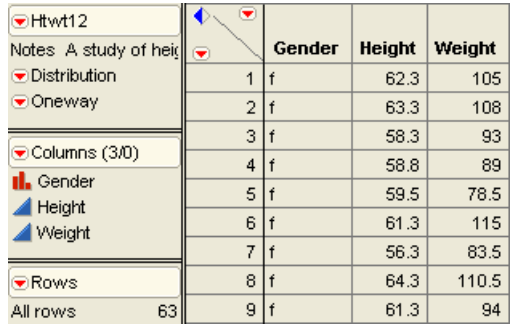
For two different groups, the goal might be to estimate the group means and determine if they are significantly different. Along the way, it is certainly advantageous to notice anything else of interest about the data.

When the Difference Isn't Significant

A study compiled height measurements from 63 children, all age 12. It's safe to say that as they get older, the mean height for males will be greater than for females, but is this the case at age 12? Let's find out:

- Open Htwt12.jmp to see the data shown (partially) below.

There are 63 rows and three columns. This example uses Gender and Height. Gender has the Nominal modeling type, with codes for the two categories, "f" and "m". Gender will be the X variable for the analysis. Height contains the response of interest, and so will be the Y variable.



	Gender	Height	Weight
1	f	62.3	105
2	f	63.3	108
3	f	58.3	93
4	f	58.8	89
5	f	59.5	78.5
6	f	61.3	115
7	f	56.3	83.5
8	f	64.3	110.5
9	f	61.3	94

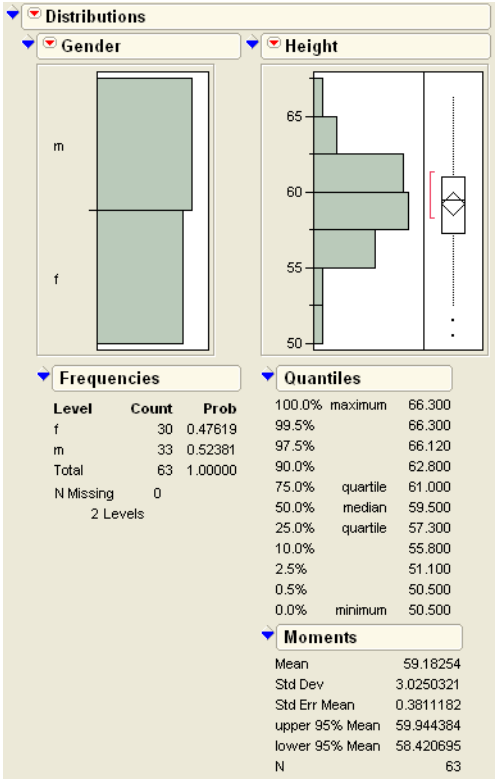
Check the Data

To check the data, first look at the distributions of both variables graphically with histograms and box plots.

- Choose **Analyze > Distribution** from the menu bar.
- In the launch dialog, select Gender and Height as Y variables.
- Click **OK** to see an analysis window like the one shown in **Figure 8.1**.

Every pilot walks around the plane looking for damage or other problems before starting up. No one would submit an analysis to the FDA without making sure that the data were not confused with data from another study. Do your kids use the same computer that you do? Then check your data. Does your data set have so many decimals of precision that it looks like it came from a random number generator? Great detectives let no clue go unnoticed. Great data analysts check their data carefully.

Figure 8.1 Histograms and Summary Tables



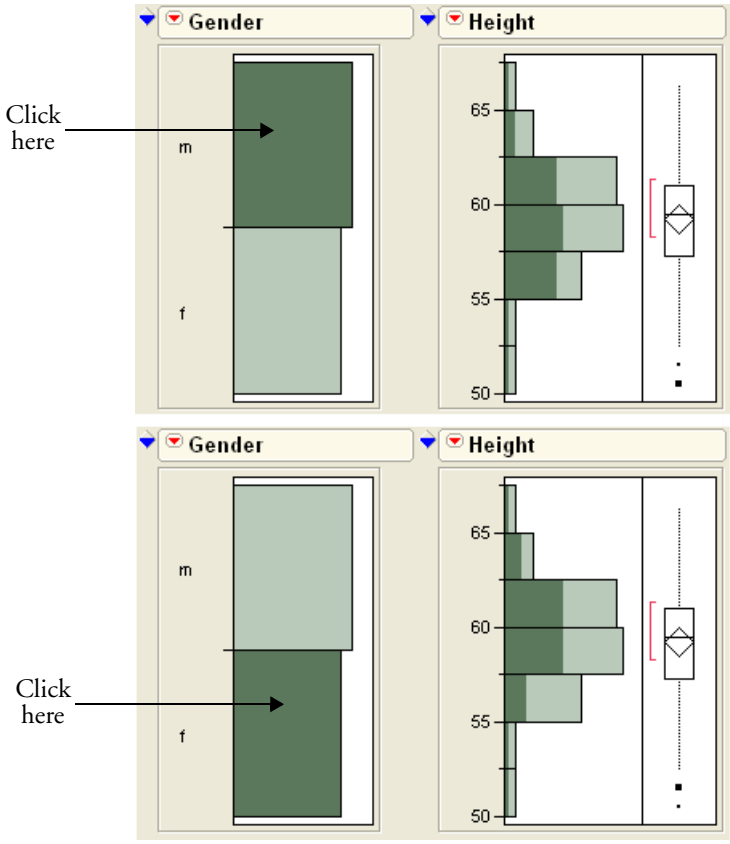
A look at the histograms for Gender and Height reveals that there are a few more males than females. The overall mean height is about 59, and there are no missing values (N is 63, and there are 63 rows in the table). The box plot indicates that two of the children seem unusually short compared to the rest of the data.

- Move the cursor to the Gender histogram, and click on the bar for "f".

Clicking the bar highlights the females in the data table and also highlights the females in the Height histogram (See **Figure 8.2**). Now click on the "m" bar, which highlights the males and un-highlights the females.

By alternately clicking on the bars for males and females, you can see the conditional distributions of each subset highlighted in the Height histogram. This gives a preliminary look at the height distribution within each group, and it is these group means we want to compare.

Figure 8.2 Interactive Histogram



Launch the Fit Y by X Platform

We know to use the Fit Y by X platform because our context is comparing two variables.

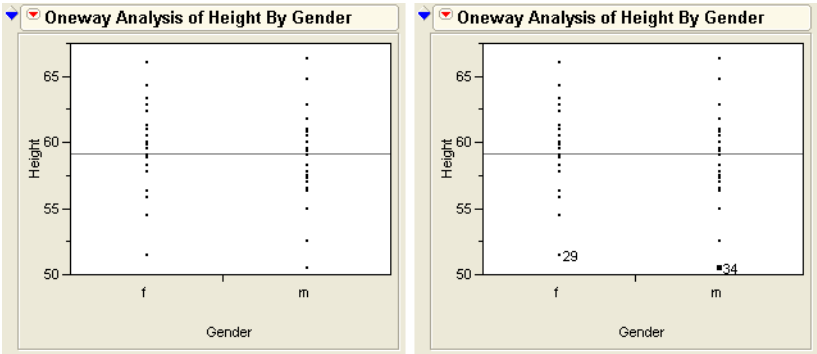
You can compare group means by assigning Height as the continuous Y variable and Gender as the nominal (grouping) X variable. Begin by launching the analysis platform:

- Choose **Analyze > Fit Y by X**.
- In the launch dialog, select Height as Y and Gender as X.

Notice that the role-prompting dialog indicates that you are doing a one-way analysis of variance (ANOVA). Because Height is continuous and Gender is categorical (nominal), the **Fit Y by X** command automatically gives a one-way layout for comparing distributions.

- Click **OK** to see the initial graphs, which are side-by-side vertical dot plots for each group (see the left picture in **Figure 8.3)**.

Figure 8.3 Plot of the Responses, Before and After Labeling Points



Examine the Plot

The horizontal line across the middle shows the overall mean of all the observations. To identify possible outliers (students with unusual values):

- Click the lowest point in the "f" vertical scatter and Shift-click in the lowest point in the "m" sample.

Shift-clicking extends a selection so that the first selection does not un-highlight.

- Choose **Rows > Label/Unlabel** to see the plot on the right in **Figure 8.2**.

Now the points are labeled 29 and 34, the row numbers corresponding to each data point.

Display and Compare the Means

The next step is to display the group means in the graph, and to obtain an analysis of them.

- Select **t test** from the red triangle popup menu that shows on the plot's title bar.
- From the same menu, select **Means/Anova/Pooled t**.

This adds analyses that estimates the group means and tests to see if they are different.

Note: Normally, you don't select both versions of the *t*-test, shown in **Figure 8.5**. We're selecting these for illustration. To determine the correct test for other situations, see "Equal or Unequal Variances?" on page 174.

The **Means/Anova/Pooled t** option automatically displays the *means diamonds* as shown in **Figure 8.4**, with summary tables and statistical test reports.

The center lines of the means diamonds are the group means. The top and bottom of the diamonds form the 95% confidence intervals for the means. You can say the probability is 0.95 that this confidence interval contains the true group mean.

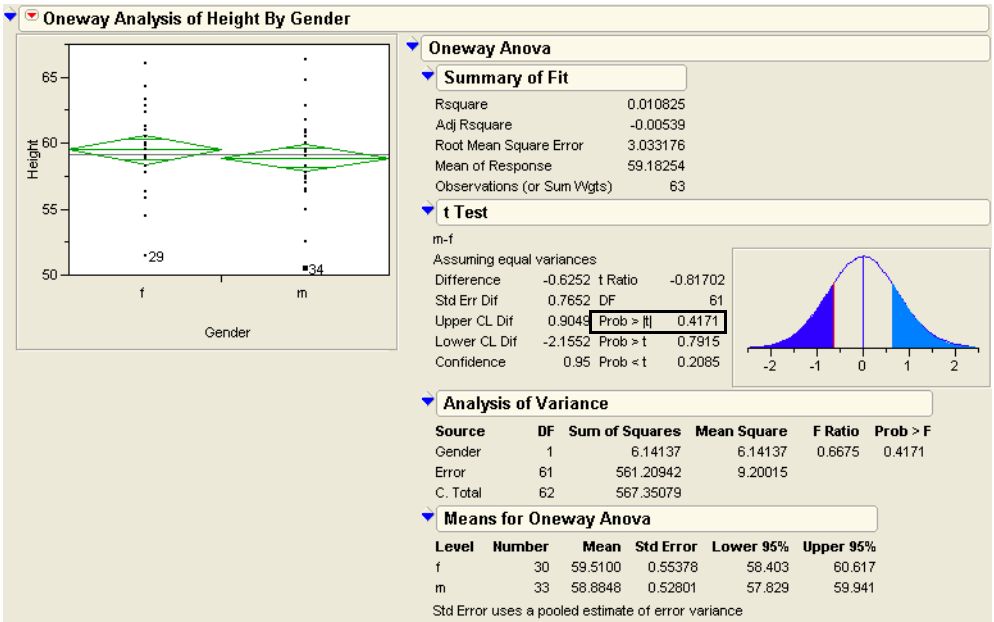
The confidence intervals show whether a mean is significantly different from some hypothesized value, but what can it show regarding whether two means are significantly different? Use the following rule.

Interpretation Rule for Means Diamonds: If the confidence intervals shown by the means diamonds do not overlap, the groups are significantly different (but the reverse is not necessarily true).

It is clear that the means diamonds in this example overlap. Therefore, you need to take a closer look at the text report beneath the plots to determine if the means are really different. The report, shown in **Figure 8.4**, includes summary statistics, *t*-test reports, an analysis of variance, and means estimates.

Note that the *p*-value of the *t*-test (shown with the label **Prob>|t|** in the **t test** section of the report) table is not significant.

Figure 8.4 Diamonds to Compare Group Means and Pooled *t* Report



Inside the Student's *t*-Test

The Student's *t*-test appeared in the last chapter to test whether a mean was significantly different from a hypothesized value. Now the situation is to test whether the difference of two means is significantly different from the hypothesized value of zero. The *t*-ratio is formed by first finding the difference between the estimate and the hypothesized value, and then dividing that quantity by its standard error.

$$t \text{ statistic} = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error of the estimate}}$$

In the current case, the estimate is the difference in the means for the two groups, and the hypothesized value is zero.

$$t \text{ statistic} = \frac{(\text{mean 1} - \text{mean 2}) - 0}{\text{standard error of the difference}}$$

For the means of two independent groups, the standard error of the difference is the square root of the sum of squares of the standard errors of the means.

$$\text{standard error of the difference} = \sqrt{s^2_{\text{mean 1}} + s^2_{\text{mean 2}}}$$

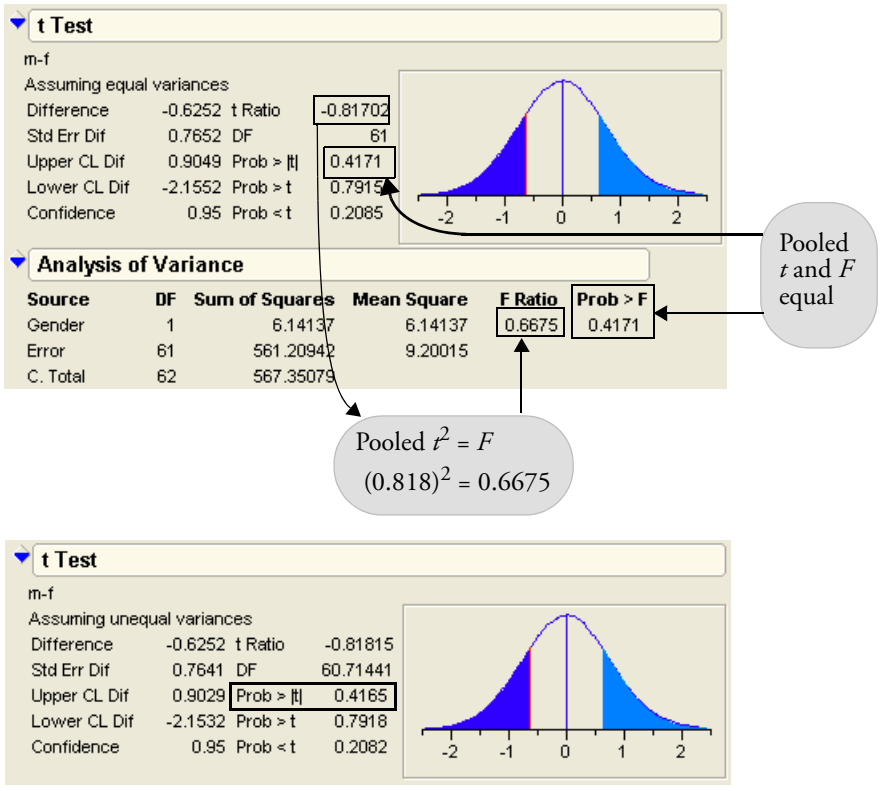
JMP does the standard error calculations and forms the tables shown in **Figure 8.4**. Roughly, you look for a *t*-statistic greater than 2 in absolute value to get significance at the 0.05 level. The significance level (the *p*-value) is determined in part by the degrees of freedom (DF) of the *t*-distribution. For this case, DF is the number of observations (63) minus two, because two means are being estimated. With the calculated *t* (0.818) and DF, the *p*-value is determined to be 0.4165. The label Prob>|t| is given to the *p*-value in the test table to indicate that this is the probability of getting an even greater absolute *t* statistic. This is the significance level. Usually a *p*-value less than 0.05 is regarded as significant.

In this example, the *p*-value of 0.4165 isn't small enough to detect a significant difference in the means. Is this to say that the means are the same? Not at all. You just don't have enough evidence to show that they are different. If you collect more data, you might be able to show a significant, albeit small, difference.

Equal or Unequal Variances?

The report shown in **Figure 8.5** contains two *t*-test reports. The uppermost of the two is labeled **Assuming equal variances**, and is generated with the **Means/Anova/Pooled t** command. The other is labeled **Assuming unequal variances**, and is generated with the **t test** command. Which is the correct report to use?

Figure 8.5 *t*-Test and ANOVA Reports



In general, the unequal-variance *t*-test (also known as the *unpooled t*-test) is the preferred test. This is because the unpooled version is quite sensitive (the opposite of robust) to departures from the equal-variance assumption (especially if the number of observations in the two groups is not the same), and often we cannot assume the variances of the two groups are equal. In addition, if the two variances are unequal, the unpooled test maintains the prescribed α-level and retains good power. For example, you may think you are conducting a test with $\alpha = 0.05$, but it may in fact be 0.10 or 0.20. What you think is a 95% confidence interval may be, in reality, an 80% confidence interval (Cryer and Wittmer, 1999). For these reasons, we recommend the unpooled (**t test** command) *t*-test for most situations. In this case, both *t*-tests are not significant.

However, the equal-variance variation is included for several reasons.

- For situations with very small sample sizes (for example, having three or fewer observations in each group), the individual variances cannot be estimated very well, but

the pooled versions can be, giving better power. In these circumstances, the pooled version has slightly enough power.

- Pooling the variances is the only option when there are more than two groups, when the *F*-test must be used. Therefore, the pooled *t*-test is a useful analogy for learning the analysis of the more general, multi-group situation. This situation is covered in the next chapter, "Comparing Many Means: One-Way Analysis of Variance" on page 209.

Rule for *t*-tests: Unless you have very small sample sizes, or a specific *a priori* reason for assuming the variances are equal, use the *t*-test produced by the **t test** command. When in doubt, use the **t test** command (*i.e.* unpooled) version.

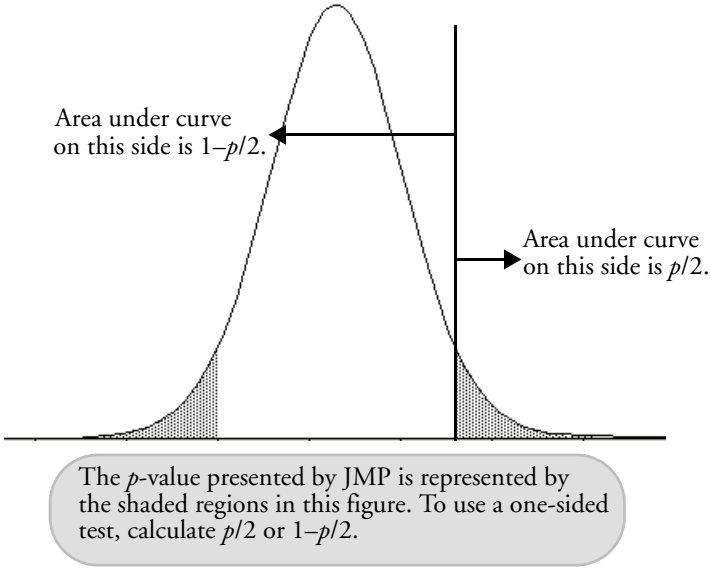
One-Sided Version of the Test

The Student's *t*-test in the previous example is for a two-sided alternative. In that situation, the difference could go either way (that is, either group could be taller), so a two-sided test is appropriate. The one-sided *p*-values are shown on the report, but you can get them by doing a a little arithmetic on the reported two-sided *p*-value, forming one-sided *p*-values by using

$$\frac{p}{2} \text{ or } 1 - \frac{p}{2},$$

depending on the direction of the alternative.

Figure 8.6 Two- and One-sided *t*-Test



In this example, the mean for males was less than the mean for females, so the significance with the alternative to conclude the females as higher is the *p*-value 0.2083, half the two-tailed *p*-value. Testing the other direction, the *p*-value is 0.7917. Both of these values are reported in **Figure 8.5** as Prob < t and Prob > t, respectively.

Analysis of Variance and the All-Purpose *F*-Test

As well as showing the *t*-test, the report in **Figure 8.5** for comparing two groups shows an analysis of variance with its *F*-test. The *F*-test surfaces many times in the next few chapters, so an introduction is in order. Details will unfold later.

The *F-test* compares variance estimates for two situations, one a special case of the other. Not only is this useful for testing means, but other things, as well. Furthermore, when there are only two groups, the *F*-test is equivalent to the pooled (equal variance) *t*-test, and the *F*-ratio is the square of the *t*-ratio: $(0.81)^2 = 0.66$, as you can see in **Figure 8.5**.

To begin, look at the different estimates of variance as reported in the Analysis of Variance table.

First, the analysis of variance procedure pools all responses into one big population and estimates the population mean (the *grand mean*). The variance around that grand mean is

estimated by taking the average sum of squared differences of each point from the grand mean.

The difference between a response value and an estimate such as the mean is called a *residual*, or sometimes the *error.*

What happens when a separate mean is computed for each group instead of the grand mean for all groups? The variance around these individual means is calculated, and this is shown in the Error line in the Analysis of Variance table. The Mean Square for Error is the estimate of this variance, called *residual variance* (also called s^2), and its square root, called the *root mean squared error* (or s), is the residual standard deviation estimate.

If the true group means are different, then the separate means give a better fit than the one grand mean. In other words, there will be less variance using the separate means than when using the grand mean. The change in the residual sum of squares from the single-mean model to the separate-means model leads us to the *F*-test shown in the Model line of the Analysis of Variance table. If the hypothesis that the means are the same is true, the Mean Square for Model also estimates the residual variance.

The *F*-ratio is the Model Mean Square divided by the Error Mean Square:

$$F\text{-Ratio} = \frac{\text{Mean Square for the Model}}{\text{Mean Square for the Error}} = \frac{6.141}{9.200} = 0.6675$$

The *F*-ratio is a measure of improvement in fit when separate means are considered. If there is no difference between fitting the grand mean and individual means, then both numerator and denominator estimate the same variance (the grand mean residual variance), so the *F*-ratio is around 1. However, if the separate-means model does fit better, the numerator (the model mean square) contains more than just the grand mean residual variance, and the value of the *F*-test increases.

If the two mean squares in the *F*-ratio are statistically independent (and they are in this kind of analysis), then you can use the *F*-distribution associated with the *F*-ratio to get a *p*-value. This tells how likely you are to see the *F*-ratio given by the analysis if there really was no difference in the means.

If the tail probability (*p*-value) associated with the *F*-ratio in the *F*-distribution is smaller than 0.05 (or the α-level of your choice), you can conclude that the variance estimates are different, and thus that the means are different.

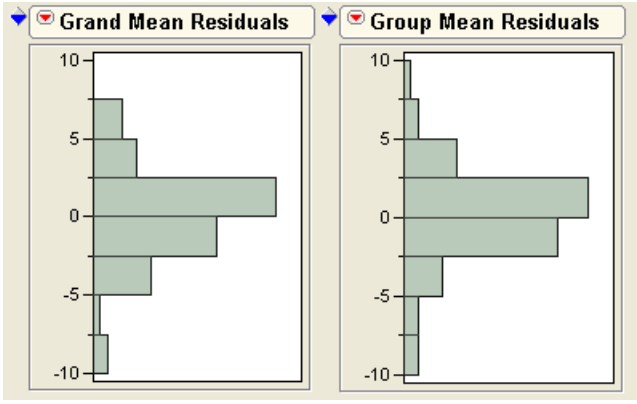
In this example, the total mean square and the error mean square are not much different. In fact, the *F*-ratio is actually less than one, and the *p*-value of 0.4171 (exactly the same as seen for the pooled *t*-test) is far from significant (much greater that 0.05).

The *F*-test can be viewed as whether the variance around the group means (the histogram on the right in **Figure 8.7**) is significantly less than the variance around the grand mean (the histogram on the left). In this case, the variance isn't much different.

In this way, a test of variances is also a test on means. The *F*-test turns up again and again because it is oriented to comparing the variation around two models. Most statistical tests can be constituted this way.

Figure 8.7 shows histograms that compare residuals from a group means model and a grand mean model. If the effect were significant, the variation showing on the left would have been much less than that on the right.

Figure 8.7 Residuals for Group Means Model (left) and Grand Mean Model (right)



Terminology for Sums of Squares: All disciplines that use statistics use analysis of variance in some form. However, you may find different names used for its components. For example, the following are different names for the same kinds of sums of squares (SS):

SS(model) = SS(regression) = SS (between)

SS(error) = SS(residual) = SS(within)

How Sensitive Is the Test?
How Many More Observations Are Needed?

So far, in this example, there is no conclusion to report because the analysis failed to show anything. This is an uncomfortable state of affairs. It is tempting to state that we have shown no significant difference, but in statistics this is the same as saying inconclusive. Our conclusions can be attributed to not having enough data as easily as to there being a very small true effect.

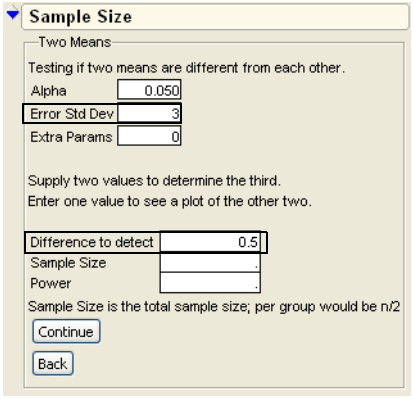
To gain some perspective on the power of the test, or to estimate how many data points are needed to detect a difference, we use the **Sample Size and Power** facility in JMP. This allows us to estimate some experimental values and graphically make decisions about the sample's data and effect sizes.

- Choose **DOE > Sample Size and Power**.

This brings up a list of prospective power and sample size calculators for several situations. In our case, we are concerned with comparing two means. From the Distribution report on height, we can see that the standard deviation is about 3. Suppose we want to detect a difference of 0.5.

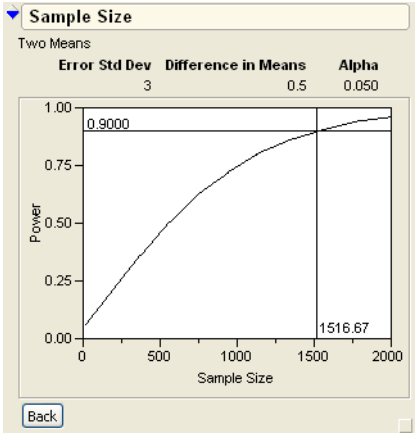
- Enter 3 for **Error Std Dev** and 0.5 as **Difference to Detect**, as shown in **Figure 8.8**.

Figure 8.8 Sample Size and Power Dialog



- Click **Continue** to see the graph shown in **Figure 8.9**.
- Use the crosshair tool to find out what sample size is needed to have a power of 90%.

Figure 8.9 Finding a Sample Size for 90% Power



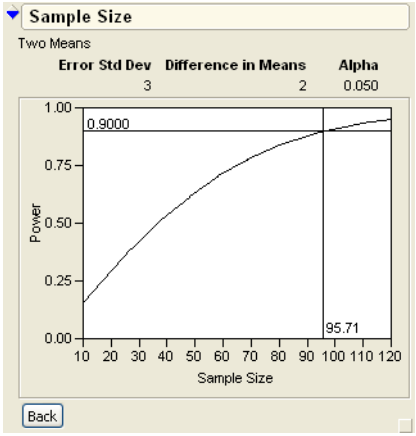
We would need around 1516 data points to detect a difference of 0.5 with the current standard deviation.

How would this change if we were interested in a difference of 2 rather than a difference of 0.5?

- Click the **Back** button and change the **Diff to Detect** from 0.5 to 2.
- Click **Continue**.
- Use the crosshair tool to find the number of data points you need for 90% power.

The results should be similar to **Figure 8.10**.

Figure 8.10 Data Points Needed When Difference Is Two



We would need only about 96 participants if we were interested in a difference of 2.

When the Difference Is Significant

The 12-year-olds in the previous example don't have significantly different average heights, but let's take a look at the 15-year-olds.

- To start, open the sample table called Htwt15.jmp.

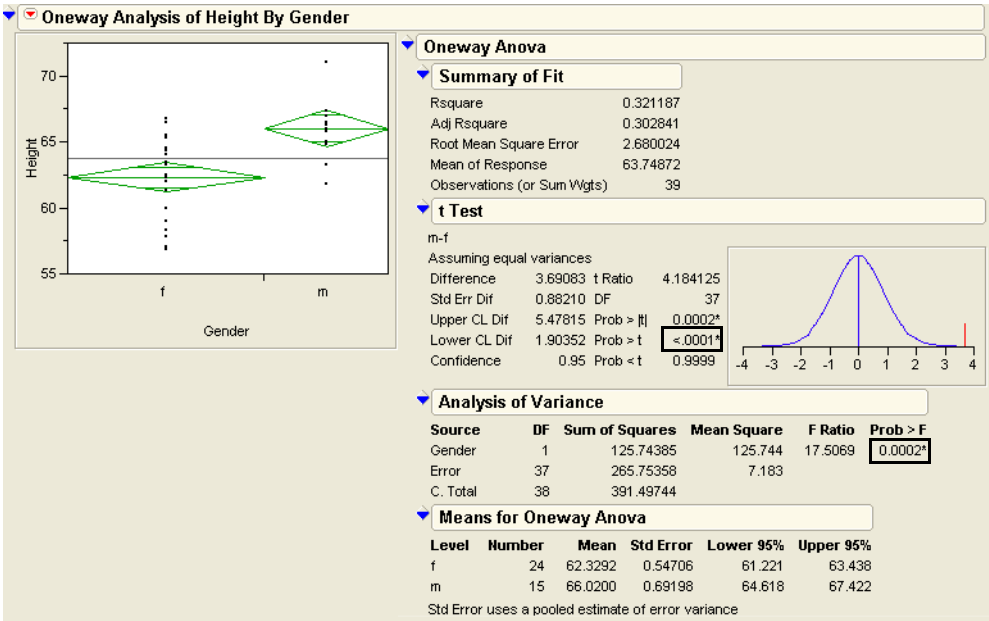
Then, proceed as before:

- Choose **Analyze > Fit Y by X**, with Gender as X and Height as Y.
- Click **OK.**
- Select **Means/Anova/t test** from the red triangle popup menu showing beside the report title.

This adds the analysis that estimates group means and tests to see if they are different.

You see the plot and tables shown in **Figure 8.11**.

Figure 8.11 Analysis for Mean Heights of 15-year-olds

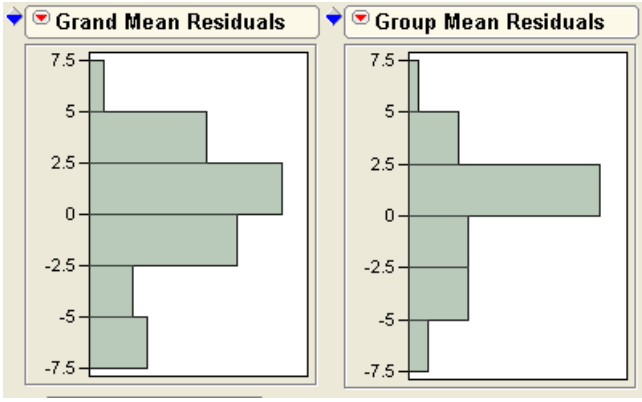


Note: We would normally recommend the unpooled (**t test** command) version of the test. We're using the pooled version as a basis for comparison.

The results for the analysis of the 15-year-old heights are completely different than the results for 12-year-olds. Here, the males are significantly taller than the females. You can see this because the confidence intervals shown by the means diamonds do not overlap. You can also see that the *p*-values for the *t*-test (< 0.001) and *F*-test (0.0002) are highly significant.

The *F*-test results say that the variance around the group means is significantly less than the variance around the grand mean. These two variances are shown in the histograms in **Figure 8.12**.

Figure 8.12 Histograms of Grand Means Variance and Group Mean Variance



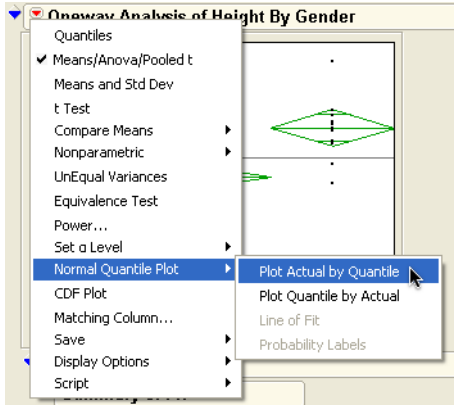
Normality and Normal Quantile Plots

The *t*-tests (and *F*-tests) used in this chapter assume that the sampling distribution for the group means is the Normal distribution. With sample sizes of at least 30 for each group, Normality is probably a safe assumption. The Central Limit Theorem says that means approach a Normal distribution as the sample size increases even if the original data are not Normal.

If you suspect non-Normality (due to small samples, or outliers, or a non-Normal distribution), consider using nonparametric methods, covered at the end of this chapter.

To assess Normality, use a Normal quantile plot. This is particularly useful when overlaid for several groups, because so many attributes of the distributions are visible in one plot.

- Return to the Fit Y by X platform showing Height by Gender for the 12-year-olds and select **Normal Quantile Plot > Plot Actual by Quantile** in the popup menu on the report title bar.
- Do the same for the 15-year-olds.



The resulting plot (**Figure 8.13**) shows the data compared to the Normal distribution. The Normality is judged by how well the

points follow a straight line. In addition, the Normal Quantile plot gives other useful information:

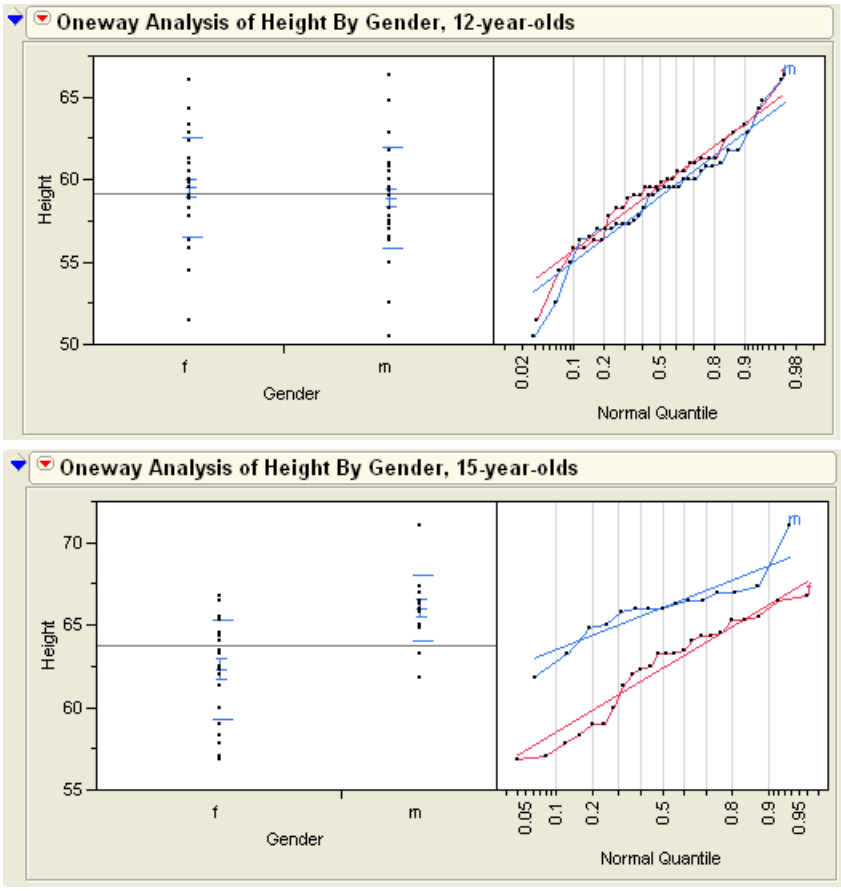
- The standard deviations are the slopes of the straight lines. Lines with steep slopes represent the distributions with the greater variances.
- The vertical separation of the lines in the middle shows the difference in the means. The separation of other quantiles shows at other points on the *x*-axis.

The first graph shown in **Figure 8.13** confirms that heights of 12-year-old males and females have nearly the same mean and variance—the slopes (standard deviations) are the same and the positions (means) are only slightly different.

The second graph in **Figure 8.13** shows 15-year-old males and females have different means and different variances—the slope (standard deviation) is higher for the females, but the position (mean) is higher for the males.

The distributions for all groups look reasonably Normal since the points (generally) cluster around their corresponding line.

Figure 8.13 Normal Quantile Plots for 12-year-olds and 15-year-olds



Testing Means for Matched Pairs

Consider a situation where two responses form a pair of measurements coming from the same experimental unit. A typical situation is a before-and-after measurement on the same subject. The responses are correlated, and if only the group means are compared—ignoring the fact that the groups have a pairing —information is lost. The statistical method called the *paired t-test* allows you to compare the group means, while taking advantage of the information gained from the pairings.

In general, if the responses are positively correlated, the paired *t*-test gives a more significant *p*-value than the *t*-test for independent means (grouped *t*-test) discussed in the previous sections. If responses are negatively correlated, then the paired *t*-test is less significant than the

grouped *t*-test. In most cases where the pair of measurements are taken from the same individual at different times, they are positively correlated, but be aware that it is possible for the correlation to be negative.

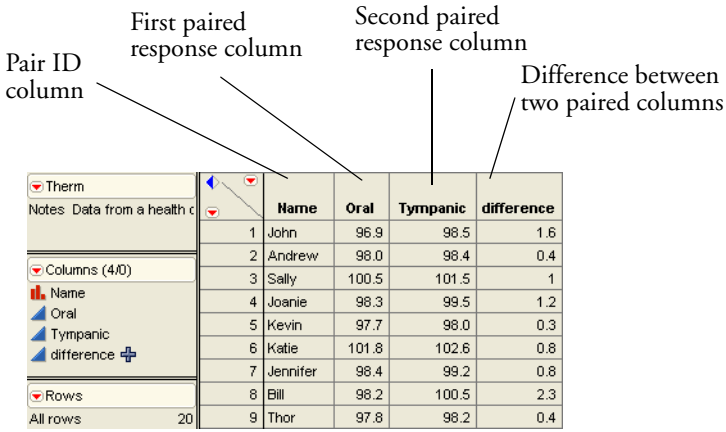
Thermometer Tests

A health care center suspected that temperature readings from a new ear drum probe thermometer were consistently higher than readings from the standard oral mercury thermometer. To test this hypothesis, two temperature readings were taken, one with the ear-drum probe, and the other with the oral thermometer. Of course, there was variability among the readings, so they were not expected to be exactly the same. However, the suspicion was that there was a systematic difference—that the ear probe was reading too high.

For this example, open the Therm.jmp data file.

A partial listing of the data table appears in **Figure 8.14**. The Therm.jmp data table has 20 observations and 4 variables. The two responses are the temperatures taken orally and tympanically (by ear) on the same person on the same visit.

Figure 8.14 Comparing Paired Scores



	Name	Oral	Tympanic	difference
1	John	96.9	98.5	1.6
2	Andrew	98.0	98.4	0.4
3	Sally	100.5	101.5	1
4	Joanie	98.3	99.5	1.2
5	Kevin	97.7	98.0	0.3
6	Katie	101.8	102.6	0.8
7	Jennifer	98.4	99.2	0.8
8	Bill	98.2	100.5	2.3
9	Thor	97.8	98.2	0.4

For paired comparisons, the two responses need to be arranged in two columns, each with a continuous modeling type. This is because JMP assumes each row represents a single experimental unit. Since the two measurements are taken from the same person, they belong in the same row. It is also useful to create a new column with a formula to calculate the difference between the two responses. (If your data table is arranged with the two responses in different rows, use the **Tables > Split** command to rearrange it. For more information, see “Juggling Data Tables” on page 50.)

Look at the Data

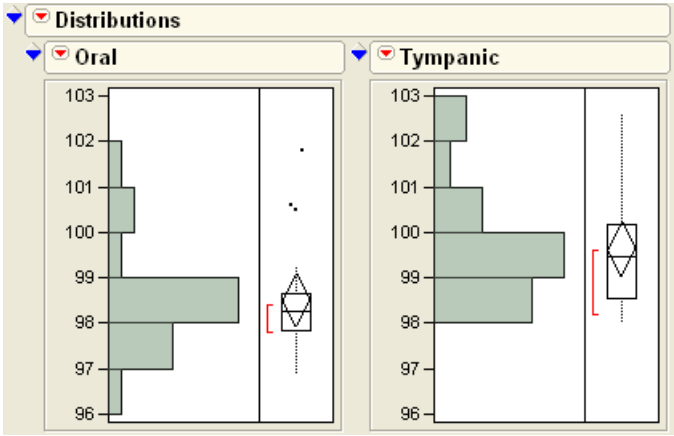
Start by inspecting the distribution of the data. To do this:

- Choose **Analyze > Distribution** with Oral and Tympanic as Y variables.
- When the results appear, select **Uniform Scaling** from the popup menu on the Distribution title bar to display the plots on the same scale.

The histograms (in **Figure 8.15**) show the temperatures to have different distributions. The mean looks higher for the Tympanic temperatures. However, as you will see later, this side-by-side picture of each distribution can be very misleading if you try to judge the significance of the difference from this perspective.

What about the outliers at the top end of the Oral temperature distribution? Are they of concern? Can you expect the distribution to be Normal? Not really. *It is not the temperatures that are of interest, but the difference in the temperatures*. So there is no concern about the distribution so far. If the plots showed temperature readings of 110 or 90, there would be concern, because that would be suspicious data for human temperatures.

Figure 8.15 Plots and Summary Statistics for Temperature



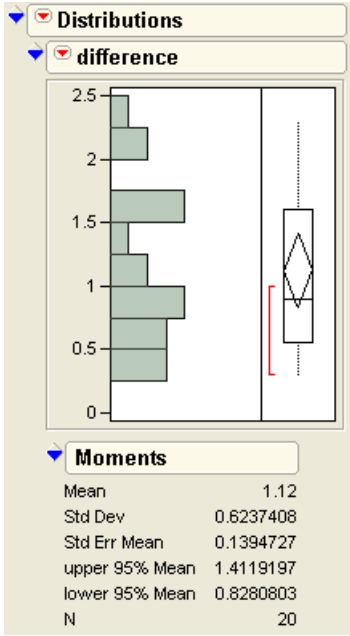
Look at the Distribution of the Difference

The comparison of the two means is actually a comparison of the difference between them. Inspect the distribution of the differences:

- Choose **Analyze > Distribution** with difference as the Y variable.

The results (shown in **Figure 8.16**) show a distribution that seems to be above zero. In the Moments table, the lower 95% limit for the mean is 0.828—greater than zero. The Student's *t*-test shows the mean to be significantly above zero.

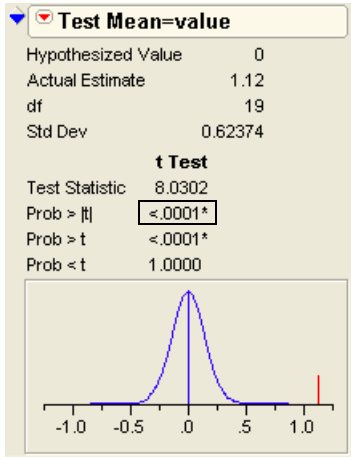
Figure 8.16 Histogram and Moments of the Difference Score



Student's *t*-Test

- Choose **Test Mean** from the popup menu on the report title bar for the histogram of the difference variable. When prompted for a hypothesized value, accept the default value of zero.
- If the box for the Wilcoxon's signed-rank test is checked, uncheck it—this is covered later.
- Click **OK**.

Now you have the *t*-test for testing that the mean over the matched pairs is the same. In this case, the results in the Test Mean table, shown to the right, show a *p*-value of less than 0.0001, which supports our visual guess that there is a significant difference between methods of temperature taking. The tympanic temperatures are significantly higher than the oral temperatures.



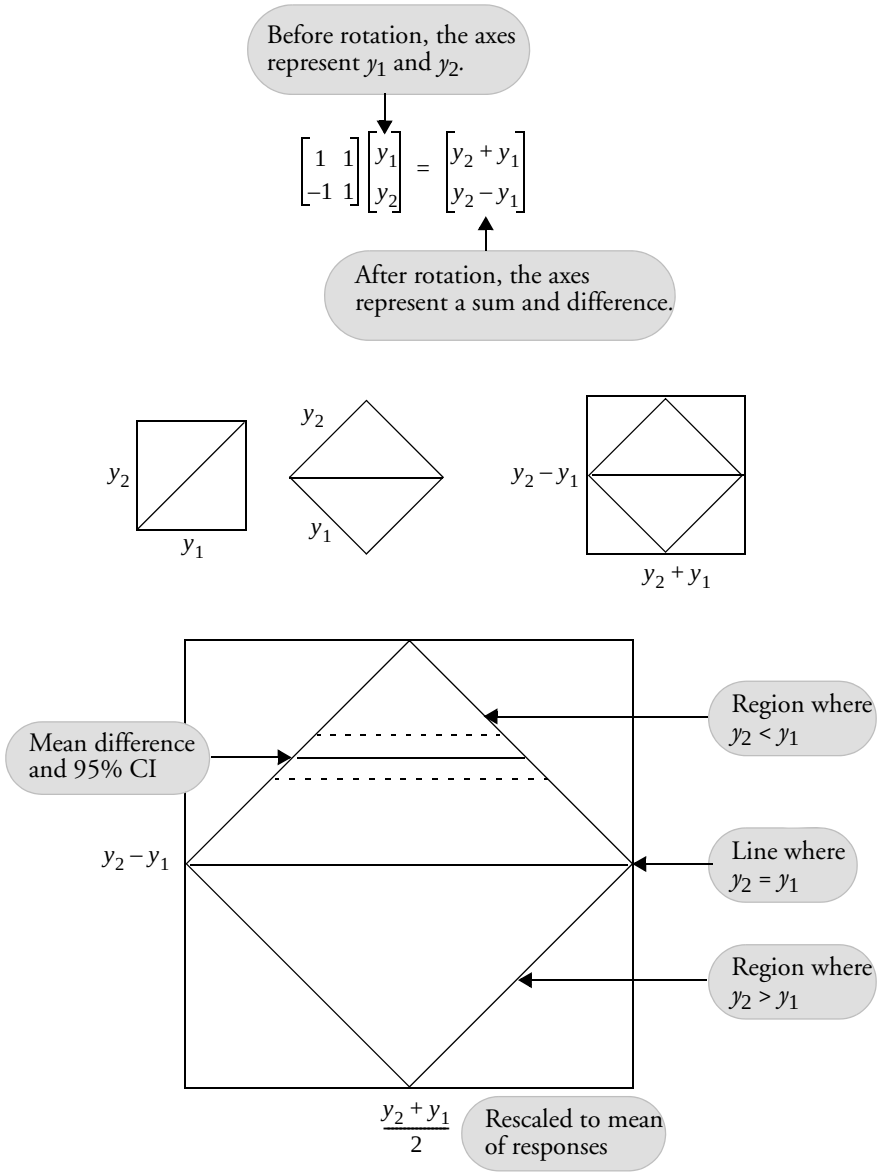
There is also a nonparametric test, the Wilcoxon signed-rank test, described at the end of this chapter, that tests the difference between two means. This test is produced by checking the appropriate box on the test mean dialog.

- Choose **Test Mean** from the popup menu next on the difference report title bar and accept zero as the hypothesized value. This time, make sure the box for Wilcoxon's signed-rank test is checked.
- Click **OK**.

The Matched Pairs Platform for a Paired *t*-Test

JMP offers a special platform for the analysis of paired data. The Matched Pairs platform compares means between two response columns using a paired *t*-test. The primary plot in the platform is a plot of the difference of the two responses on the *y*-axis, and the mean of the two responses on the *x*-axis. This graph is the same as a scatterplot of the two original variables, but rotated 45°clockwise. A 45°rotation turns the original coordinates into a difference and a sum. By rescaling, this plot can show a difference and a mean, as illustrated in **Figure 8.17**.

Figure 8.17 Transforming to Difference by Sum Is a Rotation by 45°



There is a horizontal line at zero, and a confidence interval is plotted using dashed lines. If the confidence interval does not contain the horizontal zero line, the test detects a significant difference.

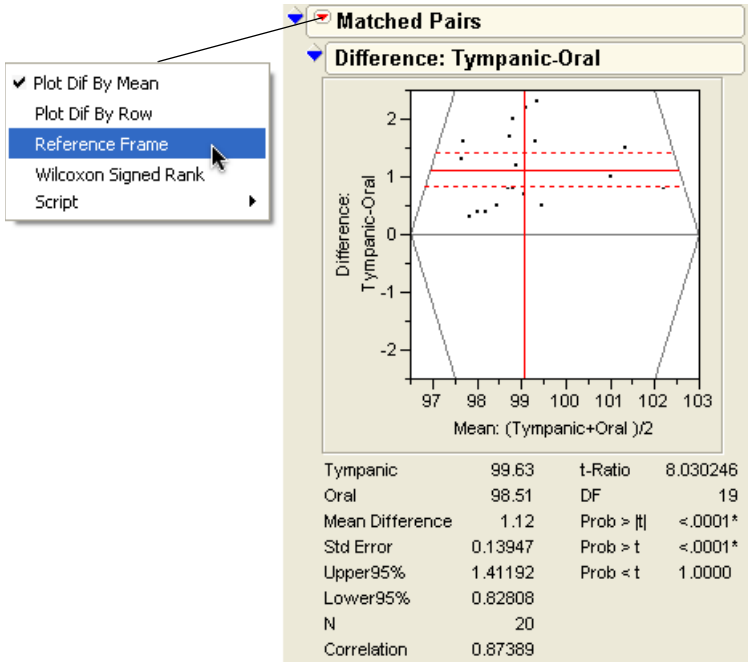
Seeing the platform in use reveals its usefulness.

- Choose **Analyze > Matched Pairs** and use Oral and Tympanic as the paired responses.
- Click **OK** to see a scatterplot of Tympanic and Oral as a matched pair.

To see the rotation of the scatterplot in **Figure 8.18** more clearly,

- Select the **Reference Frame** option from the popup menu on the Matched Pairs title bar.
- Right-click on the vertical axis and choose **Revert Axis** from the menu that appears.

Figure 8.18 Scatterplot of Matched Pairs Analysis



The analysis first draws a reference line where the difference is equal to zero. This is the line where the two columns are equal. If the means are equal, then the points should be evenly distributed around this line. You should see about as many points above this line as below it. If a point is above the reference line, it means that the difference is greater than zero. In this example, points above the line show the situation where the Tympanic temperature is greater than the Oral temperature.

Parallel to the reference line at zero is a solid red line that is displaced from zero by an amount equal to the difference in means between the two responses. This red line is the line of fit for the sample. The test of the means is equivalent to asking if the red line through the points is significantly separated from the reference line at zero.

The dashed lines around the red line of fit show the 95% confidence interval for the difference in means.

This scatterplot gives you a good idea of each variable's distribution, as well as the distribution of the difference.

Interpretation Rule for the Paired *t*-test Scatterplot: If the confidence interval (represented by the dashed lines around the red line) contains the reference line at zero, then the two means are not significantly different.

Another feature of the scatterplot is that you can see the correlation structure. If the two variables are positively correlated, they lie closer to the line of fit, and the variance of the difference is small. If the variables are negatively correlated, then most of the variation is perpendicular to the line of fit, and the variance of the difference is large. It is this variance of the difference that scales the difference in a *t*-test and determines whether the difference is significant.

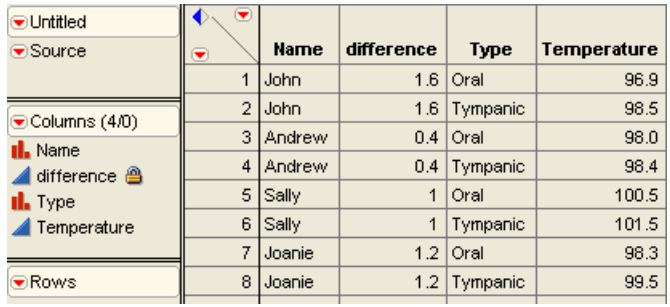
The paired *t*-test table beneath the scatterplot of **Figure 8.18** gives the statistical details of the test. The results should be identical to those shown earlier in the Distribution platform. The table shows that the observed difference in temperature readings of 1.12 degrees is significantly different from zero.

Optional Topic: An Equivalent Test for Stacked Data

There is a third approach to the paired *t*-test. Sometimes, you receive grouped data with the response values stacked into a single column instead of having a column for each group. Here is how to rearrange the Therm.jmp data table and see what a stacked table looks like:

- Choose **Tables > Stack**.
- When the Stack dialog appears, select Oral and Tympanic as the **Stack Columns**. Name the Stacked Data Column Temperature and the Source Label column Type.
- Click **OK** to see the data shown here.

The response values (temperatures) are in the Temperature column, identified as "Oral" or "Tympanic" by the Type column.



	Name	difference	Type	Temperature
1	John	1.6	Oral	96.9
2	John	1.6	Tympanic	98.5
3	Andrew	0.4	Oral	98.0
4	Andrew	0.4	Tympanic	98.4
5	Sally	1	Oral	100.5
6	Sally	1	Tympanic	101.5
7	Joanie	1.2	Oral	98.3
8	Joanie	1.2	Tympanic	99.5

If you choose **Analyze > Fit Y by X** with Temperature (the response of both temperatures) as Y and Type (the classification) as X, then select **t test** from the popup menu, you get the *t*-test designed for independent groups. However, this is inappropriate for paired data.

However, fitting a model that includes an adjustment for each person fixes the independence problem because the correlation is due to temperature differences from person to person. To do this, you need to use the **Fit Model** command, covered in detail in "Fitting Linear Models" on page 345. The response is modeled as a function of both the category of interest (Type—Oral or Tympanic) and the Name category that identifies the person.

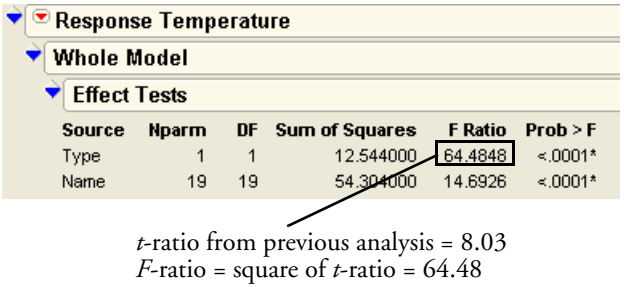
- Choose **Analyze > Fit Model**.
- When the Fit Model dialog appears, add Temperature as Y, and both Type and Name as Model Effects.
- Click **Run Model**.

When you get the results, the *p*-value for the category effect is identical to the *p*-value that you would get from the ordinary paired *t*-test. In fact, the *F*-ratio in the effect test is exactly the square of the *t*-test value in the paired *t*-test. In this case the formula is

(Paired *t*-test statistic) ^{2} = 8.032^2 = 64.4848 = (stacked *F*-test statistic)

The Fit Model platform gives you a plethora of information, but for this example you need to only open the Effect Test table (**Figure 8.19**). It shows an *F*-ratio of 64.48, which is exactly the square of the *t*-ratio of 8.03 found with the previous approach. It's just another way of doing the same test.

Figure 8.19 Equivalent *F*-test on Stacked Data



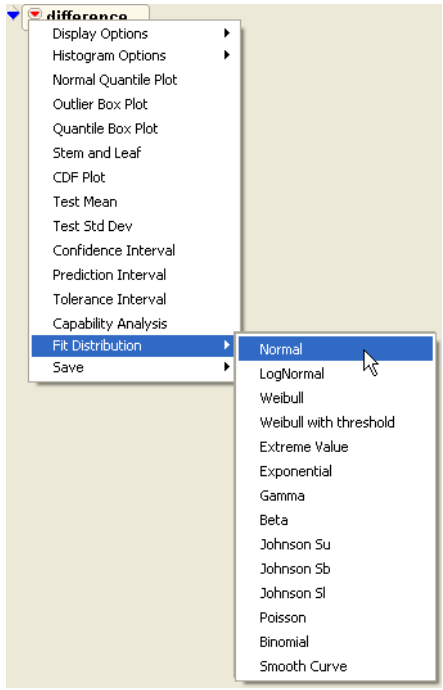
The alternative formulation for the paired means covered in this section is important for cases in which there are more than two related responses. Having many related responses is a *repeated-measures* or *longitudinal* situation. The generalization of the paired *t*-test is called the *multivariate* or T^2 approach, whereas the generalization of the stacked formulation is called the *mixed-model* or *split-plot* approach.

The Normality Assumption

The paired *t*-test assumes the differences are Normally distributed. With 30 pairs or more, this is probably a safe assumption. The results are reliable even if the distribution departs from Normal. The temperatures example only has 20 observations, so some people may like to check the Normality. To do this:

- Use the Therm.jmp data table and select **Analyze > Distribution** on the variable difference.
- Select **Normal Quantile Plot** and **Quantile Box Plot** from the popup menu on the Difference title bar.

- Select **Fit Distribution > Normal** from the same menu as shown here.

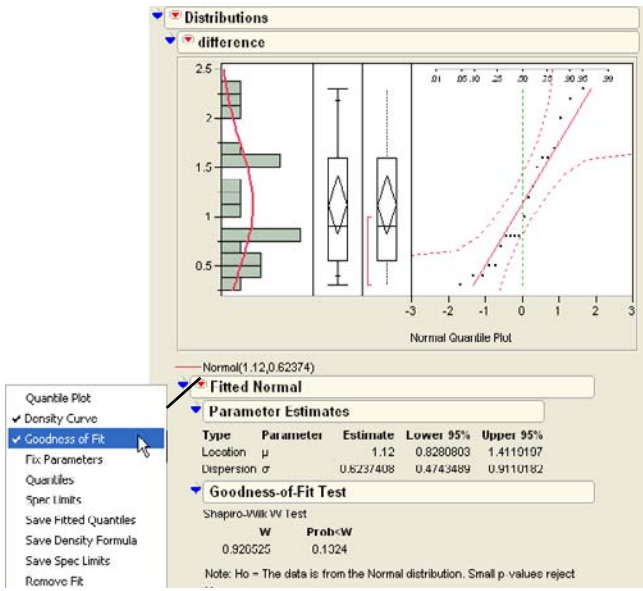


- Scroll down to the Fitted Normal report, and select **Goodness of Fit** from the red triangle popup menu found on its title bar (See **Figure 8.20**).

The quantile plot, outlier box plot, and the S-shaped Normal quantile plot are all indicative of a slightly skewed distribution. The Goodness of Fit table in **Figure 8.20**, with a *p*-value of 0.1326, also indicates that the distribution might not be Normal. (Remember: the null hypothesis for Normality tests is that the distribution is Normal. In this case, we don't reject the hypothesis, and therefore assume Normality.)

If you are concerned with non-Normality, nonparametric methods or a data transformation should be considered.

Figure 8.20 Looking at the Normality of the Difference Score



Two Extremes of Neglecting the Pairing Situation: A Dramatization

What happens if you do the wrong test? What happens if you do a *t*-test for independent groups on highly correlated paired data?

Consider the following two data tables:

- Open the sample data table called Bptime.jmp to see the left-hand table in **Figure 8.27**.

This table represents blood pressure measured for ten people in the morning and again in the afternoon. The hypothesis is that, on average, the blood pressure in the morning is the same as it is in the afternoon.

- Open the sample data table called BabySleep.jmp to see the right-hand table in **Figure 8.21**.

In this table, a researcher examined ten two-month-old infants at 10 minute intervals over a day to count the intervals in which a baby was asleep or awake. The hypothesis is that at two months old, the asleep time is equal to the awake time.

Figure 8.21 The Bptime and Babysleep Data Tables

BpTime

Notes This table repr

side by side

univ. test Dif

paired t-test

bivariate

Columns (3/0)

BP AM

BP PM

Dif

Rows

All rows 10

	BP AM	BP PM	Dif
1	70	94	24
2	85	100	15
3	92	106	14
4	97	113	16
5	110	130	20
6	110	131	21
7	126	142	16
8	137	149	12
9	140	156	16
10	148	170	22

BabySleep

Notes A researcher

side by side

univ. test Dif

paired t-test

bivariate

Columns (3/0)

Awake

Asleep

Dif

Rows

All rows 10

	Awake	Asleep	Dif
1	110	131	21
2	126	113	-13
3	85	156	71
4	140	100	-40
5	92	149	57
6	70	170	100
7	148	94	-54
8	97	142	45
9	137	106	-31
10	110	130	20

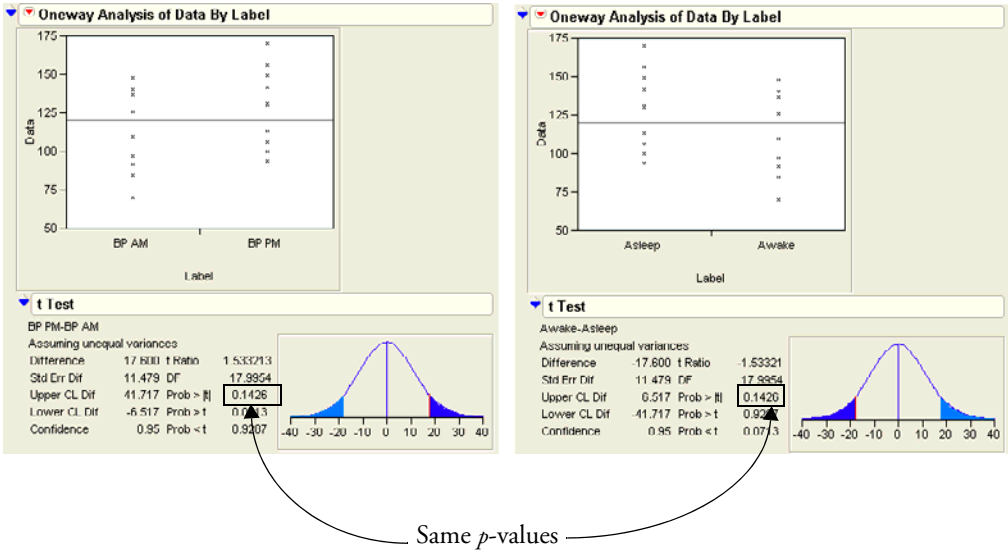
Let's do the incorrect *t*-test: the *t*-test for independent groups. Before conducting the test, we need to reorganize the data using the **Stack** command.

- In two separate tables, stack Awake and Asleep to form a single column, and BP AM and BP PM to form a single column.
- Select **Analyze > Fit Y by X** on both new tables, using the Label column as Y and the Data column as X.

- Choose **t test** from the popup menu on the title bar of each plot.

The results for the two analyses are shown in **Figure 8.22**. The conclusions are that there is no significant difference between Awake and Asleep time, nor is there a difference between time of blood pressure measurement. The summary statistics are the same in both analysis and the probability is the same, showing no significance (p = 0.1426).

Figure 8.22 Results of t-test for Independent Means



Now do the proper test, the paired t-test.

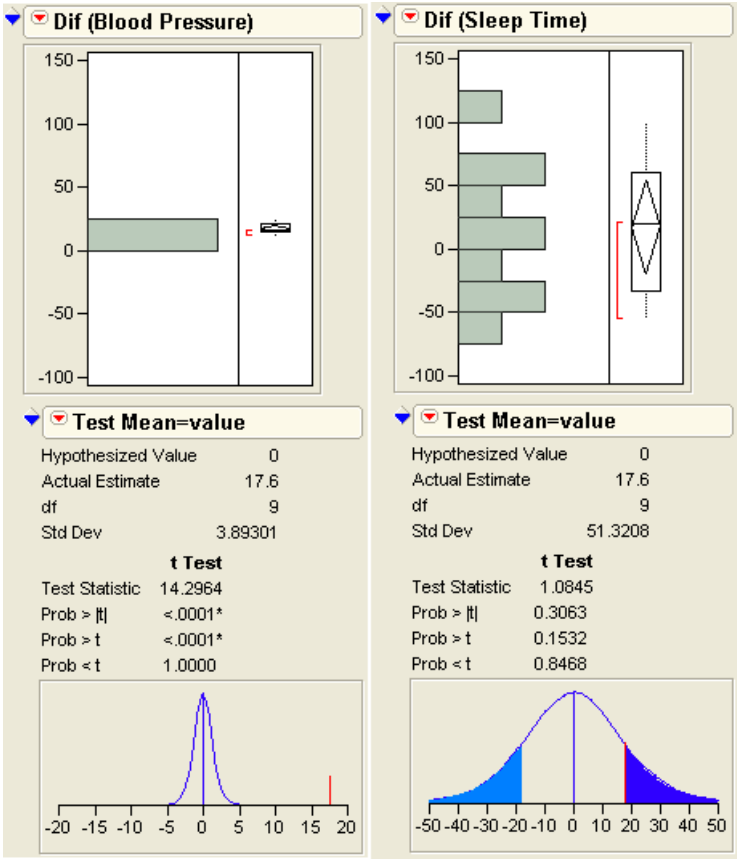
- Using the **Distribution** command, examine a distribution of the Dif variable in each table.
- Double click on the axis of the blood pressure histogram and make its scale match the scale of the baby sleep axis.
- Then, test that each mean is zero (see **Figure 8.23**).

In this case the analysis of the differences leads to very different conclusions.

- The mean difference between time of blood pressure measurement is highly significant because the variance is small (Std Dev=3.89).
- The mean difference between awake and asleep time is not significant because the variance of that difference is large (Std Dev=51.32).

So don't judge the mean of the difference by the difference in the means without noting that the variance of the difference is the measuring stick, and that the measuring stick depends on the correlation between the two responses.

Figure 8.23 Histograms and Summary Statistics Show the Problem



The scatterplots produced by the Bivariate platform (**Figure 8.24**) and the Matched Pairs platform (**Figure 8.25**) show what is happening. The first pair is highly positively correlated, leading to a small variance for the difference. The second pair is highly negatively correlated, leading to a large variance for the difference.

Figure 8.24 Bivariate Scatterplots of Blood Pressure Data and Baby Sleep Data

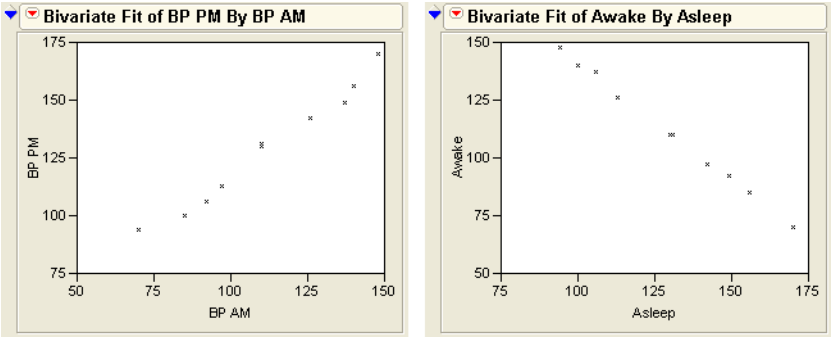
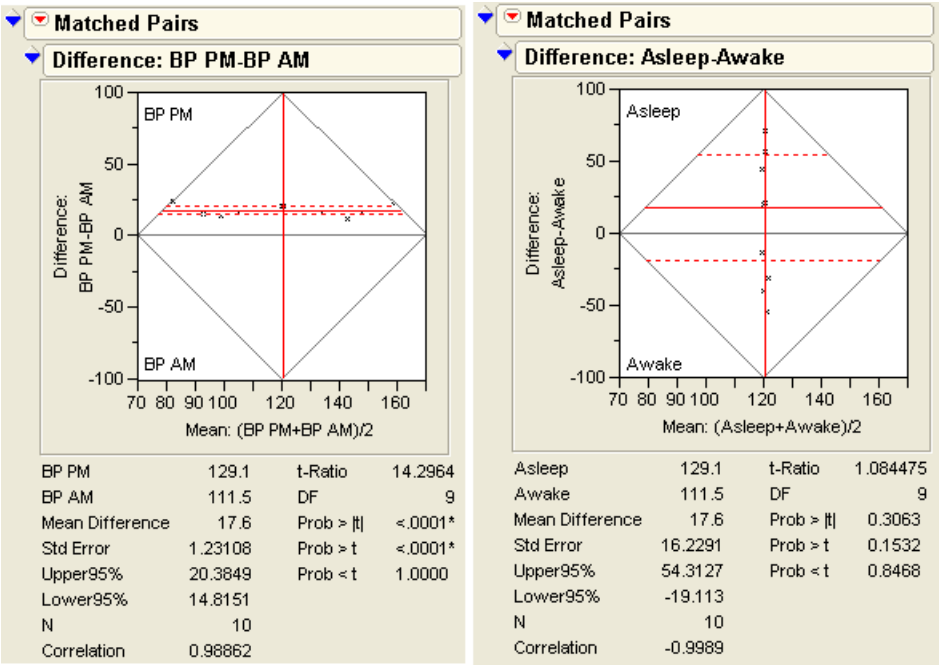


Figure 8.25 Paired *t*-test for Positively Correlated and Negatively Correlated Data



To review, make sure you can answer the following question:

What is the reason that you use a different *t*-test for matched pairs?

a. Because the statistical assumptions for the *t*-test for groups are not satisfied with correlated data.

b. Because you can detect the difference much better with a paired *t*-test. The paired *t*-test is much more sensitive to a given difference.

c. Because you might be overstating the significance if you used a group *t*-test rather than a paired *t*-test.

d. Because you are testing a different thing.

Answer: All of the above.

a. The grouped *t*-test assumes that the data are uncorrelated and paired data are correlated. So you would violate assumptions using the grouped *t*-test.

b. Most of the time the data are positively correlated, so the difference has a smaller variance than you would attribute if they were independent. So the paired *t*-test is more powerful—that is, more sensitive.

c. There may be a situation in which the pairs are negatively correlated, and if so, the variance of the difference would be greater than you expect from independent responses. The grouped *t*-test would overstate the significance.

d. You are testing the same thing in that the mean of the difference is the same as the difference in the means. But you are testing a different thing in that the variance of the mean difference is different than the variance of the differences in the means (ignoring correlation), and the significance for means is measured with respect to the variance.

Mouse Mystery

Comparing two means is not always straightforward. Consider this story.

A food additive showed promise as a dieting drug. An experiment was run on mice to see if it helped control their weight gain. If it proved effective, then it could be sold to millions of people trying to control their weight.

After the experiment was over, the average weight gain for the treatment group was significantly less than for the control group, as hoped for. Then someone noticed that the treatment group had fewer observations than the control group. It seems that the food additive caused the obese mice in that group to tend to die young, so the thinner mice had a better survival rate for the final weighing.

These tables are set up such that the values are identical for the two responses, as a marginal distribution, but the values are paired differently so that the bpTime difference is highly significant and the babySleep difference is non-significant. This illustrates that it is the distribution of the difference that is important, not the distribution of the original values. If you don't look at the data correctly, the data can appear the same even when they are dramatically different.

A Nonparametric Approach

Introduction to Nonparametric Methods

Nonparametric methods provide ways to analyze and test data that do not depend on assumptions about the distribution of the data. In order to ignore Normality assumptions, nonparametric methods disregard some of the information in your data. Typically, instead of using actual response values, you use the *rank ordering* of the response.

Most of the time you don't really throw away much relevant information, but you avoid information that might be misleading. A nonparametric approach creates a statistical test that ignores all the spacing information between response values. This protects the test against distributions that have very non-Normal shapes, and can also provide insulation from data contaminated by rogue values.

In many cases, the nonparametric test has almost as much power as the corresponding parametric test and in some cases has more power. For example, if a batch of values is Normally distributed, the rank-scored test for the mean has 95% efficiency relative to the most powerful Normal-theory test.

The most popular nonparametric techniques are based on functions (scores) of the ranks:

- the rank itself, called a *Wilcoxon score*
- whether the value is greater than the median; whether the rank is more than $\frac{n+1}{2}$, called the *Median test*
- a Normal quantile, computed as in Normal quantile plots, called the *van der Waerden* score

Nonparametric methods are not contained in a single platform in JMP, but are available through many platforms according to the context where that test naturally occurs.

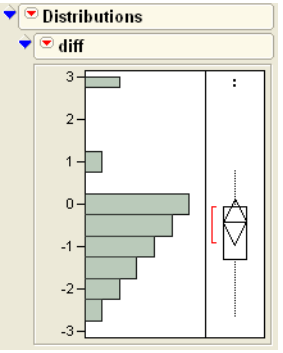
Paired Means: The Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is the nonparametric analog to the paired *t*-test. You do a signed-rank test by testing the distribution of the difference of matched pairs, as discussed previously. The following example shows the advantage of using the signed-rank test when data are non-Normal.

Open the Chamber.jmp table.

The data represent electrical measurements on 24 wiring boards. Each board is measured first when soldering is complete, and again after three weeks in a chamber with a controlled environment of high temperature and humidity (Iman 1995).

- Examine the diff variable (difference between the outside and inside chamber measurements) with **Analyze > Distribution**.



- Select the **Fit Distribution > Normal** from the popup menu on the title bar of the diff histogram.
- Select **Goodness of Fit** from the popup menu on the Fitted Normal Report.

The Shapiro-Wilk *W*-test that results tests the assumption that the data are Normal. The probability of 0.0090 given by the Normality test indicates that the data are significantly non-Normal. In this situation, it might be better to use signed ranks for comparing the mean of diff to zero. Since this is a matched pairs situation, we use the Matched Pairs platform.

Figure 8.26 The Chamber Data and Test For Normality

Chamber
Notes Data from Imar

Columns (4/0)
board
outside
inside
diff

Rows
All rows 24

	board	outside	inside	diff
1	1	0.23	0.29	-0.06
2	2	1.38	1.32	0.06
3	3	0.75	0.7	0.05
4	4	1.46	2.88	-1.42
5	5	0.45	1.28	-0.83
6	6	0.48	1.36	-0.88
7	7	0.07	0.37	-0.3
8	8	0.57	3.23	-2.66
9	9	0.47	2.14	-1.67

Distributions

diff

Normal(-0.4333,1.27974)

Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	-0.433333	-0.973722	0.1070552
Dispersion	σ	1.2797441	0.9946345	1.7951746

Goodness-of-Fit Test

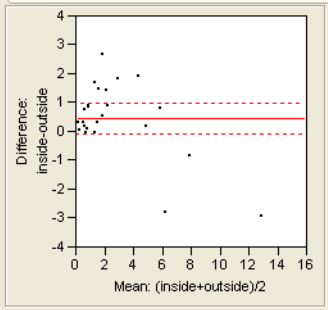
Shapiro-Wilk W Test

W	Prob<W
0.881636	0.0090*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Matched Pairs

Difference: inside-outside



inside	2.90667	t-Ratio	1.65884
outside	2.47333	DF	23
Mean Difference	0.43333	Prob > |t|	0.1107
Std Error	0.26123	Prob > t	0.0554
Upper95%	0.97372	Prob < t	0.9446
Lower95%	-0.1071		
N	24		
Correlation	0.94059		

Wilcoxon Sign-Rank

	inside-outside
Test Statistic	86.000
Prob > |z|	0.011
Prob > z	0.005
Prob < z	0.995

Select **Analyze > Matched Pairs**.

Assign outside and inside as the paired responses.

Click **OK**.

When the report appears,

Select **Wilcoxon Signed Rank** from the report's popup menu.

Note that the standard *t*-test probability is insignificant ($p = 0.1107$). However, in this example, the signed-rank test detects a difference between the groups with a *p*-value of 0.011.

Independent Means: The Wilcoxon Rank Sum Test

If you want to nonparametrically test the means of two independent groups, as in the *t*-test, then you can rank the responses and analyze the ranks instead of the original data. This is the *Wilcoxon rank sum test*. It is also known as the *Mann-Whitney U test* because there is a different formulation of it that was not discovered to be equivalent to the Wilcoxon rank sum test until after it had become widely used.

- Open Htwt15 again, and choose **Analyze > Fit Y by X** with Height as Y and Gender as X.

This is the same platform that gave the *t*-test.

- Select the **Nonparametric-Wilcoxon** command from the popup menu on the title bar at the top of the report.

The result is the report in **Figure 8.27**. This table shows the sum and mean ranks for each group, then the Wilcoxon statistic along with an approximate *p*-value based on the large-sample distribution of the statistic. In this case, the difference in the mean heights is declared significant, with a *p*-value of 0.0002. If you have small samples, you should consider also checking the tables of the Wilcoxon to obtain a more exact test, because the Normal approximation is not very precise in small samples.

Figure 8.27 Wilcoxon Rank Sum Test for Independent Groups

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
f	24	350.500	14.6042	-3.728
m	15	429.500	28.6333	3.728

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
429.5	3.72806	0.0002*

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
14.0064	1	0.0002*

Exercises

1. The file On-Time Arrivals.jmp (*Aviation Consumer Home Page,* 1999) contains the percentage of airlines' planes that arrived on time in 29 airports (those that the Department of Transportation designates "reportable"). You are interested in seeing if there are differences between certain months.

(a) Suppose you want to examine the differences between March and June. Is this a situation where a grouped test of two means is appropriate or would a matched pairs test be a better choice?

(b) Based on your answer in (a), determine if there is a difference in on-time arrivals between the two months.

(c) Similarly, determine if there is a significant difference between the months June and August, and also between March and August.

2. William Gossett was a pioneer in statistics. In one famous experiment, he wanted to investigate the yield from corn planted from two different types of seeds. One type of seed was dried in the normal way, while the other was kiln-dried. Gossett planted one of each seed in eleven different plots and measured the yield for each one. The drying methods are represented by the columns Regular or Kiln in the data file Gosset's Corn (Gosset 1908).

   (a) This is a matched pairs experiment. Explain why it is inappropriate to use the grouped-means method of determining the difference between the two seeds.

   (b) Using the matched pairs platform, determine if there is a difference in yield between kiln-dried and regular-dried corn.

3. The data file Companies.jmp (*Fortune* Magazine, 1990) contains data on sales, profits, and employees for two different industries (Computers and Pharmaceutical). This exercise is interested in detecting differences between the two types of companies.

   (a) Suppose you wanted to test for differences in sales amounts for the two business types. First, examine histograms of the variables Type and Sales and comment on the output.

   (b) In comparing sales for the two types of companies, should you use grouped means or matched pairs for the test?

   (c) Using your answer in part (b), determine if there is a difference between the sales amount of the two types of companies.

   (d) Should you remove any outliers in your analysis of part (c)? Comment on why this would or would not be appropriate in this situation.

4. Automobile tire manufacturing companies are obviously interested in the quality of their tires. One of their measures of quality is tire treadwear. In fact, all major manufacturers regularly sample their production and conduct tire treadwear tests. There are two accepted methods of measuring tread-wear: one based on weight loss during use, the other based on groove wear. A scientist at one of

these manufacturers decided to see if the two methods gave different results for the wear on tires. He set up an experiment that measured 16 tires, each by the two methods. His data is assembled in the file Tire Tread Measurement.jmp (Stichler, Richey, and Mandel, 1953).

(a) Determine if there is a difference in the two methods by using the matched pairs platform.

(b) Now, determine if there is a difference in the two methods by using group means. To do this, you will need to "stack" the data. Select **Tables**, then **Stack**, and select both columns to be stacked. After pressing **OK**, you will have a data table with two columns: one with the Weight/Groove identifier, the other with the measurement. At this point you are ready to carry out your analysis with the Fit Y By X platform.

(c) Which of the two methods (matched pairs or grouped means) is correct?

(d) Would the scientist have had different results by using the wrong method?

5. The data table Cars.jmp (Henderson and Velleman, 1981) contains information on several different brands of cars, including number of doors and impact compression for various parts of the body during crash tests.

   (a) Is there a difference between two- and four-door cars when it comes to impact compression on left legs?

   (b) Is there a difference between two- and four-door cars when it comes to compression on right legs?

   (c) Is there a difference between two- and four-door cars when it comes to head impact compression?

6. The data in Chamber.jmp represent electrical measurements on 24 electrical boards (this is the same data used in "Paired Means: The Wilcoxon Signed-Rank Test" on page 202). Each measurement was taken when soldering was complete, then again three weeks later after sitting in a temperature- and humidity-controlled chamber. The investigator wants to know if there is a difference between the measurements.

   (a) Why is this a situation that calls for a matched pairs analysis?

   (b) Determine if there is a significant difference between the means when the boards were outside vs. inside the chamber.

9 Comparing Many Means: One-Way Analysis of Variance

Overview

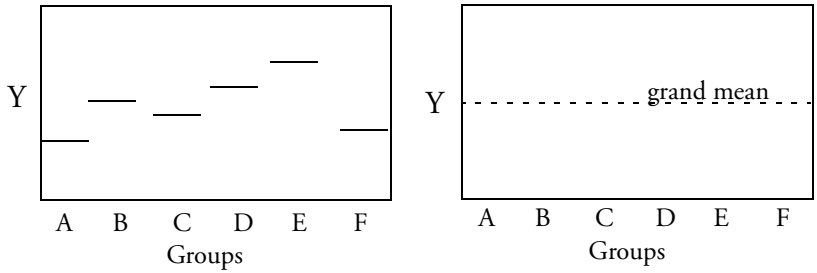
In Chapter 8, "The Difference between Two Means," the *t*-test was the tool used to compare the means of two groups. However, if you need to test the means of more than two groups, the *t*-test can't handle the job because it is only defined for two groups. This chapter shows how to compare more than two means using the one-way *analysis of variance*, or ANOVA for short. The *F*-test, which has made brief appearances in previous chapters, is the key element in an ANOVA. It is the statistical tool necessary to compare many groups, just as the *t*-test compares two groups. This chapter also introduces multiple comparisons and power calculations, reviews the topic of unequal variances, and extends nonparametric methods to the one-way layout.

What Is a One-Way Layout?

A one-way layout is the organization of data when a response is measured across a number of groups, and the distribution of the response may be different across the groups. The groups are labeled by a classification variable, which is a column in the JMP data table with the nominal or ordinal modeling type.

Usually, one-way layouts are used to compare group means. **Figure 9.1** shows a schematic that compares two models. The model on the left fits a different mean for each group, and the model on the right indicates a single grand mean (a single-mean model).

Figure 9.1 Different Mean for Each Group Versus a Single Overall Mean



The previous chapter showed how to use the *t*-test and the *F*-test to compare two means. When there are more than two means, the *t*-test is no longer applicable; the *F*-test must be used.

An *F*-test has the following features:

- An *F*-test compares two models, one constrained and the other unconstrained. The constrained model fits one grand mean. The unconstrained model for the one-way layout fits a mean for each group.
- The measurement of fit is done by accumulating the sum of the squares of *residuals*, where the residual is the difference between the actual response and the fitted response.
- A Mean Square is calculated by dividing a sum of squares by its degrees of freedom (DF). Mean Squares are estimates of variance, sometimes under the assumption that certain hypotheses are true.
- Degrees of Freedom (DF) are numbers, based on the number of parameters and number of data points in the model, that you divide by to get an unbiased estimate of the variance (see the chapter "What Are Statistics?" on page 95 for a definition of bias).

- An *F*-statistic is a ratio of Mean Squares (MS) that are independent and have the same expected value. In our discussion, this ratio is

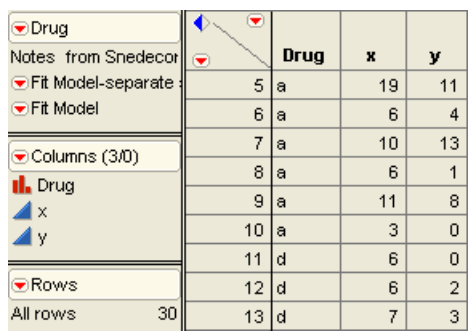
$$\frac{\text{Model MS}}{\text{Total MS}}$$

- If the null hypothesis that there is no difference between the means is true, this *F*-statistic has an *F distribution*.
- If the hypothesis is not true (if there is a difference between the means), the mean square for the model in the numerator of the *F*-ratio has some effect in it besides the error variance. This numerator produces a large (and significant) *F* if there is enough data.
- When there is only one comparison, the *F*-test is equivalent to the pooled (equal-variance) *t*-test. In fact, when there is only one comparison, the *F*-statistic is the square of the pooled *t*-statistic. This is true despite the fact that the *t*-statistic is derived from the distribution of the estimates, whereas the *F*-test is thought of in terms of the comparison of variances of residuals from two different models.

Comparing and Testing Means

The file Drug.jmp contains the results of a study that measured the response of 30 subjects to treatment by one of three drugs (Snedecor and Cochran, 1967). To begin,

Open Drug.jmp.

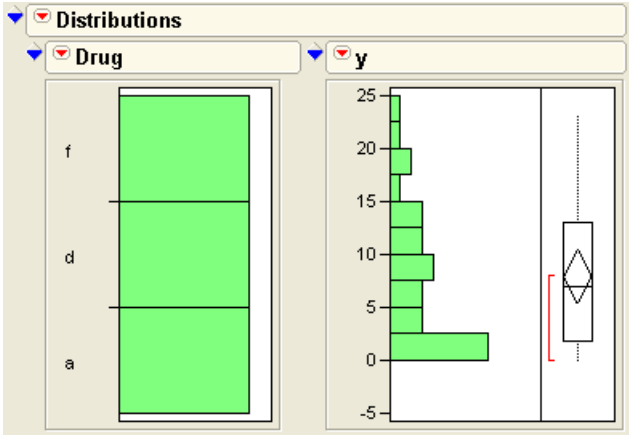


The three drug types are called “a”, “d”, and “f.” The y column is the response measurement. (The x column is used in a more complex model, covered in Chapter 14, “Fitting Linear Models.”)

For a quick look at the data, choose **Analyze > Distribution** and select Drug and y as Y variables.

Note in the histogram on the left in **Figure 9.2** that the number of observations is the same in each of the three drug groups; that is what is meant by a *balanced design*.

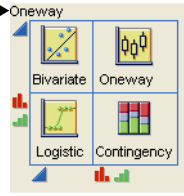
Figure 9.2 Distributions of Model Variables



Next, choose **Analyze > Fit Y by X** with Drug as X and y as Y.

Notice that the launch dialog displays the message that you are requesting a one-way analysis.

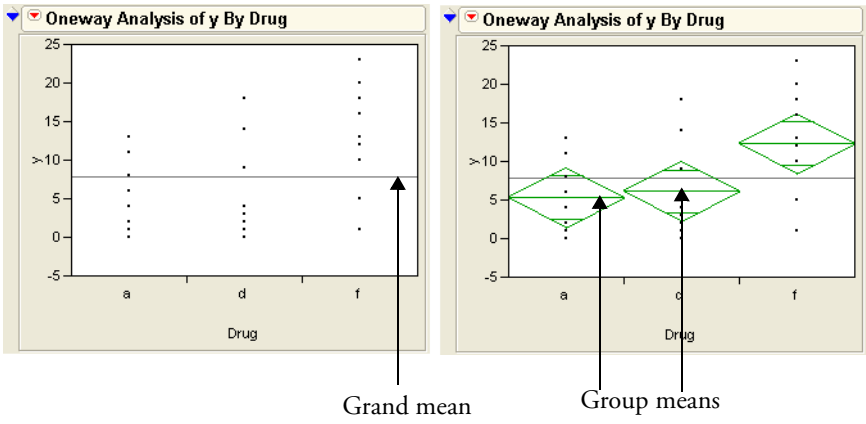
The launch dialog shows the appropriate analysis for the selected variables.



Click **OK.**

The results window in **Figure 9.3** appears. The initial plot on the left shows the distribution of the response in each drug group. The line across the middle is the grand mean. We want to test the null hypothesis that there is no difference in the response among the groups.

Figure 9.3 Distributions of Drug Groups

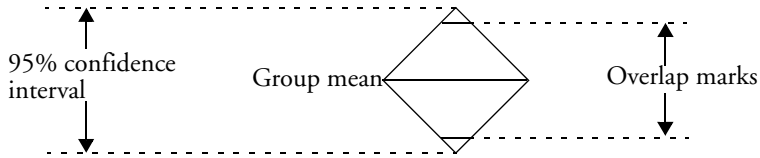


Means Diamonds: A Graphical Description of Group Means

Select **Means/Anova** from the popup menu showing on the title bar of the plot.

This adds means diamonds to the plot and also adds a set of text reports. The plot on the right in **Figure 9.3** shows means diamonds:

- The middle line in the diamond is the response group mean for the group.
- The vertical endpoints form the 95% confidence interval for the mean.
- The *x*-axis is divided proportionally by group sample size.



If the means are not much different, they are close to the grand mean. If the confidence intervals (the points of the diamonds) don't overlap, the means are significantly different.

See the section “Display and Compare the Means” on page 171 for details and interpretation rules for means diamonds.

Statistical Tests to Compare Means

The **Means/Anova** command produces a report composed of the three tables shown in **Figure 9.4**:

- The **Summary of Fit** table gives an overall summary of how well the model fits.
- The **Analysis of Variance** table gives sums of squares and an *F*-test on the means.
- The **Means for Oneway Anova** table shows the group means, standard error, and upper and lower 95% confidence limits on the mean.

Figure 9.4 One-Way ANOVA Report

Oneway Anova

Summary of Fit

Rsquare	0.227826
Adj Rsquare	0.170628
Root Mean Square Error	6.070878
Mean of Response	7.9
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Drug	2	293.6000	146.800	3.9831	0.0305*
Error	27	995.1000	36.856		
C. Total	29	1288.7000			

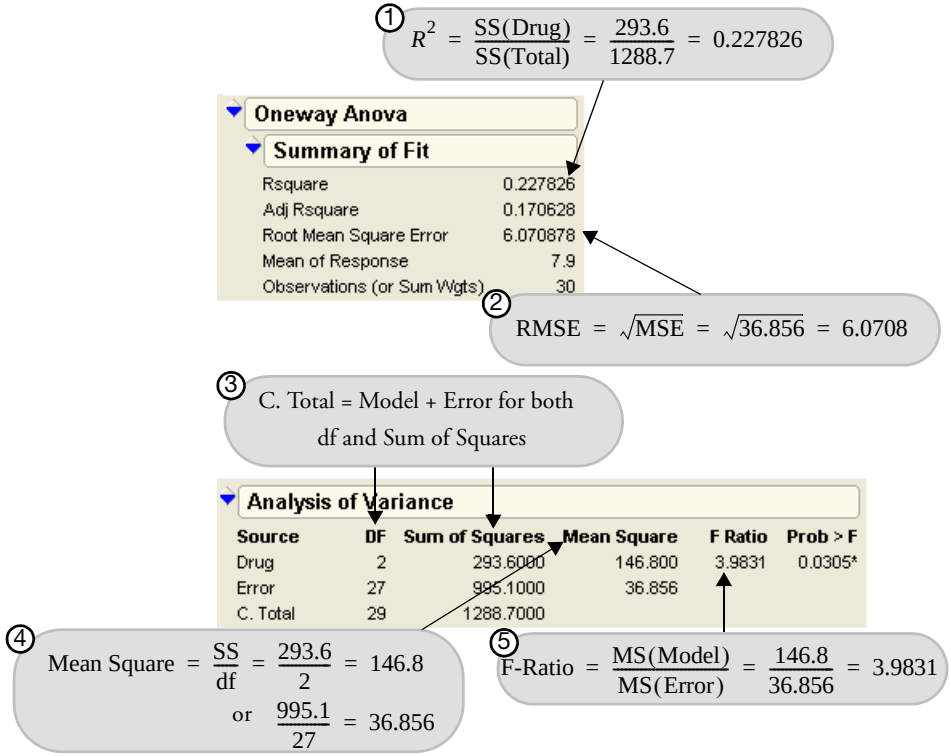
Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
a	10	5.3000	1.9198	1.3609	9.239
d	10	6.1000	1.9198	2.1609	10.039
f	10	12.3000	1.9198	8.3609	16.239

Std Error uses a pooled estimate of error variance

The Summary of Fit and the Analysis of Variance tables may look like a hodgepodge of numbers, but they are all derived by a few simple rules. **Figure 9.5** illustrates how the statistics relate.

Figure 9.5 Summary of Fit and ANOVA Tables



The Analysis of Variance table (**Figure 9.4** and **Figure 9.5**) describes three source components:

C. Total

The C. Total Sum of Squares (SS) is the sum of the squares of residuals around the grand mean. C. Total stands for *corrected total* because it is corrected for the mean. The C. Total degrees of freedom is the total number of observations in the sample minus 1.

Error

After you fit the group means, the remaining variation is described in the Error line. The Sum of Squares is the sum of squared residuals from the individual means. The remaining unexplained variation is C. Total minus Model (labeled Drug in this example) and is called Error for both the sum of squares and the degrees of freedom. The Error Mean Square estimates the variance.

Model

The Sum of Squares for the Model line is the difference between C. Total and Error. It is a measure of how much the residuals' sum of squares is accounted for by fitting the model rather than fitting only the grand mean. The degrees of freedom in the drug example is the number of parameters in the model (the number of groups, 3) minus 1.

Everything else in the Analysis of Variance table and the Summary of Fit table is derived from these quantities.

Mean Square

Mean Square is the sum of squares divided by their respective degrees of freedom.

F-ratio

The *F-ratio* is the model mean square divided by the error mean square. The *p*-value for this *F*-ratio comes from the *F*-distribution.

RSquare

The *RSquare* (R^2) is the proportion of variation explained by the model. In other words, it is the model sum of squares divided by the total sum of squares.

Adjusted RSquare

The *Adjusted RSquare* is more comparable over models with different numbers of parameters (degrees of freedom). It is the error mean square divided by the total mean square, subtracted from 1:

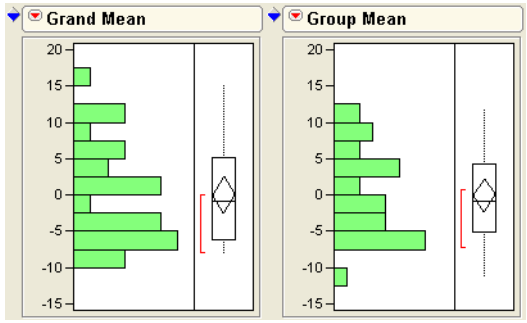
$$1 - \frac{\text{Error MS}}{\text{Total MS}}$$

Root Mean Square Error

The Root Mean Square Error is the square root of the Mean Square for Error in the Analysis of Variance table. It estimates the standard deviation of the error.

So what's the verdict for the null hypothesis that the groups are the same? The *F*-value of 3.98 is significant with a *p*-value of 0.03, which rejects the null hypothesis and confirms that there is a significant difference in the means. The *F*-test does not give any specifics about which means are different, only that there is at least one pair of means that is statistically different.

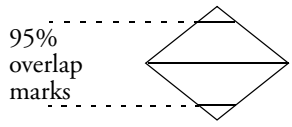
The *F*-test shows whether the variance of residuals from the model is smaller than the variance of the residuals from only fitting a grand mean. In this case, the answer is yes, but just barely. The histograms shown to the right compare the residuals from the grand means (left) with the group mean residuals (right).



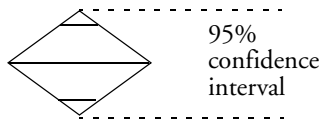
Means Comparisons for Balanced Data

At this point, we know that there is at least one pair of means that are different. Which means are significantly different from which other means? It looks like the mean for the drug "placebo" is separate from the other two. However, since all the confidence intervals for the means intersect, it takes further digging to see significance.

If two means were composed from the same number of observations, then you could use the overlap marks to get a more precise graphical measure of which means are significantly different. Two means are significantly different when their overlap marks don't overlap. The overlap marks are placed into the confidence interval at a distance of $1/\sqrt{2}$, a distance given by the Student's *t*-test of separation.



For balanced data, to be significantly different, two means must not overlap their overlap marks.



When two means do not have the same number of observations, the design is unbalanced and the overlap marks no longer apply. For these cases, JMP provides another technique using *comparison circles* to compare means. The next section describes comparison circles and shows you how to interpret them.

Means Comparisons for Unbalanced Data

Suppose, for the sake of this example, that the drug data are unbalanced. That is, there is not the same number of observations in each group. The following steps unbalance the Drug.jmp

data in an extreme way to illustrate an apparent paradox, as well as introduce a new graphical technique.

- Change Drug in rows 1, 4–7, and 9 from “a” to “f”. Change Drug in rows 2 and 3 to “d”. Change y in row 10 to “4.” (Be careful not to save this modified table over the original copy in your sample data.)

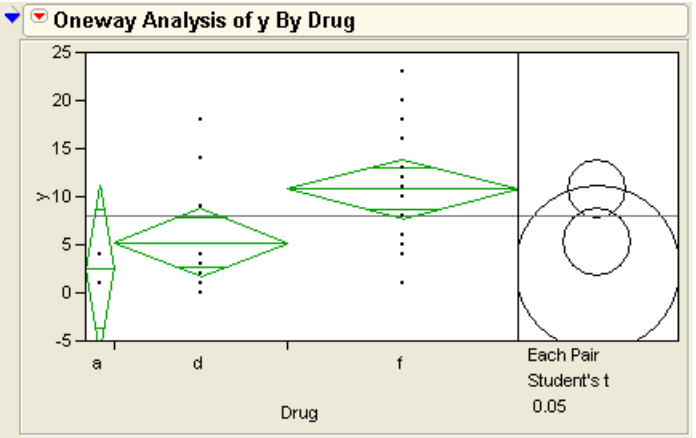
Now drug “a” has only two observations, whereas “d” has 12 and “placebo” has 16. The mean for “a” will have a very high standard error because it is supported by so few observations compared with the other two levels.

Again, use the **Fit Y by X** command to look at the data:

- Choose **Analyze > Fit Y by X** for the modified data and select the **Means/Anova** option from the Oneway menu.
- Select **Compare Means > Each Pair, Student's t** from the platform menu on the scatterplot title bar.

The modified data should give results like those illustrated in **Figure 9.6**. The *x*-axis divisions are proportional to the group sample size, which causes drug “a” to be very thin, because it has fewer observations. The confidence interval on its mean is large compared with the others. Comparison circles for Student's *t*-tests appear to the right of the means diamonds.

Figure 9.6 Comparison Circles to Compare Group Means



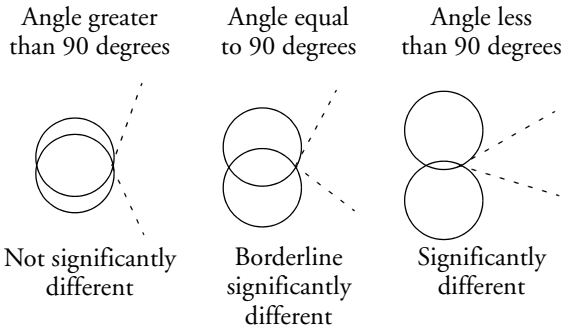
Comparison circles are a graphical technique that let you see significant separation among means in terms of how circles intersect. This is the only graphical technique that works in general with both equal and unequal sample sizes. The plot displays a circle for each group,

with the centers lined up vertically. The center of each circle is aligned with its corresponding group mean. The radius of a circle is the 95% confidence interval for its group mean, as you can see by comparing a circle with its corresponding means diamond. The non-overlapping confidence intervals shown by the diamonds for groups that are significantly different correspond directly to the case of non-intersecting comparison circles.

When the circles intersect, the angle of intersection is the key to seeing if the means are significantly different. If the angle of intersection is exactly a right angle (90°) then the means are on the borderline of being significantly different. See the *JMP Statistics and Graphics Guide* for details on the geometry of comparison circles.

If the circles are farther apart than the right angle case, then the outside angle is more acute and the means are significantly different. If the circles are closer together, the angle is larger than a right angle, and the means are not significantly different. **Figure 9.7** illustrates these angles of intersection.

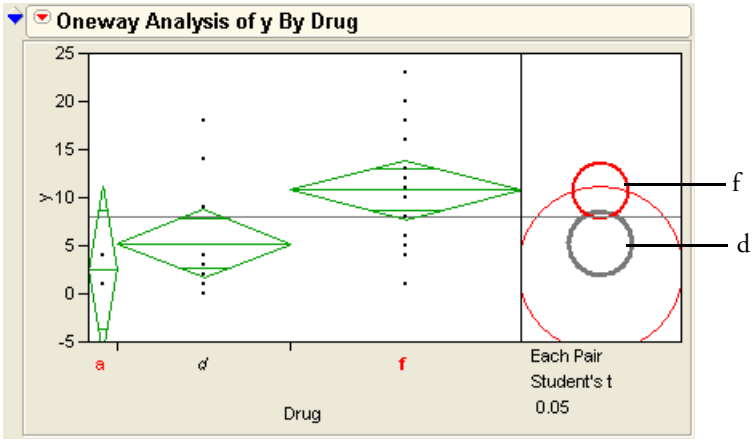
Figure 9.7 Diagram of How to Interpret Comparison Circles



So what are the conclusions for the drug example shown in **Figure 9.6**?

You don't need to hunt down a protractor to figure out the size of the angles of intersection. Click on a circle and see what happens. The circle highlights and becomes red. Groups that are not different from it also show in red. All groups that are significantly different remain black.

- Click on the “f” circle and use the circles to compare group means.



- The “f” and “d” means are represented by the smaller circles. The circles are farther separated than would occur with a right angle. The angle is acute, so these two means are significantly different. This is shown by their differing color.
- The circle for the “d” mean is completely nested in the circle for “a”, so they are not significantly different.
- The “a” mean is well below the “d” mean, which is significantly below “f.” By transitivity, one might expect “a” to be significantly different than “f.” The problem with this logic is that the standard error around the “a” mean is so large that it is not significantly different from “f”, even though it is farther away than “d.” The angle of intersection is greater than a right angle, so they are not significantly different.

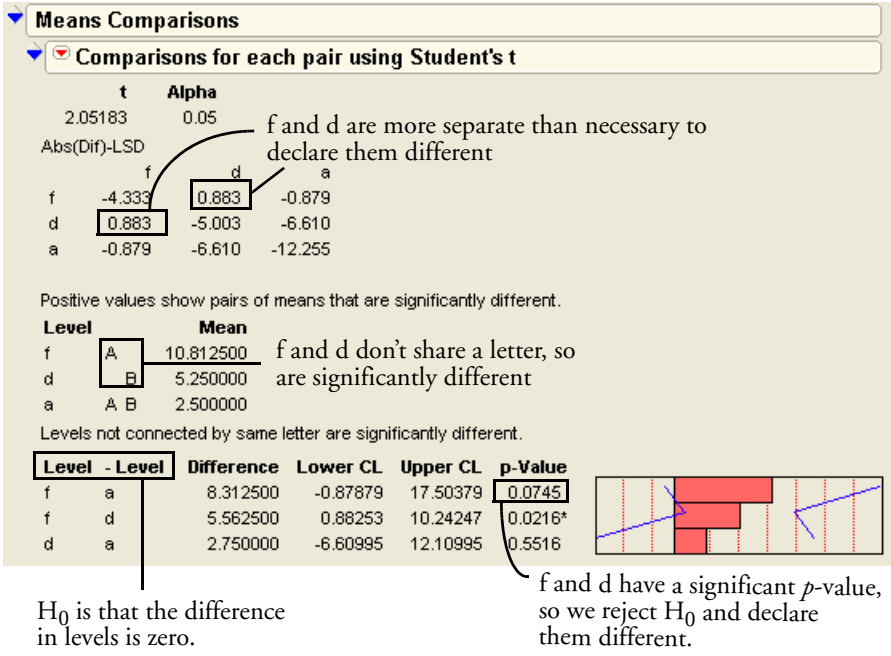
A more complex graphic is needed to show relationships due to the complexity in relationships when the data are unbalanced. The Fit Y by X platform lets you see the difference with the comparison circles and verify group differences statistically with the Means Comparisons tables shown in **Figure 9.8**.

The Means Comparisons table uses the concept of Least Significant Difference (LSD). In the balanced case, this is the separation that any two means must have from each other to be significantly different. In the unbalanced case, there is a different LSD for each pair of means.

The Means Comparison report shows all the comparisons of means ordered from high to low. The elements of the table show the absolute value of the difference in two means minus the LSD. If the means are farther apart than the LSD, they are significantly different and the element is positive. For example, the element that compares “f” and “d” is +0.88, which says the means are 0.88 more separate than needed to be significantly different. If the means are not significantly different, the LSD is greater than the difference. Therefore the element in the

table is negative. The elements for the other two comparisons are negative, showing no significant difference.

Figure 9.8 Statistical Text Reports to Compare Groups



In addition, a table shows the classic SAS-style means comparison with letters. Levels that share a letter are not significantly different from each other. For example, both levels “d” and “a” share the letter B, so d and a are not significantly different from each other.

An Ordered Differences report appears below the text reports. It lists the differences between groups in decreasing order, with confidence limits of the difference. A bar chart displays the differences with blue lines representing the confidence limits. The *p*-value tests the H_0 that there is no difference in the means.

The last thing to do in this example is to restore your copy of the Drug.jmp table to its original state so it can be used in other examples. To do this,

- Choose **File > Revert** or reopen the data table.

Adjusting for Multiple Comparisons

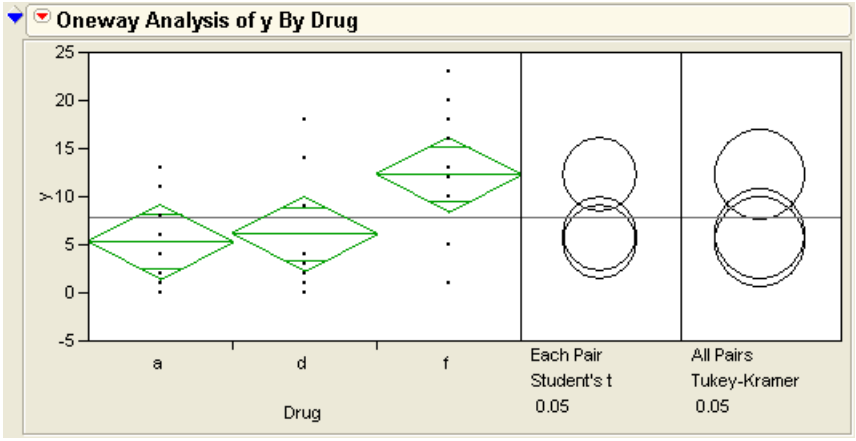
Making multiple comparisons, such as comparing many pairs of means, increases the possibility of committing a Type I error. Remember, a Type I error is the error of declaring a difference significant (based on statistical test results) that is actually not significant. We are satisfied with a 1 in 20 (5%) chance of committing a Type I error. However, the more tests you do, the more likely you are to happen upon a significant difference occurring by chance alone. If you compare all possible pairs of means in a large one-way layout with many different levels, there are many possible tests, and a Type I error becomes very likely.

There are many methods that modify tests to control for an overall error rate. This section covers one of the most basic, the *Tukey-Kramer Honestly Significant Difference* (HSD). The Tukey-Kramer HSD uses the distribution of the maximum range among a set of random variables to attempt to control for the multiple comparison problem.

- After reverting to the original copy of Drug.jmp, choose **Analyze > Fit Y by X** for the variables y as Y Drug as X. To expedite this, click the **Recall** button in the Fit Y by X dialog.
- Select the following three commands from the popup menu on the title bar: **Means/Anova**, **Compare Means > Each Pair, Student's t**, and **Compare Means > All Pairs, Tukey HSD**.

These commands should give you the results shown in **Figure 9.9**.

Figure 9.9 *t*-tests and Tukey-Kramer Adjusted *t*-tests for One-Way ANOVA



The comparison circles work as before, but have different kinds of error rates.

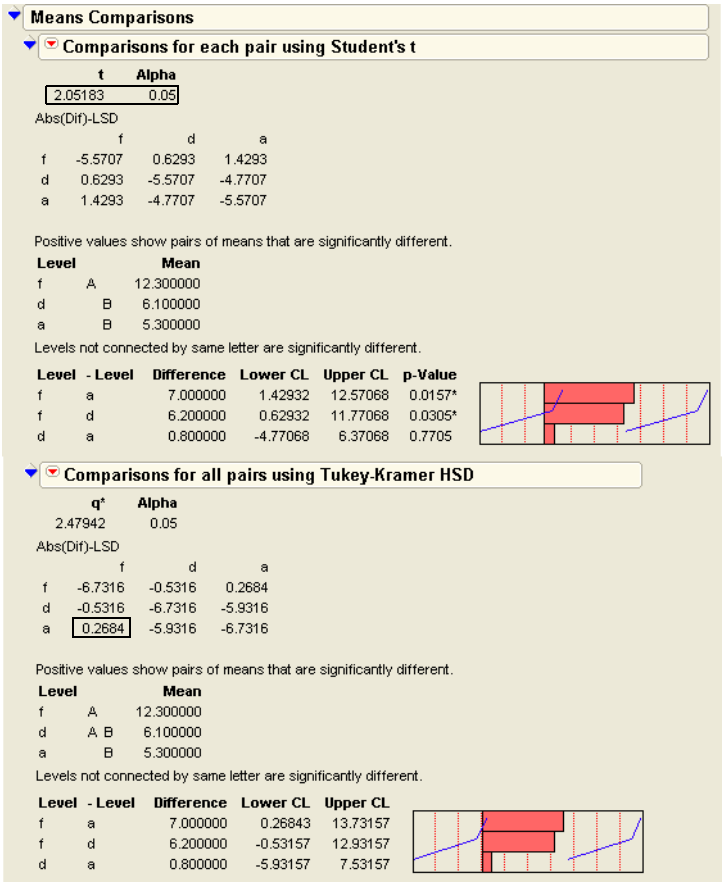
The Tukey-Kramer comparison circles are larger than the Student's *t* circles. This protects more tests from falsely declaring significance, but this protection makes it harder to declare two means significantly different.

If you click on the top circle, you see that the conclusion is different between the Student's *t* and Tukey-Kramer's HSD for the comparison of "placebo" and "d." This comparison is significant for Student's *t*-test but not for Tukey's test.

The difference in significance occurs because the quantile that is multiplied into the standard errors to create a Least Significant Difference has grown from 2.05 to 2.48 between Student's *t*-test and the Tukey-Kramer test.

The only positive element in the Tukey table is the one for the "a" versus "placebo" comparison (**Figure 9.10**).

Figure 9.10 Means Comparisons Table for One-Way ANOVA



Means Comparisons

Comparisons for each pair using Student's t

t	Alpha
2.05183	0.05

Abs(Dif)-LSD

	f	d	a
f	-5.5707	0.6293	1.4293
d	0.6293	-5.5707	-4.7707
a	1.4293	-4.7707	-5.5707

Positive values show pairs of means that are significantly different.

Level		Mean
f	A	12.300000
d	B	6.100000
a	B	5.300000

Levels not connected by same letter are significantly different.

Level	- Level	Difference	Lower CL	Upper CL	p-Value
f	a	7.000000	1.42932	12.57068	0.0157*
f	d	6.200000	0.62932	11.77068	0.0305*
d	a	0.800000	-4.77068	6.37068	0.7705

Comparisons for all pairs using Tukey-Kramer HSD

q*	Alpha
2.47942	0.05

Abs(Dif)-LSD

	f	d	a
f	-6.7316	-0.5316	0.2684
d	-0.5316	-6.7316	-5.9316
a	0.2684	-5.9316	-6.7316

Positive values show pairs of means that are significantly different.

Level		Mean
f	A	12.300000
d	A B	6.100000
a	B	5.300000

Levels not connected by same letter are significantly different.

Level	- Level	Difference	Lower CL	Upper CL
f	a	7.000000	0.26843	13.73157
f	d	6.200000	-0.53157	12.93157
d	a	0.800000	-5.93157	7.53157

Are the Variances Equal Across the Groups?

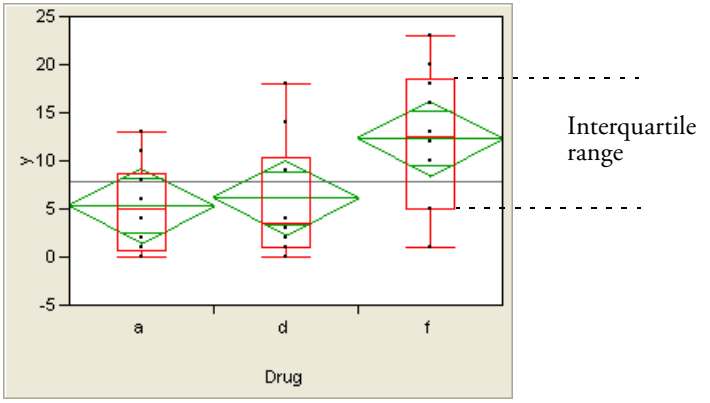
The one-way ANOVA assumes that each group has the same variance. The Analysis of Variance table shows the note "Std Error uses a pooled estimate of error variance." When testing the difference between two means, as in the previous chapter, JMP provides separate reports for both equal and unequal variance assumptions. This is why, when there are only two groups, the command is **Means/Anova/Pooled t**. ANOVA pools the variances like the pooled *t*-test does.

Before you get too concerned about the equal-variance issue, be aware that there is always a list of issues to worry about; it is not usually useful to be overly concerned about this one.

Select **Quantiles** from the red triangle popup menu showing beside the report title.

This command displays quantile box plots for each group as shown in **Figure 9.11**. Note that the interquartile range (the height of the boxes) is not much different for drugs a and d, but is somewhat different for placebos. The placebo group seems to have a slightly larger interquartile range.

Figure 9.11 Quantile Box Plots



Select **Quantiles** again to turn the box plots off.

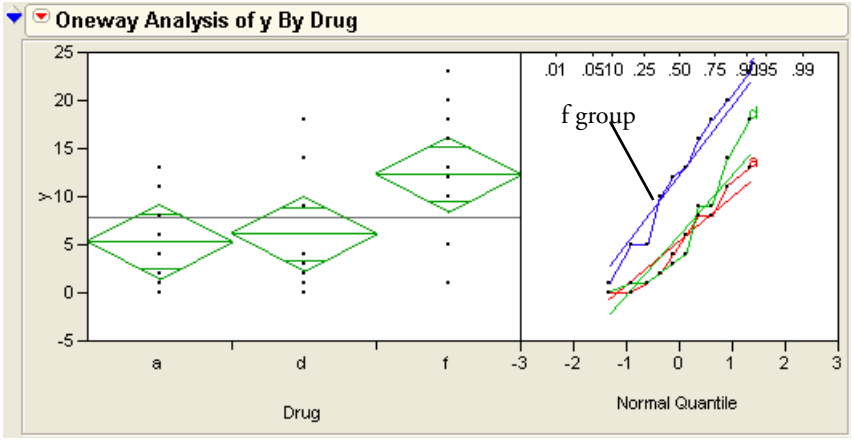
A more effective graphical tool to check the variance assumption is the Normal Quantile plot.

Select **Normal Quantile Plot > Plot Actual By Quantile** from the popup menu on the title bar.

This option displays a plot next to the Means Diamonds as shown in **Figure 9.12**. The Normal Quantile plot compares mean, variance, and shape of the group distributions.

There is a line on the Normal Quantile plot for each group. The height of the line shows the location of the group. The slope of the line shows the group's standard deviation. So, lines that appear to be parallel have similar standard deviations. The straightness of the line segments connecting the points shows how close the shape of the distribution is to the Normal distribution. Note that the "f" group is both higher and has a greater slope, which indicates a higher mean and a higher variance, respectively.

Figure 9.12 Normal Quantile Plot



It's easy to get estimates of the standard deviation within each group:

- Select **Means and Std Dev** from the red triangle popup menu showing on the report title to see the reports in **Figure 9.13**.

Figure 9.13 Mean and Standard Deviation Report

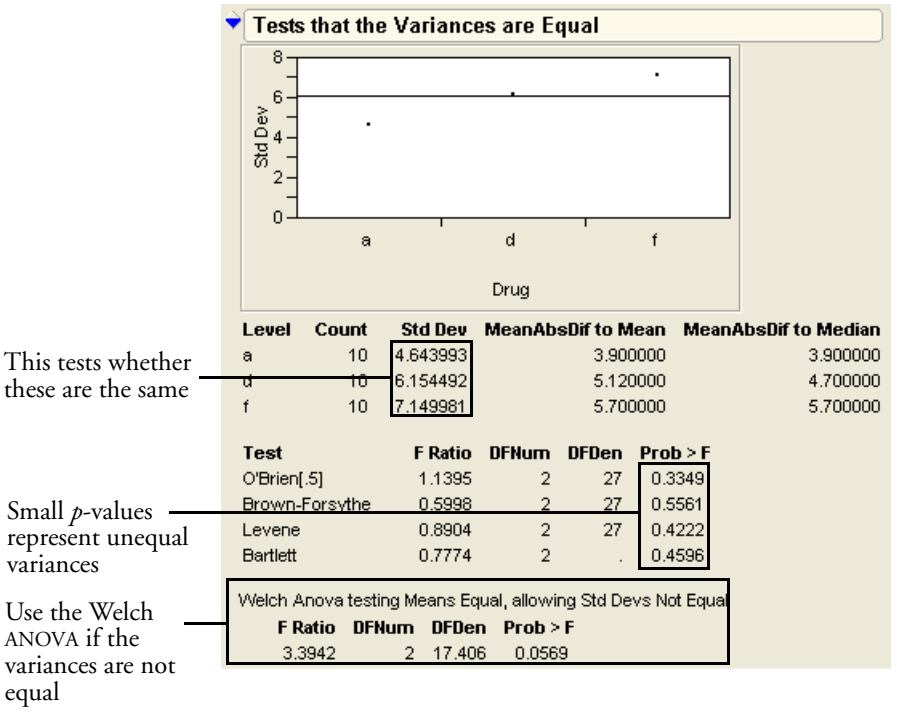
Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
a	10	5.3000	4.64399	1.4686	1.9779	8.622
d	10	6.1000	6.15449	1.9462	1.6973	10.503
f	10	12.3000	7.14998	2.2610	7.1852	17.415

You can conduct a statistical test of the equality of the variances as follows:

- Select **UnEqual Variance** from the popup menu on the title bar to see the Tests the Variances are Equal tables in **Figure 9.14**.

Figure 9.14 Tests the Variances are Equal Report



This tests whether these are the same

Small *p*-values represent unequal variances

Use the Welch ANOVA if the variances are not equal

To interpret these reports, note that the **Std Dev** column lists the estimates you are testing to be the same. The null hypothesis is that the standard deviations are equal. Then note the results listed under **Prob>F**. These small *p*-values suggest that variances are not equal.

As expected, there is no evidence that the variances are unequal. None of the *p*-values are small.

Each of the four tests in **Figure 9.14** (O'Brien, Brown-Forsythe, Levene, and Bartlett) tests whether the variances are equal, but each uses a different method for measuring variability. They all have null hypothesis that the variances are equal.

One way to evaluate dispersion is to take the absolute value of the difference of each response from its group mean. Mathematically, this means to look at $|x_i - \bar{x}|$ for each response.

- *Levene's Test* estimates the mean of these absolute differences for each group (shown in the table as **MeanAbsDif to Mean**), and then does a *t*-test (or equivalently, an *F*-test) on these estimates.

- The *Brown-Forsythe Test* measures the differences from the median instead of the mean and then tests these differences.
- *O'Brien's Test* tricks the *t*-test by telling it that the means were really variances.
- *Bartlett's Test* derives the test mathematically, using an assumption that the data are Normal. Though powerful, Bartlett's test is sensitive to departures from the Normal distribution.

Statisticians have no apologies for offering different tests with different results. Each test has its advantages and disadvantages.

Testing Means with Unequal Variances

If you think the variances are different, you should consider avoiding the standard *t*-and *F*-tests that assume the variances are equal. Instead, use the Welch ANOVA *F*-test that appears with the unequal variance tests (**Figure 9.14**). The test can be interpreted as an *F*-test in which the observations are weighted by an amount inversely proportional to the variance estimates. This has the effect of making the variances comparable.

The *p*-values may disagree slightly with those obtained from other software providing similar unequal-variance tests. These differences arise because some methods round or truncate the denominator degrees of freedom for computational convenience. JMP uses the more accurate fractional degrees of freedom.

In practice, the hope is that there are not conflicting results from different tests of the same hypothesis. However, conflicting results do occasionally occur, and there is an obligation to report the results from all reasonable perspectives.

Nonparametric Methods

JMP also offers nonparametric methods in the Fit Y by X platform. Nonparametric methods, introduced in the previous chapter, use only the rank order of the data and ignore the spacing information between data points. Nonparametric tests do not assume the data have a Normal distribution. This section first reviews the rank-based methods, then generalizes the Wilcoxon rank-sum method to the *k* groups of the one-way layout.

Review of Rank-Based Nonparametric Methods

Nonparametric tests are useful to test whether means or medians are the same across groups. However, the usual assumption of Normality is not made. Nonparametric tests use functions of the response ranks, called *rank scores* (Hajek 1969).

JMP offers the following nonparametric tests for testing the null hypothesis that distributions across factor levels are centered at the same location. Each is the most powerful rank test for a certain distribution, as indicated in Table 9.1.

- Wilcoxon rank scores are the ranks of the data.
- Median rank scores are either 1 or 0 depending on whether a rank is above or below the median rank.
- Van der Waerden rank scores are the quantiles of the standard Normal distribution for the probability argument formed by the rank divided by n–1. This is the same score that is used in the Normal quantile plots.

Table 9.1. Guide for Using Nonparametric Tests

Fit Y By X Analysis Option	Two Levels	Two or More Levels	Most Powerful for Errors Distributed as
Nonpar-Wilcoxon	Wilcoxon rank-sum (Mann-Whitney *U*)	Kruskal-Wallis	Logistic
Nonpar-Median	Two-Sample Median	*k*-Sample Median (Brown-Mood)	Double Exponential
Nonpar-VW	Van der Waerden	*k*-sample Van der Waerden	Normal

The Three Rank Tests in JMP

As an example, use the Drug.jmp example and request nonparametric tests to compare the Drug group means of LBS:

- Choose **Analyze > Fit Y by X** with y as **Y** and Drug as **X**.

The one-way analysis of variance platform appears showing the distributions of the three groups, as seen previously in **Figure 9.3**.

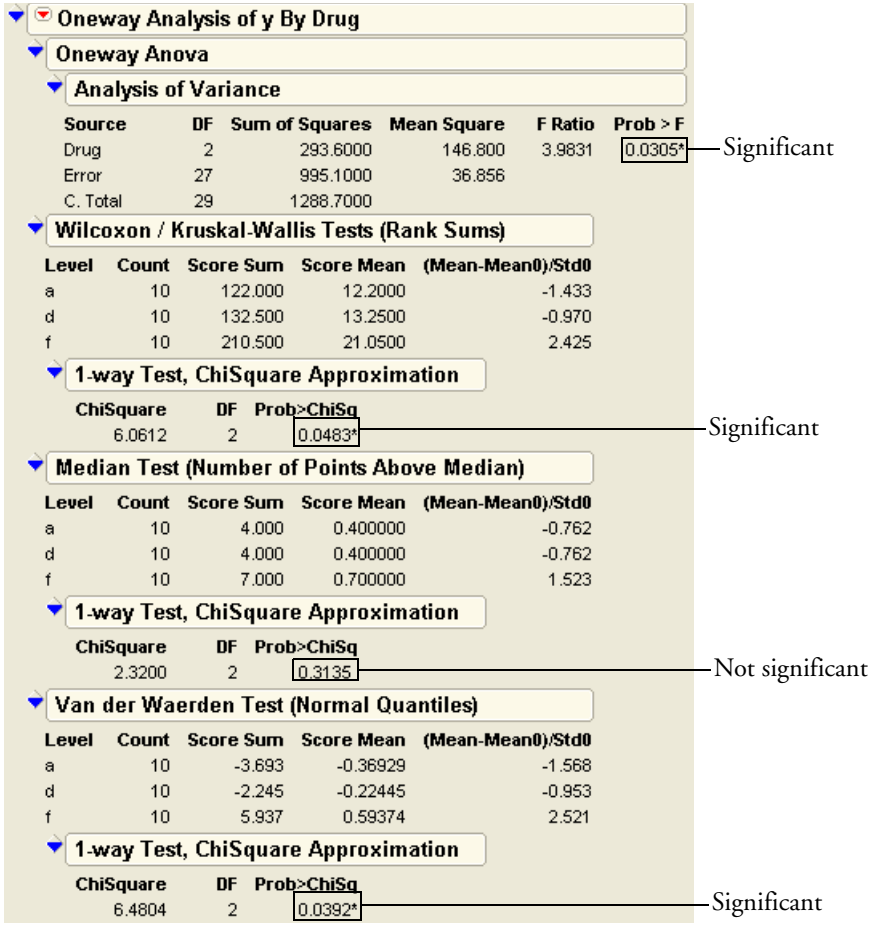
- In the popup menu on the title bar of the scatterplot, select the four tests that compare groups, all of which have the null hypothesis that the groups do not differ:

- **Means/Anova**, producing the *F*-test from the standard parametric approach
- **Nonparametric > Wilcoxon Test**, also known as the Kruskal-Wallis test when there are more than two groups
- **Nonparametric > Median Test** for the median test
- **Nonparametric > Van der Waerden Test** for the Van der Waerden test

Figure 9.15 shows the results of the four tests that compare groups. In this example, the Wilcoxon and the Van der Waerden agree with the parametric *F*-test in the ANOVA and show borderline significance for a 0.05 α-level, despite a fairly small sample and the possibility that the data are not Normal. These tests reject the null and detect a difference in the groups.

The median test is much less powerful than the others and does not detect a difference in this example.

Figure 9.15 Parametric and Non-Parametric Tests for Drug Example



Exercises

1. This exercise uses the Movies.jmp data set. You are interested in discovering if there is a difference in earnings between the different classifications of movies.

   (a) Use the Distribution platform to examine the variable Type. How many levels are there? Are there roughly equal numbers of movies of each type?

   (b) Use the Fit Y by X platform to perform an ANOVA, with Type as X and Worldwide $ as Y. State the null hypothesis you are testing. Does the test show differences among the different types?

   (c) Use comparison circles to explore the differences. Does it appear that Action and Drama are significantly different than all other types?

   (d) Examine a Normal Quantile Plot of this data and comment on the equality of the variances of the groups. Are they different enough to require a Welch ANOVA? If so, conduct one and comment on its results.

2. The National Institute of Standards and Technology (NIST) references research involving the doping of silicon wafers with phosphorus. The wafers were doped with phosphorus by neutron transmutation doping in order to have nominal resistivities of 200 ohm/cm. Each data point is the average of six measurements at the center of each wafer. Measurements of bulk resistivity of silicon wafers were made at NIST with five probing instruments on each of five days, with the data stored in the table Doped Wafers.jmp (see Ehrstein and Croarkin). The experimenters are interested in testing differences among the instruments.

   (a) Examine a histogram of the resistances reported for the instruments. Do the data appear to be Normal? Test for Normality by stating the null hypothesis, and then conduct a statistical test.

   (b) State the null hypothesis and conduct an ANOVA to determine if there is a difference between the probes in measuring resistance.

   (c) Comment on the sample sizes involved in this investigation. Do you feel confident with your results?

3. The data table Michelson.jmp contains data (as reported by Stigler, 1977) collected to determine the speed of light in air. Five separate collections of data were made in 1879 by Michelson and the speed of light was recorded in km/sec. The values for velocity in this table have had 299,000 subtracted from them.

(a) The true value (accepted today) for the speed of light is 299,792.5 km/sec. What is the mean of Michelson's responses?

(b) Is there a significant statistical difference between the trials? Use an ANOVA or a Welch ANOVA (whichever is appropriate) to justify your answer.

(c) Using Student's *t* comparison circles, find the group of observations that is statistically different from all the other groups.

(d) Does excluding the result in part (c) improve Michelson's prediction?

4. *Run-Up* is a term used in textile manufacturing to denote waste. Manufacturers often use computers to lay out designs on cloth in order to minimize waste, and the percentage difference between human layouts and computer-modeled layouts is the run-up. There are some cases where humans get better results than the computers, so don't be surprised if there are a few negative values for run-up in the table Levi Strauss Run-Up.jmp (Koopmans, 1987). The data was gathered from five different supplier plants to determine if there were differences among the plants.

   (a) Produce histograms for the values of Run Up for each of the five plants.

   (b) State a null hypothesis, and then test for differences between supplier plants by using the three non-parametric tests provided by JMP. Do they have similar results?

   (c) Compare these results to results given by an ANOVA and comment on the differences. Would you trust the parametric or nonparametric tests more?

5. The data table Scores.jmp contains a subset of results from the *Third International Mathematics and Science Study*, conducted in 1995. The data contain information on scores for Calculus and Physics, divided into four regions of the country.

   (a) Is there a difference among regions in Calculus scores?

   (b) Is there a difference among regions for Physics scores?

   (c) Do the data fall into easily definable groups? Use comparison circles to explore their groupings.

6. Three brands of typewriters were tested for typing speed by having expert typists type identical passages of text. The results are stored in Typing Data.jmp.

   (a) Complete an analysis of variance comparing the typing speed of the experts on each brand of typewriter.

   (b) Does one typewriter brand yield significantly higher speeds than the other two? Defend your answer.

7. A manufacturer of widgets determined the quality of its product by measuring abrasion on samples of finished products. The manufacturer was concerned that there was a difference in the abrasion measurement for the two shifts of workers that were employed at the factory. Use the data stored in Abrasion.jmp to compute a *t*-test of abrasion comparing the two shifts. Is there statistical evidence for a difference?

8. The manufacturers of a medication were concerned about adverse reactions in patients treated with their drug. Data on adverse reactions is stored in AdverseR.jmp. The duration of the adverse reaction is stored in the ADR DURATION variable.

   (a) Patients given a placebo are noted with PBO listed in the treatment group variable, while those that received the standard drug regimen are designated with ST_DRUG. Test whether there is a significant difference in adverse reaction times between the two groups.

   (b) Test whether there is a difference in adverse reaction times based on the sex (gender) of the patient.

   (c) Test whether there is a difference in adverse reaction time based on the race of the patient.

   (d) Redo the analyses in parts (a) through (c) using a nonparametric test. Do the results differ?

   (e) A critic of the study claims that the weights of the patients in the placebo group are not the same as those of the treatment group. Do the data support the critic's claim?

9. To judge the efficacy of a three pain relievers, a consumer group conducted a study of the amount of relief each patient received. The amount of relief is measured by each participant rating the amount of relief on a scale of 0 (no relief) to 20 (complete relief). Results of the study are stored in the file Analgesics.jmp.

   (a) Is there a difference in relief between the males and females in the study? State a null hypothesis, a test statistic, and a conclusion.

   (b) Conduct an analysis of variance comparing the three types of drug. Is there a significantly significant difference among the three types of drug?

   (c) Find the mean amount of pain relief for each of the three types of pain reliever.

   (d) Does the amount of relief differ for males and females? To investigate this question, conduct an analysis of variance on relief vs. treatment for each of the two genders. (Conduct a separate analysis for males and females. Assign gender as a By-variable.) Is there a significant difference in relief for the female subset? For the male subset?

10 Fitting Curves through Points: Regression

Overview

Regression is a method of fitting curves through data points. It is a straightforward and useful technique, but people new to statistics often ask about its rather strange name—regression.

Sir Francis Galton, in his 1885 Presidential address before the anthropology section of the British Association for the Advancement of Science (Stigler, 1986), described his study of how tall children are compared with the height of their parents. In this study, Galton defined ranges of parents' heights, and then calculated the mean child height for each range. He drew a straight line that went through the means (as best as he could), and thought he had made a discovery when he found that the child heights tended to be more moderate than the parent heights. For example, if a parent was tall, the children similarly tended to be tall, but not as tall as the parent. If a parent was short, the child tended to be short, but not as short as the parent. This discovery he called *regression to the mean*, with the regression meaning "to come back to."

Somehow, the term regression became associated with the technique of fitting the line, rather than the process describing inheritance. Galton's data are covered later in this chapter.

This chapter covers the case where there is only one factor and one response—the kind of regression situation that you can see on a scatterplot.

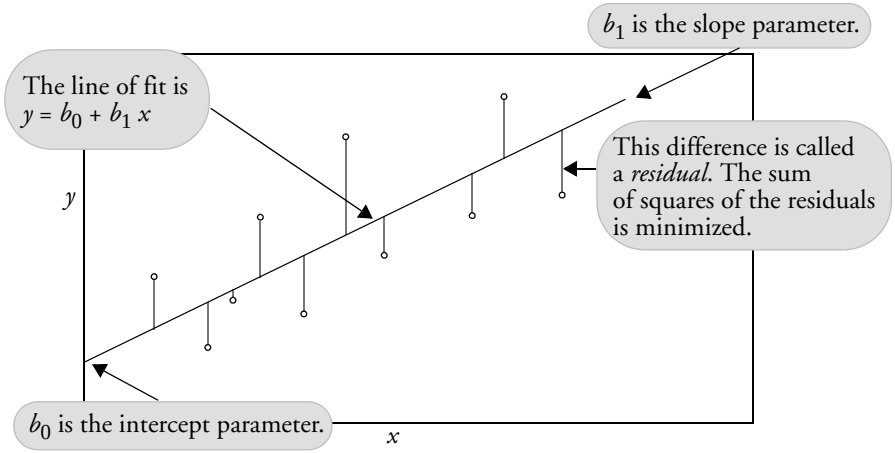
Regression

Fitting one mean is easy. Fitting several means is not much harder. How do you fit a mean when it changes as a function of some other variable? In essence, how do you fit a line or a curve through data?

Least Squares

In regression, you pick an equation type (linear, polynomial, and so forth) and allow the fitting mechanism to determine some of its parameters (coefficients). These parameters are determined by the method of least squares, which finds the parameter values that minimize the sum of squared distances from each point to the line of fit. **Figure 10.1** illustrates a least squares regression line.

Figure 10.1 Straight-Line Least Squares Regression



For any regression model, the term *residual* denotes the difference between the actual response value and the value predicted by the line of fit. When talking about the true (unknown) model rather than the estimated one, these differences are called the *errors* or *disturbances*.

Least squares regression is the method of fitting of a model to minimize the sum of squared residuals.

The regression line has interesting balancing properties with regard to the residuals. The sum of the residuals is always zero, which was also true for the simple mean fit. You can think of the fitted line as balancing data in the up-and-down direction. If you add the product of the residuals times the x (regressor) values, this sum is also zero. This can be interpreted as the line balancing the data in a rotational sense. Chapter 21, "Machines of Fit," shows how these least

squares properties can be visualized in terms of the data acting like forces of springs on the line of fit.

An important special case is when the line of fit is constrained to be horizontal (flat). The equation for this fit is a constant; if you constrain the slope of the line to zero, the coefficient of the *x* term (regressor) is zero, and the *x* term drops out of the model. In this situation, the estimate for the constant is the sample mean. This special case is important because it leads to the statistical test of whether the regressor really affects the response.

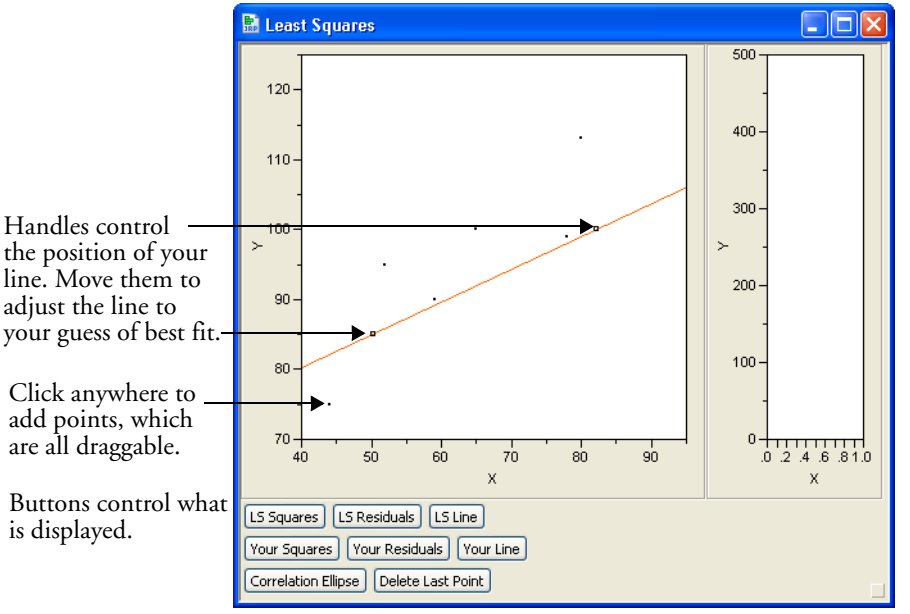
Seeing Least Squares

The principle of least squares can be seen with one of the sample scripts included in the Sample Scripts folder.

- Open and run the demoLeastSquares.jsl script.

Opening and running scripts is covered in "Working with Scripts" on page 57.

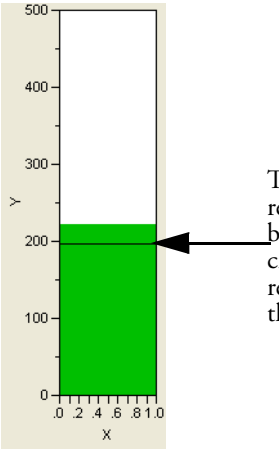
Figure 10.2 demoLeastSquares Display



There is a line in the scatterplot with two small squares on it. These two rectangular handles are draggable, and are used in this case to move the line to a position that you think best summarizes the data.

- Press the **Your Residuals** button. Use the handles to move the line around until you think the residuals (in blue) are as small as they can be.
- Press the **Your Squares** button and try to minimize the total area covered by the blue squares.

To assist you in minimizing the area of the squares, a second graph is displayed to the right of the scatterplot. Think of it as a "thermometer" that measures the sum of the area of the squares in the scatterplot. The least squares criterion selects the line that minimizes this area. To see the least squares line as calculated by JMP,



- Press the button labeled **LS Line** to display the least squares line.
- Press the buttons **LS Residuals** and **LS Squares** to display the residuals and squares for the least squares line.

Notice that a horizontal line has been added in the graph that displays the sum of squares. This represents the sum of the squared residuals from the line calculated by the least squares criterion.

To illustrate that the least squares criterion performs as it claims to:

- Using the handles, drag your line so that it coincides with the least squares line.

The sum of squares is now the same as the sum calculated by the least squares criterion.

- Using one of the handles, move your line off of the least squares line, first in one direction, then in the other.

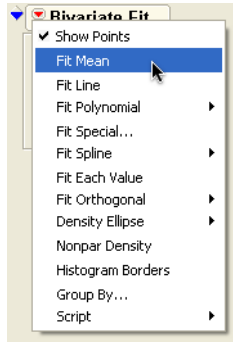
Notice that as your line moves off the line of least squares in any way, the sum of the squares increases. Therefore, the least squares line is truly the line that minimizes the sum of the squared residuals.

Fitting a Line and Testing the Slope

Eppright *et al.* (1972) as reported in Eubank (1988) measured 72 children from birth to 70 months. You can use regression techniques to examine how the weight-to-height ratio changes as kids grow up.

- Open the Growth.jmp sample table, which holds the Eppright data.
- Choose **Analyze > Fit Y by X** and select ratio as Y and age as X. When you click **OK**, the result is a scatterplot of ratio by age.

Look for the triangular popup menu icon on the title bar of the plot. Click this icon to see fitting commands for adding regression fits to the scatterplot.



- Select **Fit Mean** and then **Fit Line** from the fitting popup menu.

- **Fit Mean** draws a horizontal line at the mean of ratio.
- **Fit Line** draws the regression line through the data.

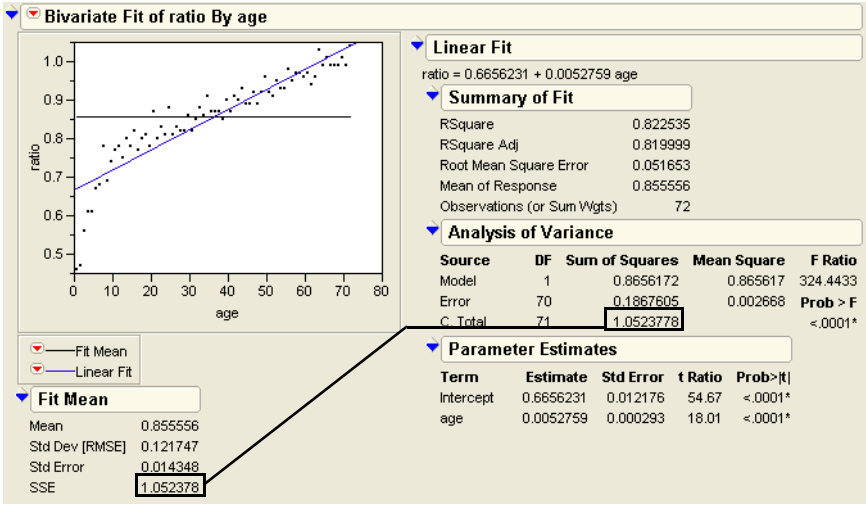
These commands also add statistical tables to the regression report. You should see a scatterplot similar to the one shown in **Figure 10.3**. The statistical tables are actually displayed beneath the scatterplot in the report window, but have been rearranged here to save space.

Each kind of fit you select has its own popup menu icon (found under the scatterplot) that lets you request fitting details.

The parameter estimates report shows the estimated coefficients of the regression line. The fitted regression is the equation

$$\text{ratio} = 0.6656 + 0.005276\ \text{age} + \text{residual}$$

Figure 10.3 Straight-line Least-squares Regression



Testing the Slope by Comparing Models

If we assume that the linear equation is adequate to describe the relationship of the weight to height ratio with growth (which turns out to be incorrect), we have some questions to answer:

- Does the regressor really affect the response?
- Does the ratio of weight to height change as a function of age? Is the true slope of the regression line zero?
- Is the true value for the coefficient of age in the model zero?
- Is the sloped regression line significantly different than the horizontal line at the mean?

Actually, these are all the same question.

"The Difference between Two Means" on page 167 presented two analysis approaches that turned out to be equivalent. One approach used the distribution of the estimates, which resulted in the *t*-test. The other approach compared the sum of squared residuals from two models where one model was a special case of the other. This model comparison approach resulted in an *F*-test. In regression, there are the same two equivalent approaches: (1) distribution of estimates and (2) model comparison.

The model comparison is between the regression line and what the line would be if the slope were constrained to be zero; that is, you compare the fitted regression line with the horizontal line at the mean. This comparison is our null hypothesis (the slope = 0). If the regression line is a better fit than the line at the mean, then the slope of the regression line is significantly

different from zero. This is often stated negatively: “If the regression line does not fit better than the horizontal fit, then the slope of the regression line does not test as significantly different from zero.”

The *F*-test in the Analysis of Variance table is the comparison that tests the null hypothesis of the slope of the fitted line. It compares the sum of squared residuals from the regression fit to the sum of squared residuals from the sample mean. **Figure 10.4** diagrams the relationship between the quantities in the statistical reports and corresponding plot.

Here are descriptions of the quantities in the statistical tables:

C Total

corresponds to the sum of squares error if you had fit only the mean. You can verify this by looking at the Fit Mean table from in the previous example. The C. Total sum of squares (SSE in the Fit Mean report) is 1.0524 for both the mean fit and the line fit.

Error

is the sum of squared residuals after fitting the line, 0.1868. This is sometimes casually referred to as the residual, or residual error. You can think of Error as leftover variation—variation that didn’t get explained by fitting a model.

Model

is the difference between the error sum of squares in the two models (the horizontal mean and the sloped regression line). It is the sum of squares resulting from the regression, 0.8656. You can think of Model as a measure of the variation in the data that was explained by fitting a regression line.

Mean Square

is a sum of squares divided by its respective degrees of freedom. The Mean Square for Error is the estimate of the error variance (0.002668 in this example).

Root Mean Square Error

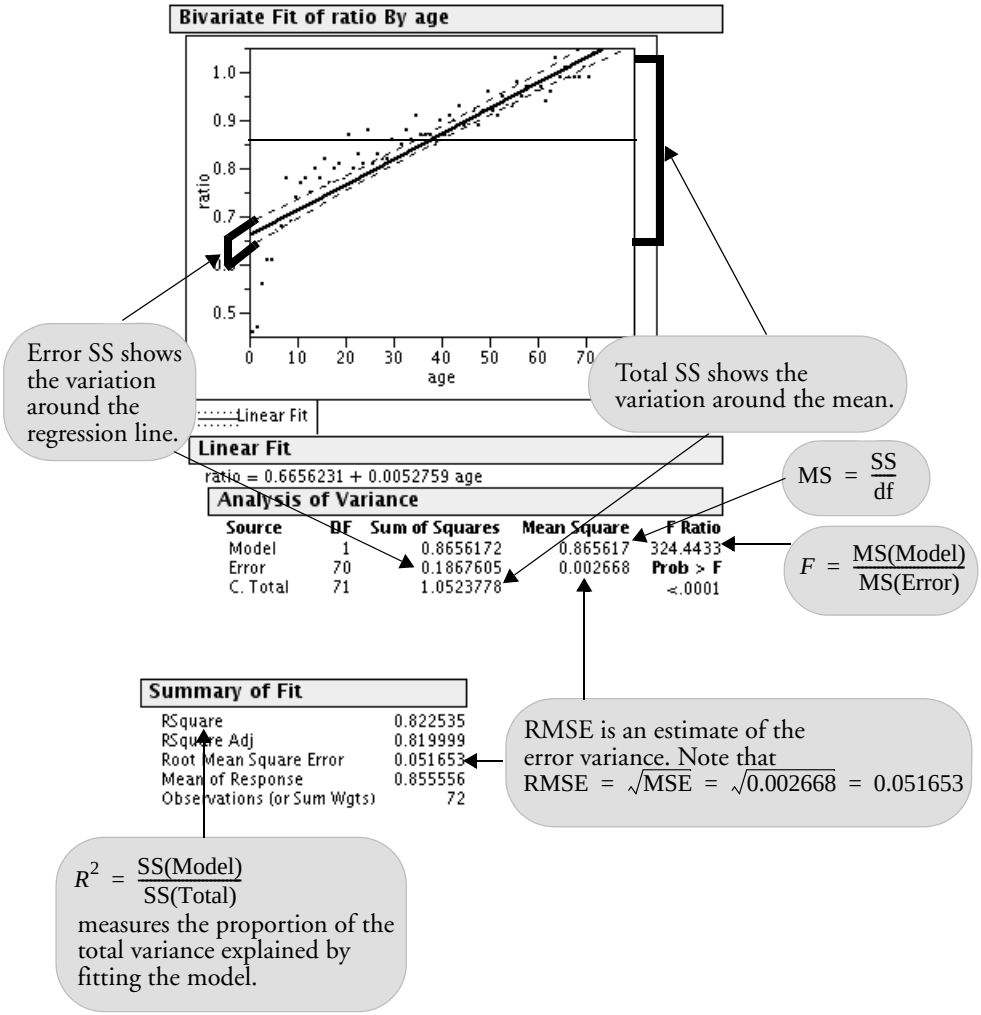
is found in the Summary of Fit report. It estimates the error standard deviation of the error and is calculated as the square root of the Mean Square for Error.

If the true regression line has a slope of zero, then the model isn't explaining any of the variation. The model mean square and the error mean square would both estimate the residual error variance, and therefore have the same expected value.

The *F*-statistic is calculated as the model mean square divided by the error mean square. If the model and error both have the same expected value, the *F*-statistic is 1. However, if the model mean square is larger than the error mean square, you suspect that the slope is not zero and

the model is explaining some variation. The *F*-ratio has an *F*-distribution under the null hypothesis that (in this example) age has no effect on ratio.

Figure 10.4 Diagram to Compare Models



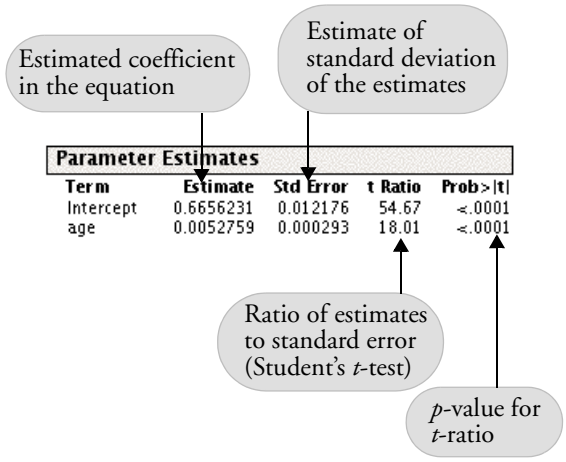
The Distribution of the Parameter Estimates

The formula for a simple straight line only has two parameters, the intercept and the slope. For this example, the model can be written as follows:

$$\text{ratio} = b_0 + b_1 \text{ age} + \text{residual}$$

where b_0 is the intercept and b_1 is the slope. In our example, b_0 is the intercept and age is the slope.

The Parameter Estimates Table also shows these quantities:



Std Error

is the estimate of the standard deviation attributed to the parameter estimates.

t-Ratio

is a test that the true parameter is zero. The *t*-ratio is the ratio of the estimate to its standard error. Generally, you are looking for *t*-ratios that are greater than 2 in absolute value, which usually correspond to significance probabilities of less than 0.05.

Prob>|t|

is the significance probability (*p*-value). You can translate this as "the probability of getting an even greater absolute *t* value by chance alone if the true value of the slope is zero."

Note that the *t*-ratio for the age parameter, 18.01, is the square root of the *F*-Ratio in the Analysis of Variance table, 324.44. You can double-click on the *p*-values in the tables to show more decimal places, and see that the *p*-values are exactly the same. This is not a surprise—the *t*-test for simple regression is testing the same null hypothesis as the *F*-test.

Confidence Intervals on the Estimates

There are several ways to look at the significance of the estimates. The *t*-tests for the parameter estimates, discussed previously, test that the parameters are significantly different from zero. A more revealing way to look at the estimates is to obtain confidence limits that show the range of likely values for the true parameter values.

- Beside **Linear Fit**, select the **Confid Curves: Fit** command from the popup menu.

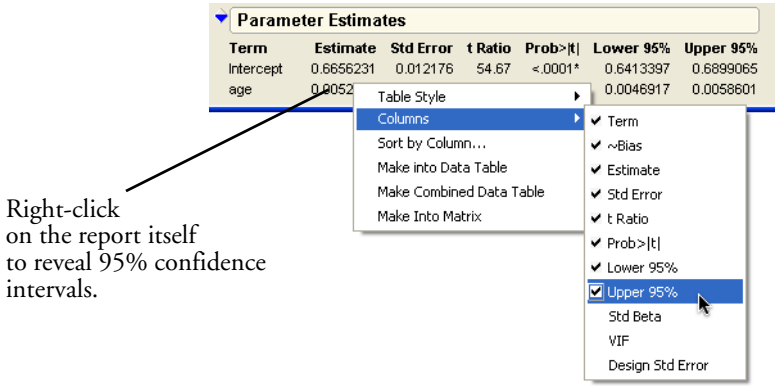
This command adds the confidence curves to the graph, as illustrated previously in **Figure 10.4**.

The 95% confidence interval is the smallest interval whose range includes the true parameter values with 95% confidence. The upper and lower confidence limits are calculated by adding and subtracting respectively the standard error of the parameter times a quantile value corresponding to a (0.05)/2 Student's *t*-test.

Another way to find the 95% confidence interval is to examine the Parameter Estimates tables. Although the 95% confidence interval values are initially hidden in the report, they can be made visible.

- Right-click on the report and select **Columns > Lower 95%** and **Columns > Upper 95%**, as shown in **Figure 10.5**.

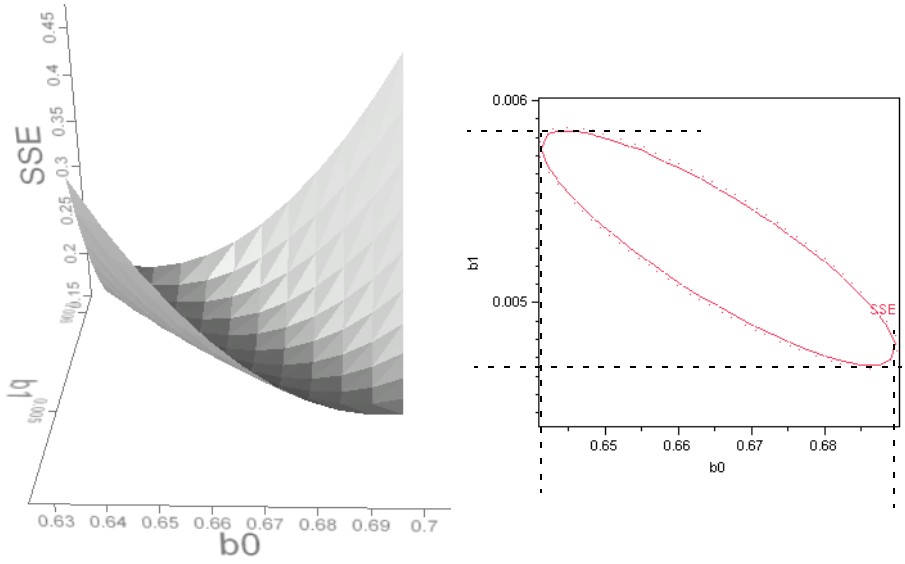
Figure 10.5 Add Confidence Intervals to Table



An interesting way to see this concept is to look from the point of view of the sum of squared errors. Imagine the sum of squared errors (SSE) as a function of the parameter values, so that as you vary the slope and intercept values you calculate the corresponding SSE. The least squares estimates are where this surface is at a minimum.

The left plot in **Figure 10.6** shows a three-dimensional view of this interpretation for the growth data regression problem. The 3-D view shows the curvature of the SSE surface as a function of the parameters, with a minimum in the center. The *x*- and *y*-axes are a grid of parameter values and the *z*-axis is the computed SSE for those values.

Figure 10.6 Representation of Confidence Limit Regions



One way to form a 95% confidence interval is to turn the F-test upside down. You take an F value that would be the criterion for a 0.05 test (3.97), multiply it by the MSE, and add that to the SSE. This gives a higher SSE of 0.19737 and forms a confidence region for the parameters. Anything that produces a smaller SSE is believable because it corresponds to an *F*-test with a *p*-value greater than 0.05.

The 95% confidence region is the inner elliptical shape in the plot on the right in **Figure 10.6**. The flatness of the ellipse corresponds to the amount of correlation of the estimates. You can look at the plot to see what parameter values correspond to the extremes of the ellipse in each direction.

- The horizontal scale corresponds to the intercept parameter. The confidence limits are the positions of the vertical tangents to the inner contour line, indicating a low point of 0.6413 and high point of 0.6899.
- The vertical scale corresponds to the slope parameter for age. The confidence limits are the positions of the vertical tangents to the inner contour line, indicating a low of 0.00469 and high point of 0.00586. These are the lower and upper 95% confidence limits for the parameters.

You can verify these numbers by looking at the confidence limits in **Figure 10.5**.

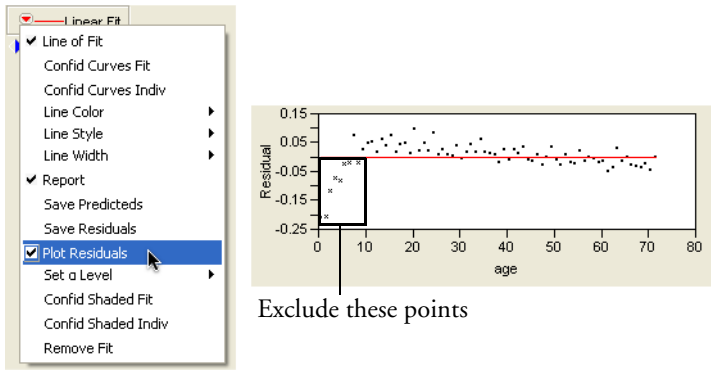
Examine Residuals

It is always a good idea to take a close look at the residuals from a regression (the difference between the actual values and the predicted values):

- Select **Plot Residuals** from the **Linear Fit** popup menu beneath the scatterplot (**Figure 10.7**).

This command appends the residual plot shown in **Figure 10.7** to the bottom of the regression report.

Figure 10.7 Scatterplot to Look at Residuals



The picture you usually hope to see is the residuals scattered randomly about a mean of zero. So, in residual plots like the one shown in **Figure 10.7**, you are looking for patterns and for points that violate this random scatter. This plot is suspicious because the left side has a pattern of residuals below the line. These points influence the slope of the regression line **(Figure 10.3**), pulling it down on the left. You can see what the regression would look like without these points by excluding them from the analysis.

Exclusion of Rows

To exclude points (rows) from an analysis, you highlight the rows and assign them the **Exclude** row state characteristic as follows.

- Get the Brush tool from the **Tools** menu or toolbar.
- Shift-drag the brush (which shows on the plot as a stretch rectangle) to highlight the points at the lower left of the plot (indicated by an x marker in **Figure 10.7**).
- Choose **Rows > Exclude / Include**.

You then see a *do not use* ( ) sign on those row number areas of the data grid.

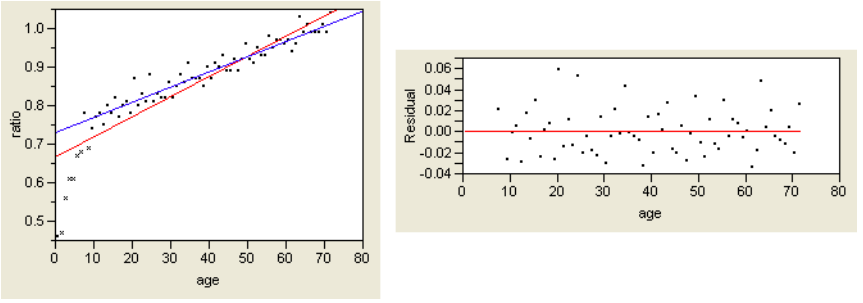
- Click the original scatterplot window and select **Remove Fit** from the **Fit Mean** popup menu to clean up the plot by removing the horizontal line.
- Again select **Fit Line** from the Bivariate menu to overlay a regression line with the lower age points excluded from the analysis.

The plot in **Figure 10.8** shows the two regression lines. Note that the new line of fit appears to go through the bulk of the points better, ignoring the points at the left that are excluded.

- To see the residuals plot for the new regression line, select **Plot Residuals** from the second Linear Fit popup menu.

Note in **Figure 10.8** that the residuals no longer have a pattern to them.

Figure 10.8 Regression with Extreme Points Excluded



Time to Clean Up

The next example uses this scatterplot, so let's clean it up:

- Choose **Row > Clear Row States**.

This removes the Excluded row state status from all points so that you can use them in the next steps.

- To finish the clean up, select **Remove Fit** from the second **Linear Fit** popup menu to remove the example regression that excluded outlying points.

Polynomial Models

Rather than excluding some points, let's try fitting a different model altogether— a quadratic curve. This is a simple curve; it only adds one term to the linear model we've been using, a term for the squared value of the regressor, age:

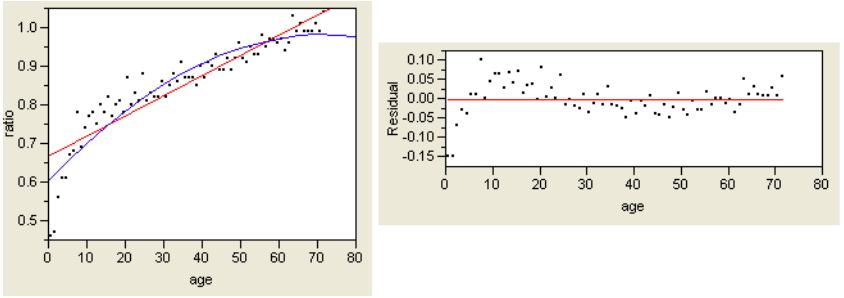
$$\text{ratio} = b_0 + b_1 \text{ age} + b_2 \text{ age}^2 + \text{residual}$$

To fit a quadratic curve to ratio by age:

- From the Bivariate report red triangle menu, select **Fit Polynomial** > **2, quadratic**.

The left plot in **Figure 10.9** shows the best-fitting quadratic fit and linear fit overlaid on the scatterplot. You can compare the straight line and curve, and also compare them statistically with the Analysis of Variance reports that show beneath the plot.

Figure 10.9 Comparison of Linear and Second-Order Polynomial Fits



Look at the Residuals

To examine the residuals,

- Select **Plot Residuals** from the **Polynomial Fit degree=2** popup menu.

You should see the plot on the right in **Figure 10.9**. There still appears to be a pattern in the residuals, so you might want to continue and fit a model with higher order terms.

Higher-Order Polynomials

To give more flexibility to the curve, specify higher-order polynomials, adding a term to the third power, to the fourth power, and so on.

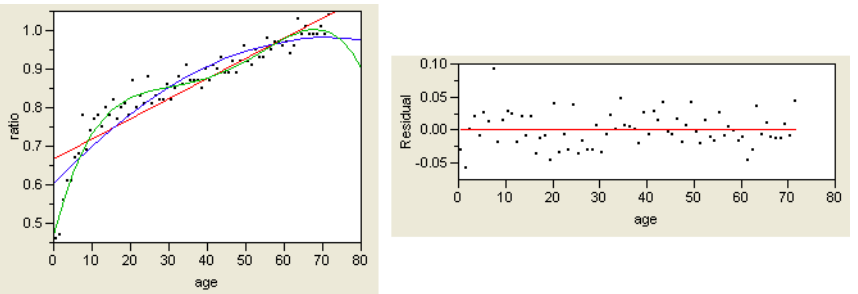
- With the scatterplot active, request a polynomial of degree 4 from the popup menu on the scatterplot title bar.

This plots the curve with linear, quadratic, cubic, and quartic terms, not simply a 4th power term. Plotting polynomials always includes lower-order terms.

- Then, select **Plot Residuals** from the **Polynomial Fit degree=4** options menu, which gives the plot on the right in **Figure 10.10**.

Note that the residuals no longer appear to have a pattern to them.

Figure 10.10 Comparison of Linear and Fourth-Order Polynomial Fits



Distribution of Residuals

It is also informative to look at the shape of the distribution of the residuals. If the distribution departs dramatically from the Normal, then you may be able to find further phenomena in the data.

Figure 10.11 shows histograms of residuals from the linear fit, the quadratic fit, and the quartic (4th degree) fit. You can see the distributions evolve toward Normality for the better-fitting models with more parameters.

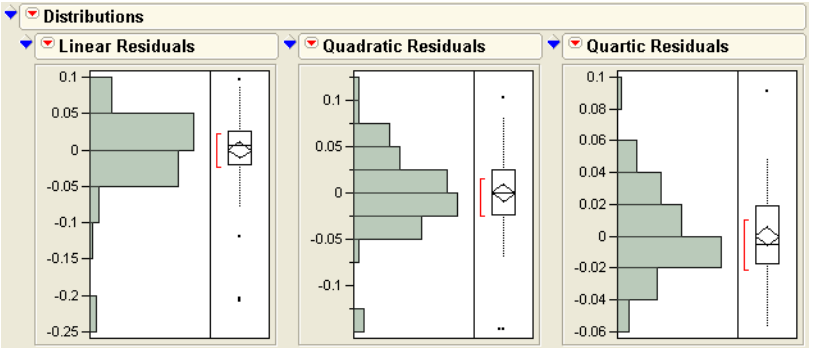
To generate these histograms,

- Select the **Save Residuals** command found in the popup menu for each regression fit.

This forms three new columns in the data table.

- Choose **Analyze > Distribution** and assign each of the three new columns of residual values as **Y, Columns**.
- Click **OK**.

Figure 10.11 Histograms for Distribution of Residuals



Transformed Fits

Sometimes, you can find a better fit if you transform either the Y or X variable (or sometimes both). When we say transform, we mean that we apply a mathematical function to it and examine these new values. For example, rather than looking at *x*, we may examine log *x*. One way of doing this is to create a new column in the data table and compute the log of age, then use **Analyze > Fit Y by X** to do a straight-line regression of ratio on this transformed age. Results from this method are shown on the right in **Figure 10.12**.

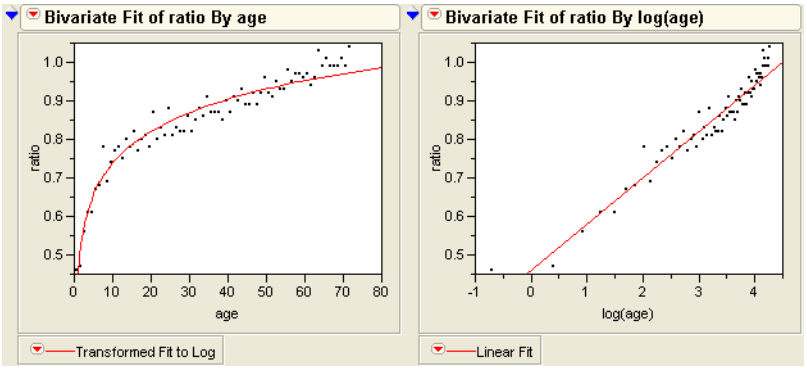
Alternatively, you can use the Fit Y by X platform to do this directly:

- Choose **Analyze > Fit Y by X** for ratio by age.
- Select **Fit Special** from the popup menu on the scatterplot title bar.

The **Fit Special** command displays a dialog that lists natural log, square, square root, exponential, and other transformations, as selections for both the X and Y variables. Try fitting ratio to the log of age:

- Click the **Natural Logarithm: (log x)** radio button for X. When you click **OK**, you see the left plot in **Figure 10.12**.

Figure 10.12 Comparison of Fits



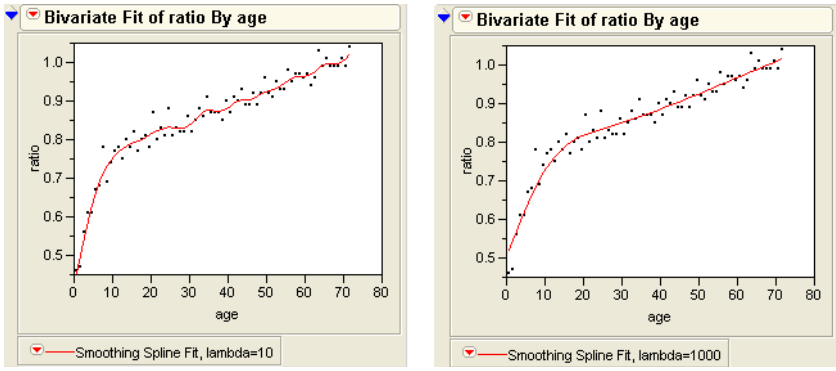
If you transform the Y variable, you can't compare the R^2 and error sums of squares of the transformed variable fit with the untransformed variable fit. You are fitting a different variable.

Spline Fit

It would be nice if you could fit a flexible leaf spring through the points. The leaf spring would resist bending somewhat, but would take gentle bends when it's needed to fit the data better. A smoothing spline is exactly that kind of fit. With smoothing splines, you specify how stiff to make the curve. If the spline is too rigid it looks like a straight line, but if it is too flexible it curves to try to fit each point. Use these commands to see the spline plots in **Figure 10.13**:

- Choose **Analyze > Fit Y by X** for ratio and age and then select **Fit Spline** from the popup menu with lambda=**10**.
- Select **Fit Spline** from the popup menu again, but with lambda=**1000**.

Figure 10.13 Comparison of Less Flexible and More Flexible Spline Fits



Are Graphics Important?

Some statistical packages don't show graphs of the regression, while others require you to make an extra effort to see a regression graph. The following example shows the kind of phenomena that you miss if you don't examine a graph.

- Open Anscombe.jmp (Anscombe, 1973).
- Click the script named **The Quartet** and select **Run Script** from the menu that appears.

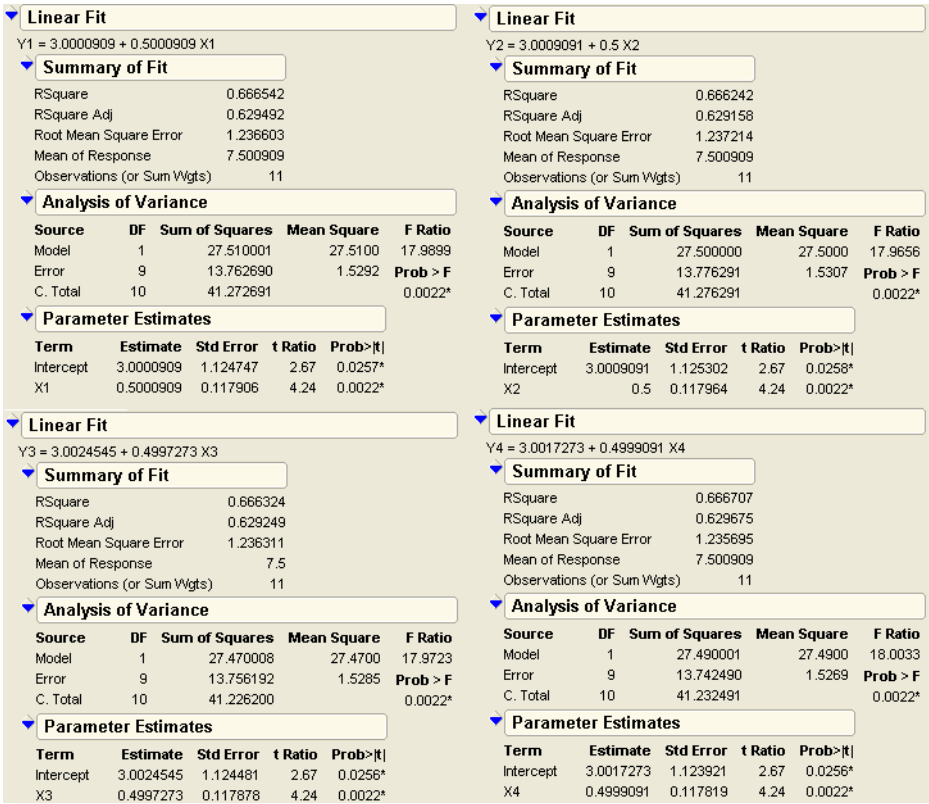
In essence, this stored script is a shortcut for the following actions. By using the script, you don't have to:

- Choose **Analyze > Fit Y by X** four times, to fit Y1 by X1, Y2 by X2, Y3 by X3, and Y4 by X4.
- For each pair, select the **Fit Line** command from the **Fitting** popup menu on the title bar above each scatterplot.

First, look at the text reports for each bivariate analysis, shown in **Figure 10.14,** and compare them. Notice that the reports are nearly identical. The R^2 values, the F-tests, the parameter estimates and standard errors—they are all the same. Does this mean the situations are the same?

Think of it another way. Suppose you are a doctor, and the four data sets represent patients. Their text reports represent the symptoms they present. Since they are identical, would you give them all the same diagnosis?

Figure 10.14 Statistical Reports for Four Analyses



Linear Fit

Y1 = 3.0000909 + 0.5000909 X1

Summary of Fit

RSquare	0.666542
RSquare Adj	0.629492
Root Mean Square Error	1.236603
Mean of Response	7.500909
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.510001	27.5100	17.9899
Error	9	13.762690	1.5292	Prob > F
C. Total	10	41.272691		0.0022*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0000909	1.124747	2.67	0.0257*
X1	0.5000909	0.117906	4.24	0.0022*

Linear Fit

Y2 = 3.0009091 + 0.5 X2

Summary of Fit

RSquare	0.666242
RSquare Adj	0.629158
Root Mean Square Error	1.237214
Mean of Response	7.500909
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.500000	27.5000	17.9656
Error	9	13.776291	1.5307	Prob > F
C. Total	10	41.276291		0.0022*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0009091	1.125302	2.67	0.0258*
X2	0.5	0.117964	4.24	0.0022*

Linear Fit

Y3 = 3.0024545 + 0.4997273 X3

Summary of Fit

RSquare	0.666324
RSquare Adj	0.629249
Root Mean Square Error	1.236311
Mean of Response	7.5
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.470008	27.4700	17.9723
Error	9	13.756192	1.5285	Prob > F
C. Total	10	41.226200		0.0022*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0024545	1.124481	2.67	0.0256*
X3	0.4997273	0.117878	4.24	0.0022*

Linear Fit

Y4 = 3.0017273 + 0.4999091 X4

Summary of Fit

RSquare	0.666707
RSquare Adj	0.629675
Root Mean Square Error	1.235695
Mean of Response	7.500909
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	27.490001	27.4900	18.0033
Error	9	13.742490	1.5269	Prob > F
C. Total	10	41.232491		0.0022*

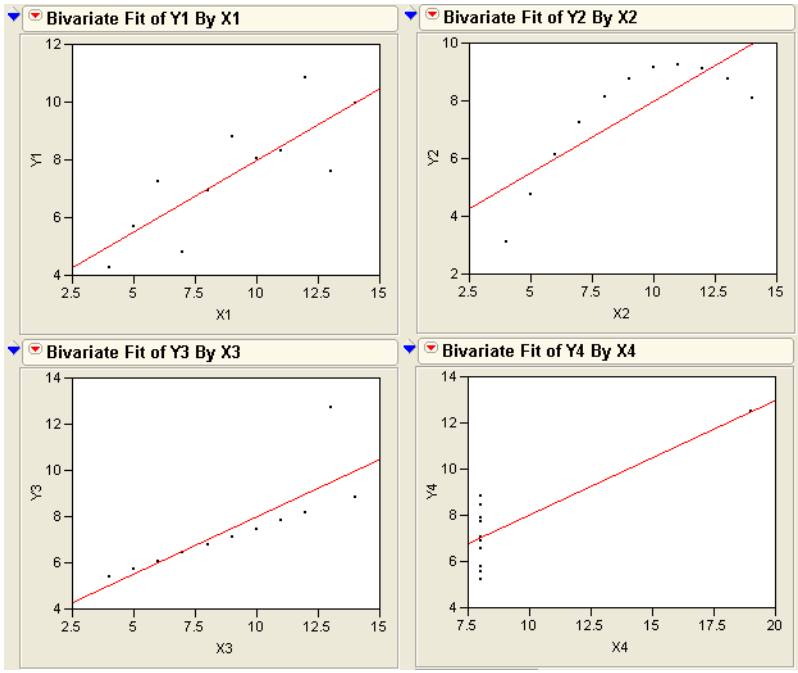
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.0017273	1.123921	2.67	0.0256*
X4	0.4999091	0.117819	4.24	0.0022*

Now look at the scatterplots of the four relations shown in **Figure 10.15**, and note the following characteristics:

- Y1 by X1 shows a Normal regression situation.
- The points in Y2 by X2 follow a parabola, so a quadratic model is appropriate, with the square of X2 as an additional term in the model. As an exercise, fit this quadratic model.
- There is an extreme outlier in Y3 by X3, which increases the slope of the line that would otherwise be a perfect fit. As an exercise, exclude the outlying point and fit another line.
- In Y4 by X4 all the *x*-values are the same except for one point, which completely determines the slope of the line. This situation is called *leverage*. It is not necessarily bad, but you ought to know about it.

Figure 10.15 Regression Lines for Four Analyses



Why It's Called Regression

Remember the story about the study done by Sir Francis Galton mentioned at the beginning of this chapter? He examined the heights of parents and their grown children to learn how much of height is an inherited characteristic. He concluded that the children's heights tended to be more moderate than the parent's heights, and used the term "regression to the mean" to name this phenomenon. For example, if a parent was tall, the children would be tall, but less so than the parents. If a parent was short, the child would tend to be short, but less so than the parent.

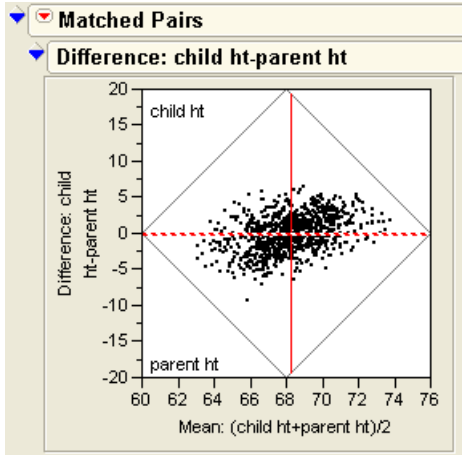
Galton's case is interesting not only because it was the first use of regression, but also because Galton failed to notice some properties of regression that would have changed his mind about using regression to draw his conclusions. To investigate Galton's data:

- Open Galton.jmp and choose **Analyze > Matched Pairs** with child ht and parent ht as the **Y, Paired Responses**.

The data in the Galton table comes from Galton's published table, but each point is *jittered* by a random amount up to 0.5 in either direction. The jittering is done so that all the points show in the plot instead of overlapping. Also, Galton multiplied the women's heights by 1.08 to make them comparable to men's. The parent ht variable is defined as the average of the two parents' heights.

The scatterplot produced by the Matched Pairs platform is the same as that given by the Fit Y by X platform, but it is rotated by 45°(See "The Matched Pairs Platform for a Paired t-Test" on page 190 for details on this plot). If the difference between the two variables is zero (our null hypothesis is that the parent and child's heights are the same), the points cluster around a horizontal reference line at zero. The mean difference is shown as a horizontal line, with the 95% confidence interval above and below. If the confidence region does not include the horizontal reference line at zero, then the means are not significantly different at the 0.05 level and we can't reject the null hypothesis. This represents the *t*-test that Galton could have hypothesized to see if the mean height of the child is the same as the parent.

Figure 10.16 Matched Pairs Output of Galton's Data



child ht	68.202	t-Ratio	-0.80923
parent ht	68.2664	DF	951
Mean Difference	-0.0644	Prob > |t|	0.4186
Std Error	0.0796	Prob > t	0.7907
Upper95%	0.09179	Prob < t	0.2093
Lower95%	-0.2206		
N	952		
Correlation	0.42095		

For this test, we do not reject the null hypothesis.

However, this is not the test that Galton considered. He invented regression to fit an arbitrary line on a plot of parent's height by child's height, and then tested to see if the slope of the line was 1. If the line has a slope of 1, then the predicted height of the child is the same as that of

the parent, except for a generational constant. A slope of less than 1 indicates that the children tended to have more moderate heights (closer to the mean) than the parents. To look at this regression,

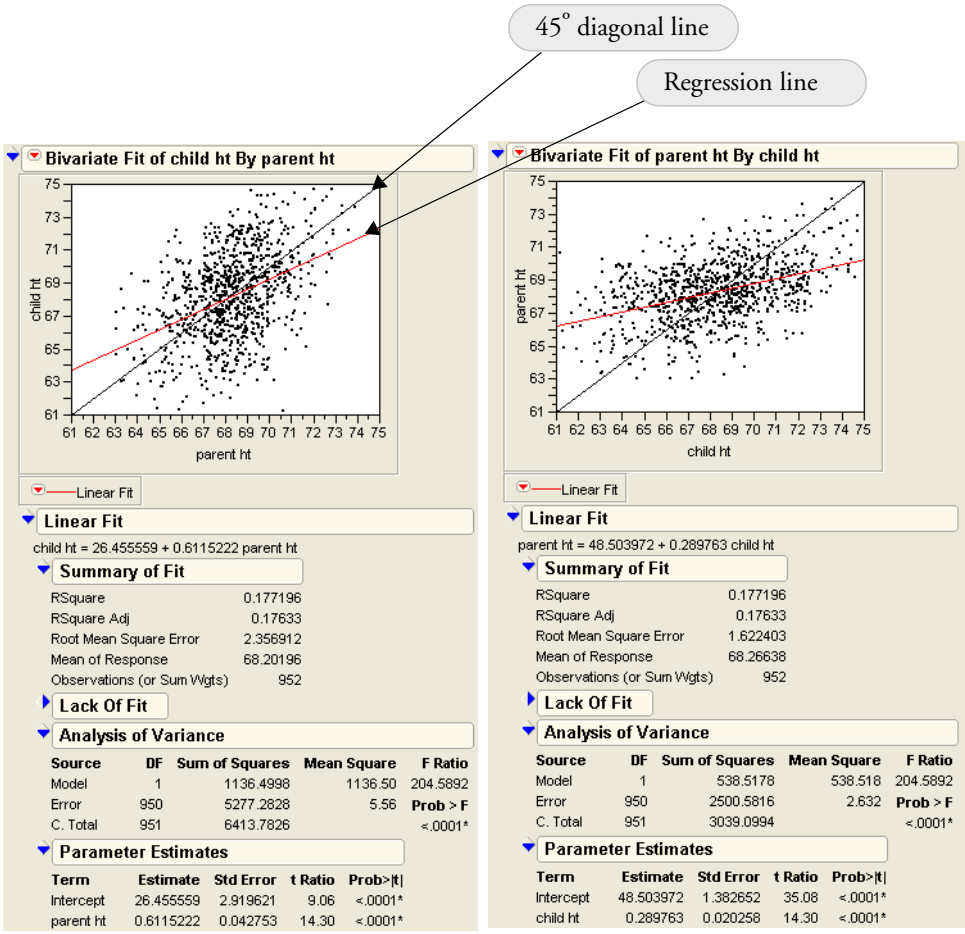
- Select **Analyze > Fit Y by X**, with parent's height as X and child's height as Y.
- Select **Fit Line** from the platform popup menu.

When you examine the Parameter Estimates table and the regression line in the left plot of **Figure 10.17**, you see that the least squares regression slope is 0.61—far below 1. This suggests the regression toward the mean.

What Happens When X and Y Are Switched?

Is Galton's technique fair to the hypothesis? Think of it in reverse: If the children's heights were more moderate than the parents, shouldn't the parent's heights be more extreme than the children's? To find out, you can reverse the model and try to predict the parent's heights from the children's heights (*i.e.* switch the roles of *x* and *y*). The analysis on the right in **Figure 10.17** shows the results when the parent's height is *y* and children's height is *x*. Because the previous slope was less than 1, you'd think that this analysis would give a slope greater than 1. Surprisingly, the reported slope is 0.28, even less than the first slope!

Figure 10.17 Child's Height and Parent's Height



Instead of phrasing the conclusion that children tended to regress to the mean, Galton could have worded his conclusion to say that there is a somewhat weak relationship between the two heights. With regression, there is no symmetry between the Y and X variables. The slope of the regression of Y on X is not the reciprocal of the slope of the regression of X on Y. You cannot calculate the X by Y slope by taking the Y by X equation and solving for the other variable.

Regression is not symmetric because the error that is minimized is only in one direction — that of the Y variable. So if the roles are switched, a completely different problem is solved.

It always happens that the slope will be smaller than the reciprocal of the inverted variables. However, there is a way to fit the slope symmetrically, so that the role of both variables is the same. This is what is you do when you calculate a *correlation*.

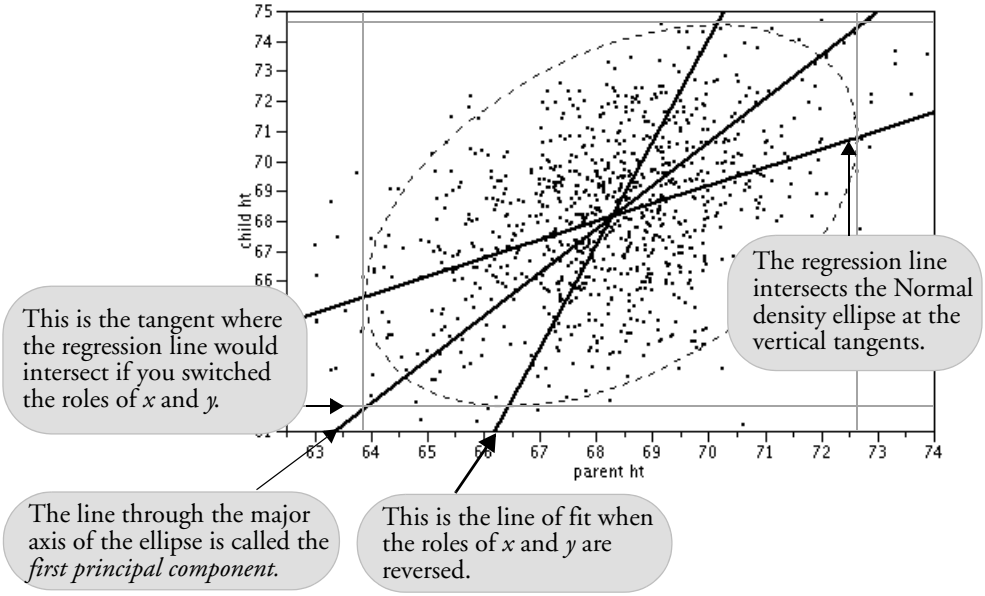
Correlation characterizes the bivariate Normal continuous density. The contours of the Normal density form ellipses, an example of which is illustrated in **Figure 10.18**. If there is a strong relationship, the ellipse becomes elongated along a diagonal axis. The line along this axis even has a name—it's called the *first principal component*.

It turns out that the least squares line is not the same as the first principal component. Instead, the least squares line bisects the contour ellipse at the vertical tangent points (see **Figure 10.18**).

If you reverse the direction of the tangents, you describe what the regression line would be if you reversed the role of the Y and X variables. If you draw the X by Y line fit in the Y by X diagram as shown in **Figure 10.18**, it intersects the ellipse at its horizontal tangents.

In both cases, the slope of the fitted line was less than 1, so Galton's phenomenon of regression to the mean was more an artifact of the method, rather than something learned from the data.

Figure 10.18 Diagram Comparing Regression and Correlation



Correlation

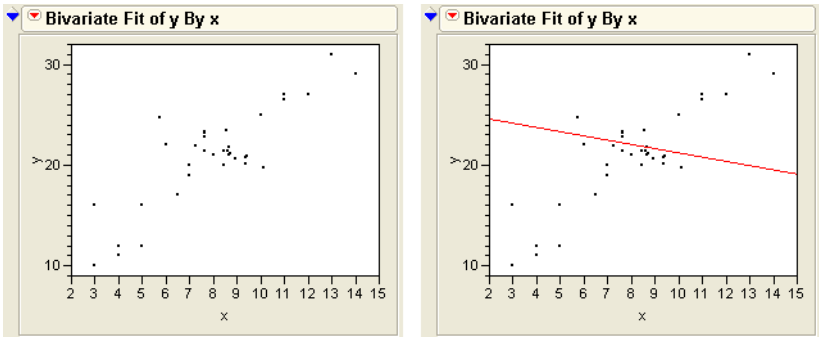
Variable	Mean	Std Dev	Correlation	Signif. Prob	Number
child ht	68.20196	2.59697	0.420947	<.0001*	952
parent ht	68.26638	1.787649			

Curiosities

Sometimes It's the Picture That Fools You

An experiment by a molecular biologist generated some graphs similar to the scatterplots in **Figure 10.19**. Looking quickly at the plot on the left, where would you guess the least squares regression line lies? Now look at the graph on the right to see where the least-squares fit really lies.

Figure 10.19 Beware of Hidden Dense Clusters



The biologist was perplexed. How did this unlikely looking regression line happen?

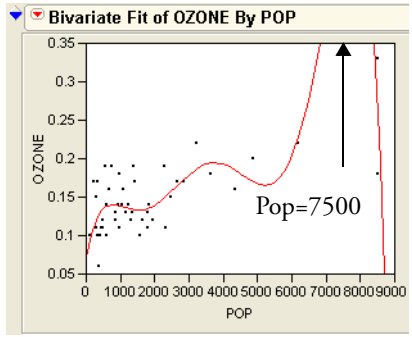
It turns out that there is a very dense cluster of points you can't see. This dense cluster of thousands of points dominated the slope estimate even though the few points farther out had more individual leverage. There was nothing wrong with the computer, the method, or the computation. It's just that the human eye is sometimes fooled, especially when many points occupy the same position. (This example uses the Slope.jmp sample data table).

High-Order Polynomial Pitfall

Suppose you want to develop a prediction equation for predicting ozone based on the population of a city. The lower-order polynomials fit fine, but why not take the "more is better" approach and try a higher order one, say, sixth degree.

- To see the next example, open the Polycity.jmp sample data table. Choose **Fit Y by X** with OZONE as Y and POP as X.
- Choose **Fit Polynomial > 6** from the **Fitting** popup menu on the scatterplot title bar.

As you can see in the bivariate fit shown to the right, the curve fits extremely well—too well, in fact. How trustworthy is the ozone prediction for a city with a population of 7500?



This overfitting phenomenon, shown to the right, occurs in higher-order polynomials when the data are unequally spaced.

More is not always better.

The Pappus Mystery on the Obliquity of the Ecliptic

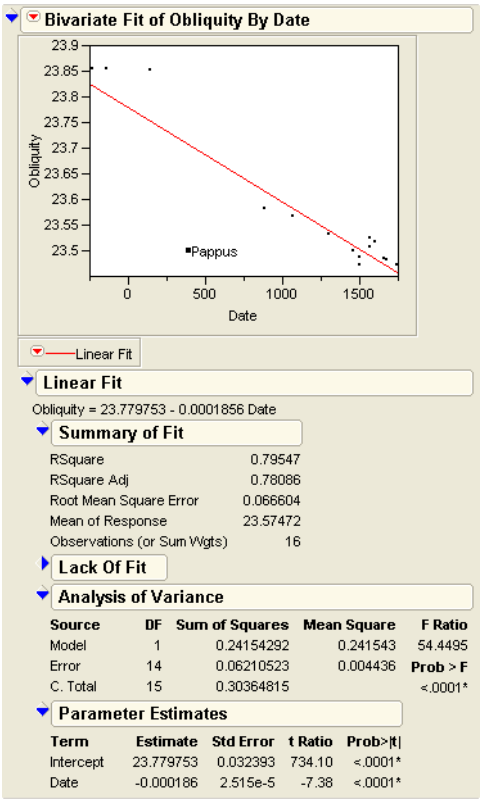
Ancient measurements of the angle of the earth's rotation disagree with modern measurements. Is this because modern ones are based on different (and better) technology, or did the angle of rotation actually change?

"Case Study: The Earth's Ecliptic" on page 140 introduced the angle-of-the-ecliptic data. The data that go back to ancient times is in the Cassini.jmp table (Stigler 1986). **Figure 10.20** shows the regression of the obliquity (angle) by time. The regression suggests that the angle has changed over time. The mystery is that the measurement by Pappus is not consistent with the rest of the line. Was Pappus's measurement flawed or did something else happen at that time? We probably will never know.

These kinds of mysteries sometimes lead to detective work that results in great discoveries. Marie Curie discovered radium because of a small discrepancy in measurements made with pitchblende experiments. If she hadn't noticed this discrepancy, the progress of physics might have been delayed.

Outliers are not to be dismissed casually. Moore and McCabe (1989) point out a situation in the 1980s where a satellite was measuring ozone levels over the poles. It automatically rejected a number of measurements because they were very low. Because of this, the ozone holes were not discovered until years later by experiments run from the earth's surface that confirmed the satellite measurements.

Figure 10.20 Measurements of the Earth's Angular Rotation



Exercises

1. This exercise deals with the data on the top-grossing box-office movies (as of June 2003) found in the data table movies.jmp. Executives are interested in predicting the amount movies will gross overseas based on the domestic gross.

   (a) Examine a scatterplot of Domestic $ vs. Worldwide $. Do you notice any outliers? Identify them.

   (b) Fit a line through the mean of the response using **Fit Mean**.

   (c) Perform a linear regression on Foreign $ vs. Domestic $. Does this linear model describe the data better than the constrained model with only the mean? Justify your answer.

(d) Exclude any outliers that you found in part (a) and re-run the regression. Describe the differences between this model and the model that included all the points. Which would you trust more?

(e) Construct a subset of this data consisting of only movies labeled "Drama" or "Comedy". Then, fit a line to this subset. Is it different than the model in part (c)?

(f) On the subsetted data, fit a line for the "Drama" movies and a separate one for "Comedy" movies. Comment on what you see, and whether you think a single prediction equation for all movie types will be useful to the executives.

2. How accurate are past elections in predicting future ones? To answer this question, open the file Presidential Elections.jmp (see Ladd and Carle). This file contains the percent of votes cast for the Democratic nominee in three recent elections.

(a) Produce histograms for the percent of votes in each election, with the three axes having uniform scaling (the **Uniform Scaling** option is in the drop-down menu at the very top of the Distribution platform.) What do you notice about the three means? If you find a difference, explain why it might be there.

(b) Find the r^2 for 1996 vs. 1980 and 1984 vs. 1980. Comment on the associations you see.

(c) Would the lines generated in these analyses be useful in predicting the percent votes cast for the Democratic nominee in the next presidential election? Justify your answer.

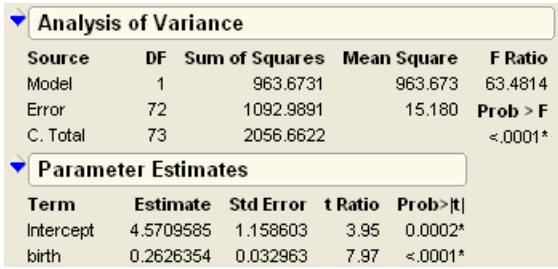
3. Open the file Birth Death.jmp to see data on the birth and death rates of several countries around the world.

(a) Identify any univariate outliers in the variables birth and death.

(b) Fit a line through the mean of the response and a regression line to the data with birth as X and death as Y. Is the linear fit significantly better than the constrained fit using just the mean?

(c) Produce a residual plot for the linear fit in part (b).

The linear fit produces the following ANOVA table and parameter estimates:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	963.6731	963.673	63.4814
Error	72	1092.9891	15.180	Prob > F
C. Total	73	2056.6622		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.5709585	1.158603	3.95	0.0002*
birth	0.2626354	0.032963	7.97	<.0001*

(d) The slope of the regression line seems to be highly significant. Why, then, is this model inappropriate for this situation?

(e) Use the techniques of this chapter to fit several transformed, polynomial, or spline models to the data and comment on the best one.

4. The table Solubility.jmp (Koehler and Dunn, 1988) contains data from a chemical experiment that tested the solubility characteristics of seven chemicals.

(a) Produce scatterplots of all the solutions vs. ether. Based on the plots, which chemical has the highest correlation with ether?

(b) Carbon tetrachloride has solubility characteristics that are highly correlated with hexane. Find a 95% confidence interval for the slope of the regression line of Carbon tetrachloride vs. hexane.

(c) Suppose the roles of the variables in part (b) were reversed. What can you say about the slope of the new regression line?

11 Categorical Distributions

Overview

When a response is categorical, a different set of tools is needed to analyze the data. This chapter focuses on simple categorical responses and introduces these topics:

- There are two ways to approach categorical data. This chapter refers to them as *choosing* and *counting*. They use different tools and conventions for analysis.
- The concept of variability in categorical responses is more difficult than in continuous responses. Monte Carlo simulation helps demonstrate how categorical variability works.
- The *chi-square test* is the fundamental statistical tool for categorical models. There are two kinds of chi-square tests. They test the same thing in a different way and get similar results.

Fitting models to categorical response data is covered in “Categorical Models” on page 283.

Categorical Situations

A *categorical response* is one in which the response is from a limited number of choices (called *response categories*). There is a probability associated with each of these choices, and these probabilities sum to 1.

Categorical responses are common:

- Consumer preferences are usually categorical: Which do you like the best— tea, coffee, juice, or soft drinks?
- Medical outcomes are often categorical: Did the patient live or die?
- Biological responses are often categorical: Did the seed germinate?
- Mechanical responses can be categorical: Did the fuse blow at 20 amps?
- Any continuous response can be converted to a categorical response: Did the temperature reach 100 degrees?

Categorical Responses and Count Data: Two Outlooks

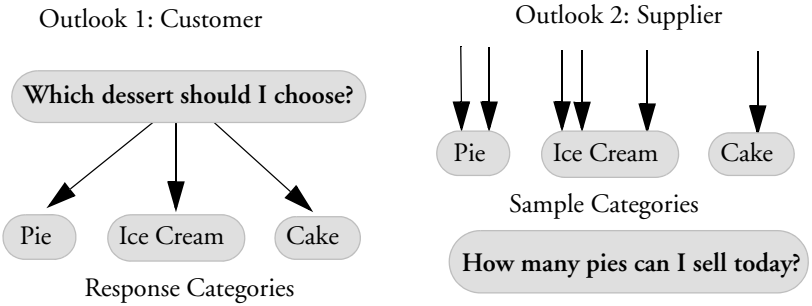
It is important to understand that there are two approaches to handling categorical responses. The two approaches generally give the same results, but they use different tools and terms.

First, imagine that each observation in a data set represents the response of a chooser. Based on conditions of the observation, the chooser is going to respond with one of the response categories. For example, the chooser might be selecting a dessert from the choices pie, ice cream, or cake. Each response category has some probability of being chosen, and that probability varies depending on other characteristics of the observational unit.

Now reverse the situation and think of yourself as the observation collector for one of the categories. For example, suppose that you sell the pies. The category "Pies" now is a sample category for the vendor, and the response of interest is how many pies can be sold in a day. Given total sales for the day of all desserts, the interest is in the market share of the pies.

Figure 11.1 diagrams these two ways of looking at categorical distributions.

Figure 11.1 Customer or Supplier?



The customer/chooser thinks in terms of *logistic regression*, where the Y variable is which dessert you choose and the X variables affect the probabilities associated with each dessert category. The supplier/counter thinks about *log-linear models*, where the Y(responses) is the count, and the X (effect) is the dessert category. There can be other effects interacting with that X.

The modeling traditions for the two outlooks are also different. Customer/chooser-oriented analyses, such as a live/die medical analysis, use continuous X's (like dose, or how many years you smoked). Supplier/counter-oriented analysts, typified by social scientists, use categorical X's (like age, race, gender, and personality type) because that keeps the data count-oriented.

The probability distributions for the two approaches are also different. This book won't go into the details of the distributions, but you can be aware of distribution names. Customer/chooser-oriented analysts refer to the *Bernoulli distribution* of the choice. Supplier/counter-oriented analysts refer to the *Poisson distribution* of counts in each category. However, both approaches refer to the *multinomial distribution.*

- To the customer/chooser analysts, the multinomial counts are aggregation statistics.
- To the supplier /counter analysts, the multinomial counts are the count distribution within a fixed total count.

The customer/chooser analyst thinks the basic analysis is fitting response category probabilities. The supplier/counter analyst thinks that basic analysis is a one-way analysis of variance on the counts and uses weights because the distribution is Poisson instead of Normal.

Both orientations are right—they just have different outlooks on the same statistical phenomenon.

In this book, the emphasis is on the customer/chooser point of view, also known as the *logistic regression* approach. With logistic regression, it is important to distinguish the responses (Y's),

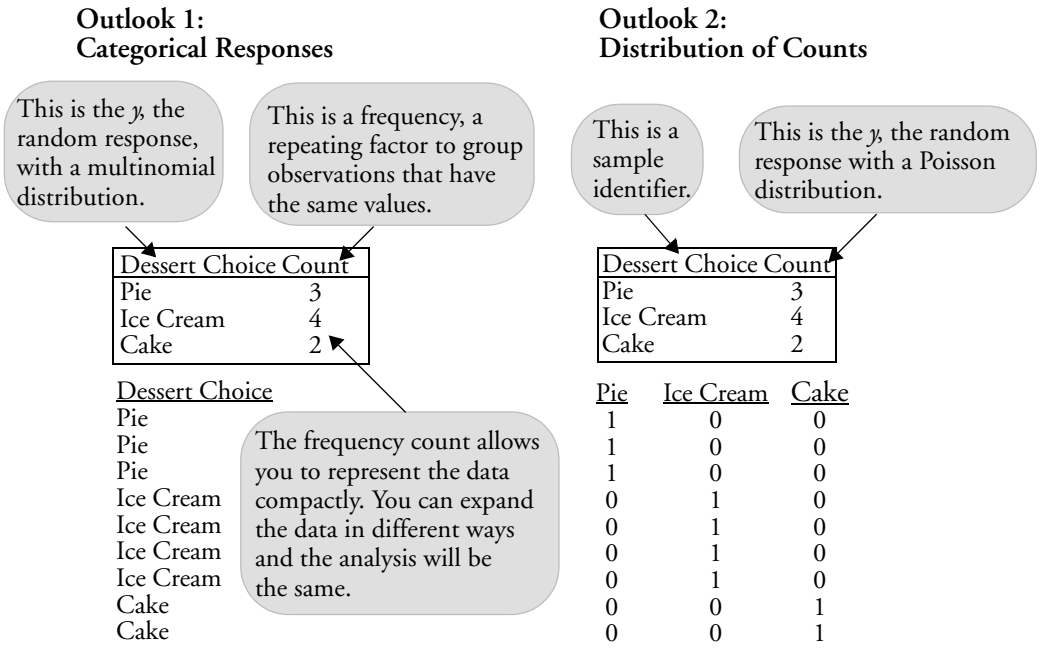
which have the random element, from the factors (X's), which are fixed from the point of view of the model. The X's and Y's must be distinguished before the analysis is started.

Let's be clear on what the X's and Y's are for the chooser point of view:

- Responses (Y's) identify a choice or an outcome. They have a random element because the choice is not determined completely by the X factors. Examples of responses are patient outcome (lived or died), or desert preference (Gobi or Sahara).
- Factors (X's) identify a sample population, an experimentally controlled condition, or an adjustment factor. They are not regarded as random even if you randomly assigned them. Examples of factors are gender, age, treatment or block.

Figure 11.2 illustrates the X and Y variables for both outlooks on categorical models.

Figure 11.2 Categories or Counts?



The other point of view is the *log-linear model* approach. The log-linear approach regards the count as the Y variable and all the other variables as X's. After fitting the whole model, the effects that are of interest are identified. Any effect that has no response category variable is discarded, since it is just an artifact of the sampling design. Log-linear modeling uses a technique called *iterative proportional fitting* to obtain test statistics. This process is also called *raking*.

A Simulated Categorical Response

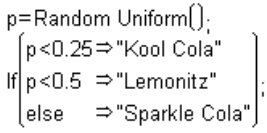
A good way to learn statistical techniques is to simulate data with known properties, and then analyze the simulation to see if you find the structure that you put into the simulation.

These steps describe the simulation process:

1. Simulate one batch of data, then analyze.
2. Simulate more batches, analyze them, and notice how much they vary.
3. Simulate a larger batch to notice that the estimates have less variance.
4. Do a batch of batches—simulations that for each run obtain sample statistics over a new batch of data.
5. Use this last batch of batches to look at the distribution of the test statistics.

Simulating Some Categorical Response Data

Let's make a world where there are three soft drinks. The most popular ("Sparkle Cola") has a 50% market share and the other two ("Kool Cola" and "Lemonitz") are tied at 25% each. To simulate a sample from this population, create a data table that has one variable (call it Drink Choice), which is drawn as a random categorical variable using the following formula:



This formula first draws a uniform random number between 0 and 1 using the **Random Uniform** function, and assigns the result to a temporary variable p. Then it compares that random number using If conditions on the random number and picks the first response where the condition is true. Each case returns the character name of a soft drink as the response value.

This table has already been created.

- Open the data table Cola.jmp, which contains the formula shown previously. It is found in the sample data directory.
- Right-click on the Drink Choice column and select **Formula** to display the stored formula.

Note that there are two statements in a single formula, which are delimited with a semicolon. The Formula Editor operations needed to construct this formula include the following operations: **Local Variables: New Local** creates a temporary variable named p, The **If** statement from the **Conditional** functions assigns soft drink names to probability conditions given by the **Random Uniform** function found in the **Random** functions.

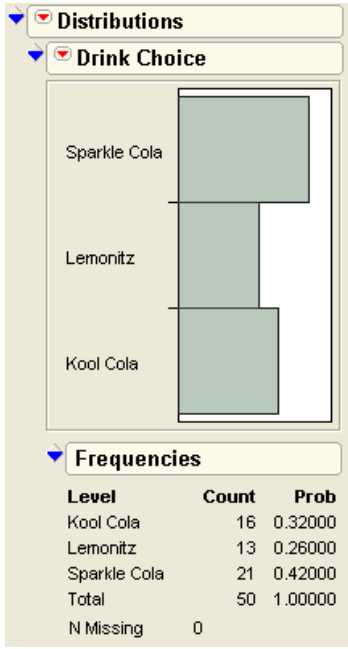
The table is stored with no rows. A data table stored with no rows and columns with formulas is called a *table template*.

- Choose **Rows > Add Rows** to add 50 rows to the table.
- Choose **Analyze > Distribution** with the Drink Choice variable as Y, which gives an analysis similar to (but not exactly like) that in **Figure 11.3**.

Don't expect to get the same numbers that we show here, because the formula generates random data. Each time the computations are performed, a different set of data is produced.

Note that even though the data are based on the true probabilities of 0.25, 0.25, and 0.50, the estimates are different (0.32, 0.26, and 0.42). Your data have random values with somewhat different probabilities.

Figure 11.3 Histogram and Frequencies of Simulated Data



Variability in the Estimates

The following sections distinguish between ρ (Greek rho), the true value of a probability, and *p*, its estimate. The true value ρ is a fixed number, an unknowable "true" value, but its estimate *p* is an outcome of a random process, so it has variability associated with it.

You cannot compute a standard deviation of the original responses—of "Kool Cola", et al., because they are character values. However, the variability in the probability estimates is well-defined and computable.

Just as with continuous variables, the variability of an estimate is expressed by its variance or its standard deviation, although the quantities are computed with different formulas. The variance of *p* is given by the formula

$$\frac{\rho(1-\rho)}{n}$$

For Sparkle Cola, having a ρ of 0.50, the variance of the probability estimate is (0.5•0.5)/50, 0.005. The standard deviation of the estimate is the square root of this variance, 0.07071. Table 11.1 compares the difference between the true ρ and its estimate *p*. Then, it compares the true standard deviation of the statistic *p*, and the standard error of *p*, which estimates the standard deviation of *p*.

Remember, the term *standard error* is used to label an estimate of the standard deviation of another estimate. Only because this is a simulation with known true values (parameters) can you see both the standard errors and the true standard deviations.

Table 11.1. Simulated Probabilities and Estimates

Level	ρ, the True Probability	*p*, the Estimate of ρ	True Standard Deviation of the Estimate	Standard Error of the Estimate
Kool Cola	0.25	0.32	0.06124	0.07065
Lemonitz	0.25	0.26	0.06124	0.06203
Sparkle Cola	0.50	0.42	0.07071	0.06980

This simulation shows a lot of variability. As with Normally distributed data, you can expect to find estimates that are two standard deviations from the true probability about 5% of the time.

Now let's see how the estimates vary with a new set of random responses.

- Right-click on the Drink Choice column heading and select **Formula**.
- Click **Apply** on the calculator to re-evaluate the random formula.
- Again perform **Analyze > Distribution** on Drink Choice.
- Repeat this evaluate/analyze cycle four times.

Each repetition results in a new set of random responses and a new set of estimates of the probabilities. Table 11.2 gives the estimates from the four Monte Carlo runs.

Table 11.2. Estimates from Monte Carlo Runs

Level	Probability	Probability	Probability	Probability
Kool Cola	0.32000	0.18000	0.26000	0.40000
Lemonitz	0.26000	0.32000	0.24000	0.18000
Sparkle Cola	0.42000	0.50000	0.50000	0.42000

With only 50 observations, there is a lot of variability in the estimates. The "Kool Cola" probability estimate varies between 0.18 and 0.40, the "Lemonitz" estimate varies between 0.18 and 0.32, and the "Sparkle Cola" estimate varies between 0.42 and 0.50.

Larger Sample Sizes

What happens if the sample size increases from 50 to 500? Remember that the variance of p is

$$\frac{\rho(1-\rho)}{n}$$

With more data, the probability estimates have a much smaller variance. To see what happens when we add observations,

- Choose **Rows > Add Rows** and enter 450, to get a total of 500 rows.
- Perform **Analyze > Distribution** for the response variable Drink Choice.

Five hundred rows give a smaller variance, 0.0005, and a standard deviation at about $\sqrt{0.005}$ =0.02. The figure to the right shows the 50-row simulation for 500 rows instead. To see the Std Err Prob column,

- Right-click in the **Frequencies** table.

Frequencies

Level	Count	Prob	StdErr Prob
Kool Cola	115	0.23000	0.01882
Lemonitz	135	0.27000	0.01985
Sparkle Cola	250	0.50000	0.02236
Total	500	1.00000	0.00000
N Missing	0		

- Select **Columns > Std Err Prob**

Now we extend the simulation.

- Repeat the evaluate/analyze cycle four times.

Table 11.3 shows the results of the next 4 simulations.

Table 11.3. Estimates from Four Monte Carlo Runs

Level	Probability	Probability	Probability	Probability
Kool Cola	0.28000	0.25000	0.25600	0.23400
Lemonitz	0.24200	0.28200	0.23400	0.26200
Sparkle Cola	0.47800	0.46800	0.51000	0.50400

Note that the probability estimates are closer to the true values and that the standard errors are smaller.

Monte Carlo Simulations for the Estimators

What do the distributions of these counts look like? Variances can be easily calculated, but what is the distribution of the count estimate? Statisticians often use Monte Carlo simulations to investigate the distribution of a statistic.

To simulate estimating a probability (which has a true value of 0.25 in this case) over a sample size (50 in this case), construct the formula shown here.

The **Random Uniform** function generates a random value distributed uniformly between 0 and 1. The random value is checked to see if it is less than 0.25. The term in the numerator evaluates to 1 or 0 depending on this comparison. It is 1 about 25% of the time, and 0 about 75% of the time. This random number is generated 50 times (look at the indices of the summation), and the sum of them is divided by 50.

$$\frac{\sum_{j=1}^{50} \text{Random Uniform}() < 0.25}{50}$$

This formula is a simulation of a Bernoulli event with 50 samplings. The result estimates the probability of getting a 1. In this case, you happen to know the true value of this probability (0.25) because you constructed the formula that generated the data.

Now, it is important to see how well these estimates behave. Theoretically, the mean (expected value) of the estimate, p, is 0.25 (the true value), and its standard deviation is the square root of

$$\frac{\rho \cdot (1 - \rho)}{n}$$

which is 0.061237.

Distribution of the Estimates

The sample data has a table template called Simprob.jmp that is a Monte Carlo simulation for the probability estimates of 0.25 and 0.5, based on 50 and 500 trials. You can add 1000 rows to the data to draw 1000 Monte Carlo trials.

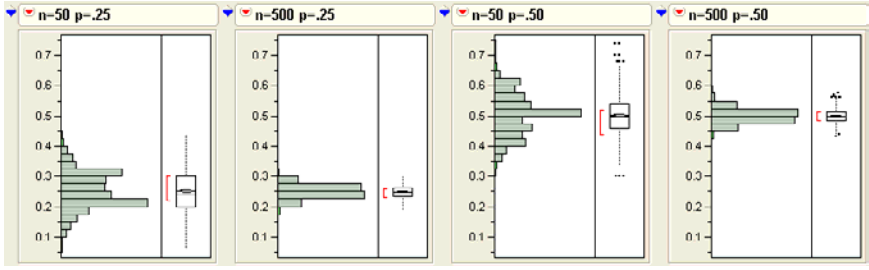
To see how the estimates are distributed:

- Open Simprob.jmp.
- Choose **Rows > Add Rows** and enter 1000.
- Next choose **Analyze > Distribution** and use all the columns as Y variables.
- When the histograms appear, select the **Uniform Scaling** option from the check-mark popup menu beside the word "Distributions" on the title bar of the report.
- Get the grabber (hand) tool from the **Tools** menu and drag the histograms to adjust the bar widths and positions.

Figure 11.4 and Table 11.4 show the properties as expected:

- The variance decreases as the sample size increases.
- The distribution of the estimates is approximately Normally distributed, especially as the sample size gets large.

Figure 11.4 Histograms for Simulations with Various *n* and *p* Values



The estimates of the probability *p* of getting response indicator values of 0 or 1 are a kind of mean. So, as the sample gets larger, the value of *p* gets closer and closer to 0.50, the mean of 0

and 1. Like the mean for continuous data, the standard error of the estimate relates to the sample size by the factor $1/\sqrt{n}$.

The Central Limit Theorem applies here. It says that the estimates approach a Normal distribution when there are a large number of observations.

Table 11.4. Summary of Simulation Results

True Value of *p*	N Used to Estimate *p*	Mean of the Trials of the Estimates of *p*	Standard Deviation of Trials of Estimates of *p*	True Mean of Estimates	True Standard Deviation of the Estimates
0.25	50	0.24774	0.06118	0.25	0.061237
0.25	500	0.24971	0.01974	0.25	0.019365
0.50	50	0.49990	0.06846	0.50	0.070711
0.50	500	0.50058	0.02272	0.50	0.022361

The X^2 Pearson Chi-Square Test Statistic

Because of the Normality of the estimates, it is reasonable to use Normal-theory statistics on categorical response estimates. Remember that the Central Limit Theorem says that the sum of a large number of independent and identically distributed random values have a nearly Normal distribution.

However, there is a big difference between having categorical and continuous responses. With categorical responses, the variances of the differences are known. They are solely a function of *n* and the probabilities. The hypothesis specifies the probabilities, so calculations can be made under the null hypothesis. Rather than using an *F*-statistic, this situation calls for the χ^2 *(chi-square) statistic*.

The standard chi-square for this model is the following scaled sum of squares:

$$\chi^2 = \frac{\sum_{j=1} (\text{Observed}_j - \text{Expected}_j)^2}{\text{Expected}_j}$$

where Observed and Expected refer to cell counts rather than probabilities.

The G^2 Likelihood-Ratio Chi-Square Test Statistic

Whereas the Pearson chi-square assumes Normality of the estimates, another kind of chi-square test is calculated with direct reference to the probability distribution of the response and so does not require Normality.

Define the *maximum likelihood estimator* to be the one that finds the values of the unknown parameters that maximize the probability of the data. In statistical language, it finds parameters that make the data that actually occurred less improbable than they would be with any other parameter values. The term *likelihood* means the probability has been evaluated as a function of the parameters with the data fixed.

It would seem that this requires a lot of guesswork in finding the parameters that maximize the likelihood of the observed data, but just as in the case of least squares, mathematics can provide short cuts to computing the ideal coefficients. There are two fortunate short cuts for finding a maximum likelihood estimator:

- Because observations are assumed to be independent, the joint probability across the observations is the product of the probability functions for each observation.
- Because addition is easier than multiplication, instead of multiplying the probabilities to get the joint probability, you add the logarithms of the probabilities, which gives the *log-likelihood*.

This makes for easy computations. Remember that an individual response has a multinomial distribution, so the probability is ρ_i for the i=1 to r probabilities over the r response categories.

Consider the first five responses of the cola example: Kool Cola, Lemonitz, Sparkle Cola, Sparkle Cola, and Lemonitz. For Kool Cola, Lemonitz, and Sparkle Cola, denote the probabilities as $\rho 1$, $\rho 2$, and $\rho 3$ respectively. The joint log-likelihood is:

$$\log(\rho 1) + \log(\rho 2) + \log(\rho 3) + \log(\rho 3) + \log(\rho 2)$$

It turns out that this likelihood is maximized by setting the probability parameter estimates to the category count divided by the total count, giving

$$p1 = n1/n = 1/5$$

$$p2 = n2/n = 2/5$$

$$p3 = n3/n = 2/5$$

where $p1$, $p2$, and $p3$ estimate $\rho 1$, $\rho 2$, and $\rho 3$. Substituting this into the log-likelihood gives the maximized log-likelihood of

$$\log(1/5) + \log(2/5) + \log(2/5) + \log(2/5) + \log(2/5)$$

At first it may seem that taking logarithms of probabilities is a mysterious and obscure thing to do, but it is actually very natural. You can think of the negative logarithm of p as the number of binary questions you need to ask to determine which of $1/p$ equally likely outcomes happens. The negative logarithm converts units of probability into units of information. You can think of the negative loglikelihood as the *surprise* value of the data because surprise is a good word for unlikeliness.

Likelihood Ratio Tests

One way to measure the credibility for a hypothesis is to compare how much surprise (–log-likelihood) there would be in the actual data with the hypothesized values compared with the surprise at the maximum likelihood estimates. If there is too much surprise, then you have reason to throw out the hypothesis.

It turns out that the distribution of twice the difference in these two surprise (-log-likelihood) values approximately follows a chi-square distribution.

Here is the setup: Fit a model twice. The first time, fit using maximum likelihood with no constraints on the parameters. The second time, fit using maximum likelihood, but constrain the parameters by the null hypothesis that the outcomes are equally likely. It happens that twice the difference in log-likelihoods has an approximate chi-square distribution (under the null hypothesis). These chi-square tests are called *likelihood ratio chi-squares*, or *LR chi-squares*.

Twice the difference in the log-likelihood is a likelihood ratio chi-square test.

The likelihood ratio tests are very general. They occur not only in categorical responses, but also in a wide variety of situations.

The G^2 Likelihood Ratio Chi-Square Test

Let's focus on Bernoulli probabilities for categorical responses. The log-likelihood for a whole sample is the sum of natural logarithms of the probabilities attributed to the events that actually occurred.

$$\text{log-likelihood} = \sum \ln(\text{probability the model gives to event that occurred in data})$$

The likelihood ratio chi-square is twice the difference in the two likelihoods, when one is constrained by the hypothesis and the other is unconstrained.

G^2 = 2 (log-likelihood(unconstrained) – log-likelihood(constrained))

This is formed by the sum over all observations

$$G^2 = \sum[\log(\rho_{y_i}) - \log(p_{y_i})]$$

where ρ_{y_i} is the hypothesized probability and p_{y_i} is the estimated rate for the events y_i that actually occurred.

If you have already collected counts for each of the responses, and bring the subtraction into the log as a division, the formula becomes

$$G^2 = 2\sum n_i \log\frac{\rho_{y_i}}{p_{y_i}}$$

To compare with the Pearson chi-square, which is written schematically in terms of counts, the LR chi-square statistic can be written

$$G^2 = 2\sum \text{observed}\left(\log\frac{\text{expected}}{\text{observed}}\right)$$

Univariate Categorical Chi-Square Tests

A company gave 998 of its employees the Myers-Briggs Type Inventory (MBTI) questionnaire. The test is scored to result in a 4-character personality type for each person. There are 16 possible outcomes, represented by 16 combinations of letters (see **Figure 11.5**). The company wanted to know if its employee force was statistically different in personality types from the general population.

Comparing Univariate Distributions

The data table Mb-dist.jmp has a column called TYPE to use as a Y response, and a Count column to use as a frequency. To see the company test results:

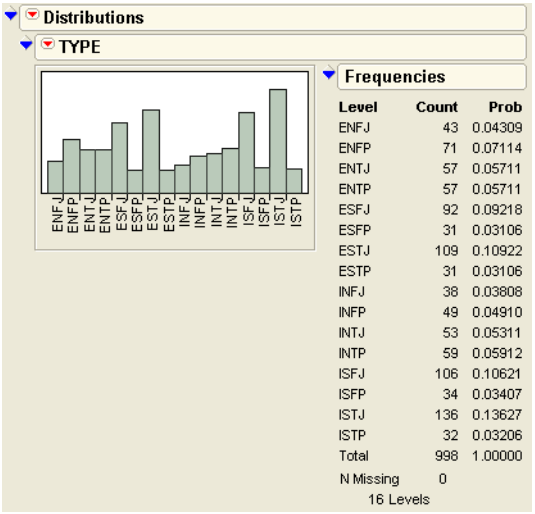
- Open the sample table called Mb-dist.jmp.

- Choose **Analyze > Distribution** and use Type as the Y variable and Count as the frequency variable.

When the report appears,

- Select **Display Options > Horizontal Layout** to see the report in **Figure 11.5**.

Figure 11.5 Histogram and Frequencies for Myers-Briggs Data



To test the hypothesis that the personalities test results at this company occur at the same rates as the general population:

- Select the **Test Probabilities** command in the popup menu on the histogram title bar at the top of the report.

A dialog then appears at the end of the report.

- Edit the Hypoth Prob (hypothesized probability) values by clicking and then entering the values as shown on the left in **Figure 11.6**.

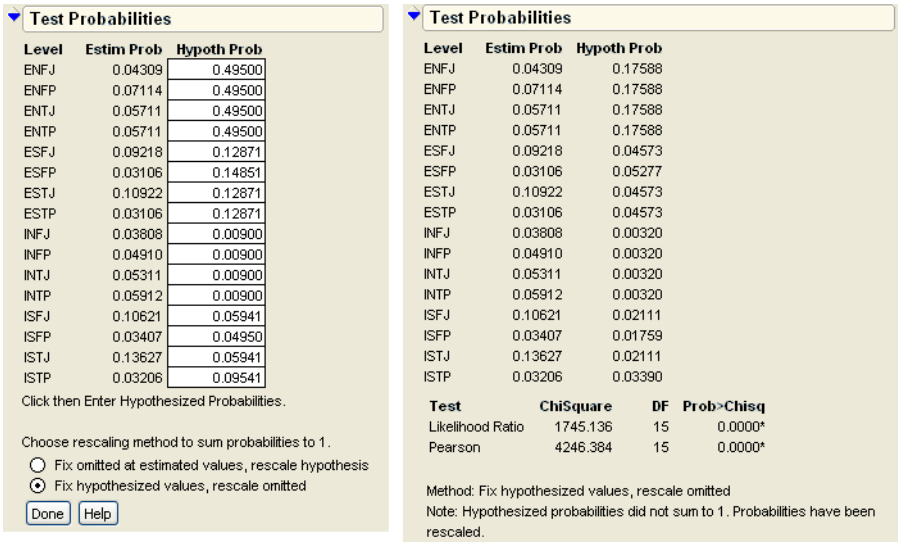
These are the general population rates for each personality type.

- Click the lower radio button in the dialog, then click **Done**.

You now see the test results appended to the Test Probabilities table, as shown in the table on the right in **Figure 11.6**.

Note that the company does have a significantly different profile than the general population. Both chi-square tests are highly significant. The company appears to have more ISTJ's (introvert sensing thinking judging) and fewer ESFP's (extrovert sensing feeling perceiving) than the general population.

Figure 11.6 Test Probabilities Report for the Myers-Briggs Data



Test Probabilities

Level	Estim Prob	Hypoth Prob
ENFJ	0.04309	0.49500
ENFP	0.07114	0.49500
ENTJ	0.05711	0.49500
ENTP	0.05711	0.49500
ESFJ	0.09218	0.12871
ESFP	0.03106	0.14851
ESTJ	0.10922	0.12871
ESTP	0.03106	0.12871
INFJ	0.03808	0.00900
INFP	0.04910	0.00900
INTJ	0.05311	0.00900
INTP	0.05912	0.00900
ISFJ	0.10621	0.05941
ISFP	0.03407	0.04950
ISTJ	0.13627	0.05941
ISTP	0.03206	0.09541

Click then Enter Hypothesized Probabilities.

Choose rescaling method to sum probabilities to 1.

○ Fix omitted at estimated values, rescale hypothesis

◉ Fix hypothesized values, rescale omitted

Done Help

Test Probabilities

Level	Estim Prob	Hypoth Prob
ENFJ	0.04309	0.17588
ENFP	0.07114	0.17588
ENTJ	0.05711	0.17588
ENTP	0.05711	0.17588
ESFJ	0.09218	0.04573
ESFP	0.03106	0.05277
ESTJ	0.10922	0.04573
ESTP	0.03106	0.04573
INFJ	0.03808	0.00320
INFP	0.04910	0.00320
INTJ	0.05311	0.00320
INTP	0.05912	0.00320
ISFJ	0.10621	0.02111
ISFP	0.03407	0.01759
ISTJ	0.13627	0.02111
ISTP	0.03206	0.03390

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	1745.136	15	0.0000*
Pearson	4246.384	15	0.0000*

Method: Fix hypothesized values, rescale omitted

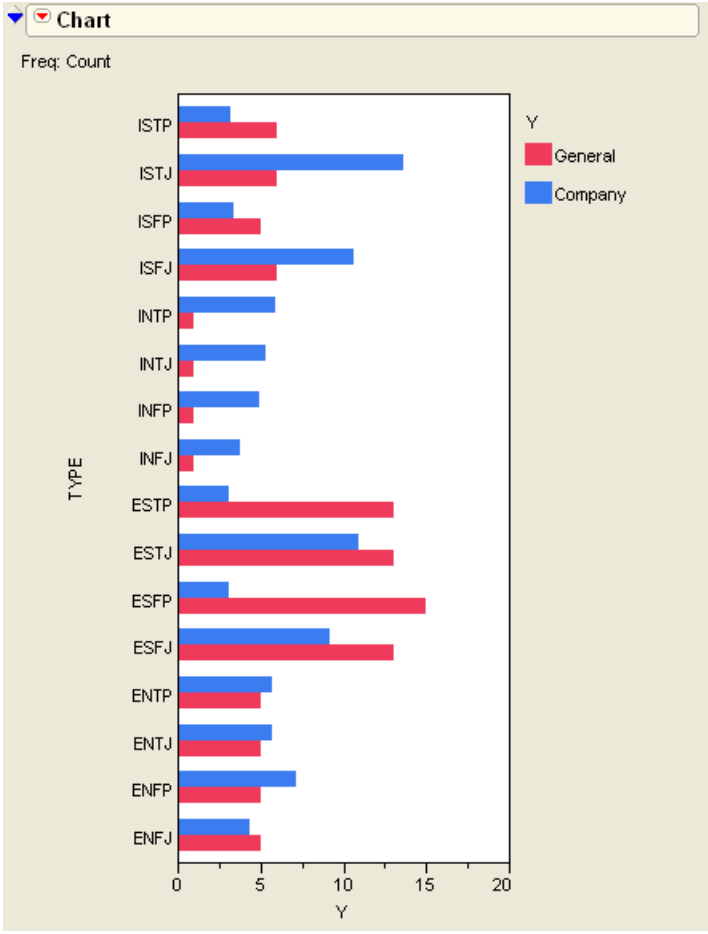
Note: Hypothesized probabilities did not sum to 1. Probabilities have been rescaled.

By the way, some people find it upsetting that different statistical methods get different results. Actually, the G^2 (likelihood ratio) and χ^2 (Pearson) chi-square statistics are usually close.

Charting to Compare Results

- The report in **Figure 11.7** was done by using **Graph > Chart** with Type as X, and the actual data from General and Company as Y.
- When the chart appears, select **Horizontal** command in the Chart title bar popup menu to rearrange the chart.

Figure 11.7 Mean Sample Personality Scores and Scores for General Population



Exercises

1. P&D Candies produces a bag of assorted sugar candies, called "Moons", in several colors. Based on extensive market research, they have decided on the following mix of Moons in each bag: Red, 20%, Yellow 10%, Brown 30%, Blue 20%, and Green, 20%. A consumer advocate suspects that the mix is not what the company claims, so he gets a bag containing 100 Moons. The 100 pieces of candy are represented in the file Candy.jmp (fictional data).

   (a) Can the consumer advocate reasonably claim that the company's mix is not as they say?

   (b) Do you think a single-bag sample is representative of all the candies produced?

2. One of the ways that public schools can make extra money is to install vending machines for students to access between classes. Suppose a high school installed three drink machines for different manufacturers in a common area of the school. After one week, they collected information on the number of visits to each machine, as shown in the following table:

Machine A	Machine B	Machine C
1546	1982	1221

Is there evidence of the students preferring one machine over another?

12 Categorical Models

Overview

Chapter 11, "Categorical Distributions," introduced the distribution of a single categorical response. You were introduced to the Pearson and the likelihood ratio chi-square tests and saw how to compare univariate categorical distributions.

This chapter covers multivariate categorical distributions. In the simplest case, the data can be presented as a two-way contingency table of frequency counts, with the expected cell probabilities and counts formed from products of marginal probabilities and counts. The chi-square test again is used for the contingency table and is the same as testing multiple categorical responses for independence.

Correspondence analysis is shown as a graphical technique useful when the response and factors have many levels or values.

Also, a more general categorical response model is used to introduce nominal and ordinal logistic regression, which allows multiple continuous or categorical factors.

Fitting Categorical Responses to Categorical Factors: Contingency Tables

When a categorical response is examined in relationship to a categorical factor (in other words, both X and Y are categorical), the question is: do the response probabilities vary across factor-defined subgroups of the population? Comparing a continuous response and a categorical factor in this way was covered in "Comparing Many Means: One-Way Analysis of Variance" on page 209. In that chapter, means were fit for each level of a categorical variable and tested using an ANOVA. When the continuous response is replaced with a categorical response, the equivalent technique is to estimate response probabilities for each subgroup and test that they are the same across the subgroups.

The subgroups are defined by the levels of a categorical factor (*X*). For each subgroup, the set of response probabilities must add up to 1. For example, consider the following:

- The probability of whether a patient lives or dies (response probabilities) depending on whether the treatment (categorical factor) was drug or placebo.
- The probability that type of car purchased (response probabilities) depending on marital status (the categorical factor).

To estimate response probabilities for each subgroup, you take the count in a given response level and divide it by the total count from that subgroup.

Testing with G^2 and X^2

You want to test whether the factor affects the response. The null hypothesis is that the response probabilities are the same across subgroups. The model comparison is to compare the fitted probabilities over the subgroups to the fitted probabilities combining all the groups into one population (a constant response model).

As a measure of fit for the models you want to compare, you can use the negative log-likelihood to compute a likelihood-ratio chi-square test. To do this, subtract the log-likelihoods for the two models and multiply by 2. For each observation, the log-likelihood is the log of the probability attributed to the response level of the observation.

Warning: When the table is *sparse*, neither the Pearson or likelihood ratio chi-square is a very good approximation to the true distribution. The Cochran criterion, used to determine if the tests are appropriate, defines sparse as when more than 20% of the cells have expected counts less than 5. JMP presents a warning when this situation occurs.

The Pearson chi-square tends to be better behaved in sparse situations than the likelihood ratio chi-square. However, G^2 is often preferred over X^2 for other reasons, specifically because it is generalizable to general categorical models where X^2 is not.

"Categorical Distributions" on page 265 discussed the G^2 and X^2 test statistics in more detail.

Looking at Survey Data

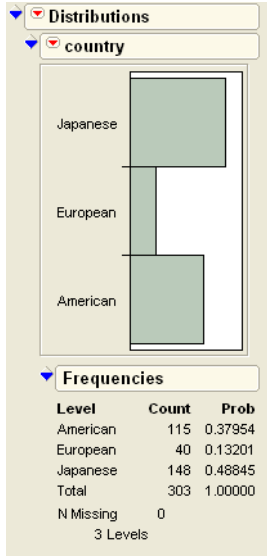
Survey data often yield categorical data suitable for contingency table analysis. For example, suppose a company did a survey to find out what factors relate to the brand of automobile people buy— in other words, what kind of people buy what kind of cars? Cars were classified into three brands: American, European, and Japanese (which included other Asian brands). This survey also contained demographic information (marital status and gender of the purchasers).

The results of the survey are in the sample table called Carpoll.jmp. A good first step is to examine probabilities for each brand when nothing else is known about the car buyer. Looking at the distribution of car brand gives this information. To see the report on the distribution of brand shown in **Figure 12.1**,

- Open Car Poll.jmp and choose **Analyze > Distribution** with country as the variable.

Overall, the Japanese brands have a 48.8% share.

Figure 12.1 Histograms and Frequencies for country in Car Poll Data



The next step is to look at the demographic information as it relates to brand of auto.

- Choose **Analyze > Fit Y by X** with country as Y, and sex, marital status, and size as X variables.
- Click **OK**.

The Fit Y by X platform displays Mosaic plots and Crosstabs tables for the combination of country with each of the X variables. By default, JMP displays Count, Total%, Col%, and Row% (listed in the upper-left corner of the table) for each cell in the contingency table.

- Right-click in the icon in the Contingency table to see the menu of the optional items. Uncheck all items except **Count** and **Col%** to see the table shown to the right.

Contingency Table

		country			
	Count Col %	American	European	Japanese	
sex	Female	54 46.96	19 47.50	65 43.92	138
	Male	61 53.04	21 52.50	83 56.08	165
		115	40	148	303

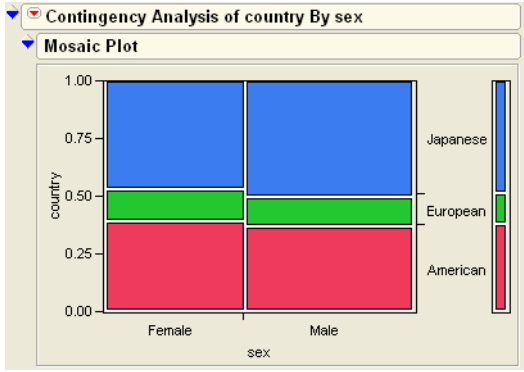
Contingency Table: Country by Sex

Is the distribution of the response levels different over the levels of other categorical variables? In principle, this is like a one-way analysis of variance, estimating separate means for each sample, but this time they are rates over response categories rather than means.

In the contingency table, you see the response probabilities as the Col% values in the bottom of each cell. The column percents are not much different between “Female” and “Male.”

Mosaic Plot

The Fit Y by X platform for a categorical variable displays information graphically with mosaic plots like the one shown to the right.



A mosaic plot is a set of side-by-side divided bar plots to compare the subdivision of response probabilities for each sample. The mosaic is formed by first dividing up the horizontal axis according to the sample proportions. Then each of these cells is subdivided vertically by the estimated response probabilities. The area of each rectangle is proportional to the frequency count for that cell.

Testing Marginal Homogeneity

Now ask the question, "Are the response probabilities significantly different across the samples (in this example, male and female)?" Specifically, is the proportion of sales by country the same for males and females? The null hypothesis that the distributions are the same across the sample's subgroup is sometimes referred to as the hypothesis of *marginal homogeneity*.

Instead of regarding the categorical X variable as fixed, you can consider it as another Y response variable and look at the relationship between two Y response variables. The test would be the same, but the null hypothesis would be known by a different name, as the *test for independence*.

When the response was continuous, there were two ways to get a test statistic that turned out to be equivalent:

- Look at the distribution of the estimates, usually leading to a *t*-test.
- Compare the fit of a model with a submodel, leading to an *F*-test.

The same two approaches work for categorical models. However, the two approaches to getting a test statistic for a contingency table both result in chi-square tests.

- If the test is derived in terms of the distribution of the estimates, then you are led to the Pearson X^2 form of the χ^2 test.
- If the test is derived by comparing the fit of a model with a submodel, then you are led to the likelihood-ratio G^2 form of the χ^2 test.

Tests

N	DF	-LogLike	RSquare (U)
303	2	0.15593790	0.0005

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	0.312	0.8556
Pearson	0.312	0.8556

For the likelihood ratio chi-square (G^2), two models are fit by maximum likelihood. One model is constrained by the hypothesis that assumes a single response population and the other is not constrained. Twice the difference of the log-likelihoods from the two models is a chi-square statistic for testing the hypothesis. The table here has the chi-square tests that test whether country of car purchased is a function of sex.

The model constrained by the null hypothesis (fitting only one set of response probabilities) has a negative log-likelihood of 298.45. After you partition the sample by the gender factor, the negative log-likelihood is reduced to 298.30. The difference in log-likelihoods is 0.1559, reported in the -LogLike line. This doesn't account for much of the variation. The likelihood ratio (LR) chi-square is twice this difference, that is, $G^2 = 0.312$, and has a nonsignificant *p*-

value of 0.8556. These statistics don't support the conclusion that the car country purchase response depends on the gender of the driver.

Note: The log-likelihood values of the null and constrained hypotheses are not shown in the Fit Y By X report. If you are interested in seeing them, launch **Analyze > Fit Model** using the same variables: country as Y, sex as an effect. The resulting report has more detail.

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.15594	2	0.311876	0.8556
Full	298.29531			
Reduced	298.45125			
RSquare (U)		0.0005		
Observations (or Sum Wgts)		303		

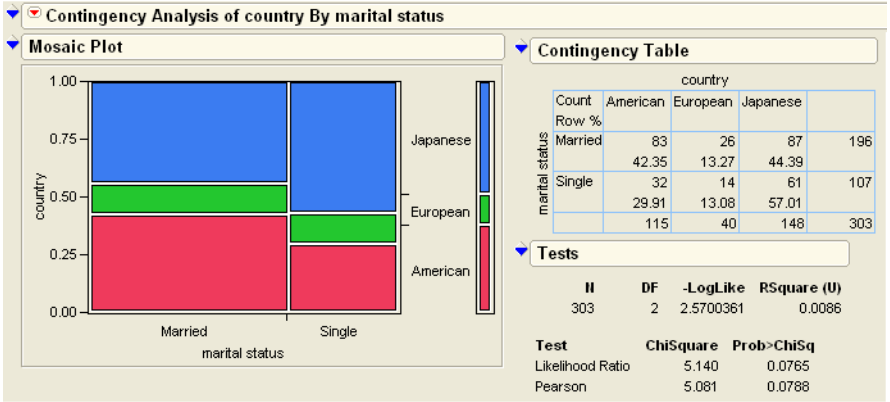
Converged by Gradient

If you want to think about the distribution of the estimates, then in each cell you can compare the actual proportion to the proportion expected under the hypothesis, square it, and divide by something close to its variance, giving a cell chi-square. The sum of these cell chi-square values is the Pearson chi-square statistic X^2, here also 0.312, which has a p-value of 0.8556. In this example, the Pearson chi-square happens to be the same as the likelihood ratio chi-square.

Car Brand by Marital Status

Let's look at the relationships of country to other categorical variables. In the case of marital status (**Figure 12.2**), there is a more significant result, with the p-value of 0.076. Married people are more likely to buy the American brands. Why? Perhaps because the American brands are generally larger vehicles, which make them more comfortable for families.

Figure 12.2 Mosaic Plot, Crosstabs, and Tests Table for country by marital status

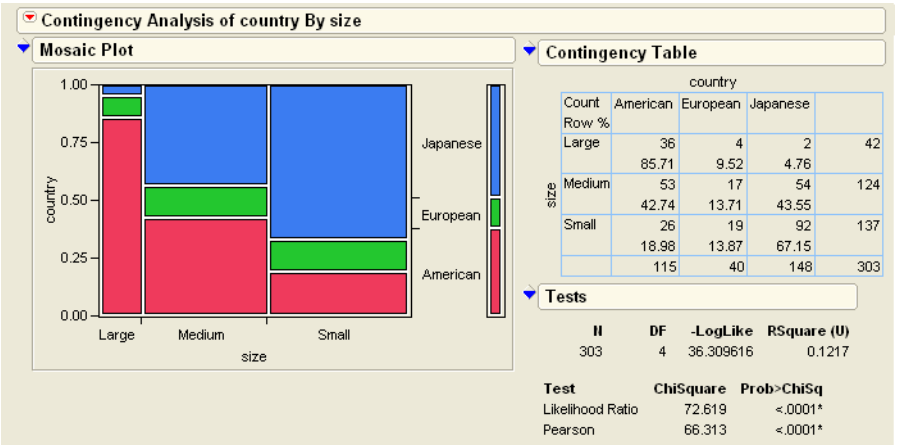


Car Brand by Size of Vehicle

If marital status is a proxy for size of vehicle, looking at country by size should give more direct information.

The Tests table for country by size (**Figure 12.3**) shows a strong relationship with a very significant chi-square. The Japanese dominate the market for small cars, the Americans dominate the market for large cars, and the European share is about the same in all three markets. The relationship is highly significant, with *p*-values less than 0.0001. The null hypothesis that car size and country are independent is easily rejected.

Figure 12.3 Mosaic Plot, Crosstabs, and Tests Table for country by size



Two-Way Tables: Entering Count Data

Often, raw categorical data is presented in textbooks in a *two-way table* like the one shown below. The levels of one variable are the rows, the levels of the other variable are the columns, with cells containing frequency counts. For example, data for a study of alcohol and smoking (based on Schiffman, 1982) is arranged in a two-way table, like this:

		Relapsed	
		Yes	**No**
Alcohol Consumption	**Consumed**	20	13
	Did Not Consume	48	96

This arrangement shows the two levels of alcohol consumption (“Consumed” or “Did Not Consume”) and levels of whether the subject relapsed and returned to smoking (reflected in the “Yes” column) or managed to stay smoke-free (reflected in the “No” column).

In the following discussion, keep the following things in mind:

- The two variables in this table do not fit neatly into independent and dependent classifications. The subjects in the study were not separated into two groups, with one group given alcohol, and the other not. The interpretation of the data, then, needs to be limited to association, and not cause-and-effect. The tests are regarded as tests of independence of two responses, rather than the marginal homogeneity of probabilities across samples.
- Because this is a 2×2 table, JMP produces Fisher’s Exact Test in its results. This test, in essence, computes exact probabilities for the data rather than relying on approximations.

Does it appear that alcohol consumption is related to the subject’s relapse status? Phrased more statistically, if you assume these variables are independent, are there surprising entries in the two-way table? To answer this question, we must know what values would be expected in this table, and then determine if there are observed results that are different than these expected values.

Expected Values Under Independence

To further examine the data, the following table shows the totals for the rows and columns of the two-way table. The row and column totals have been placed along the right and bottom margins of the table, and are therefore called *marginal totals.*

		Relapsed		
		Yes	**No**	**Total**
Alcohol Consumption	**Consumed**	20	13	33
	Did Not Consume	48	96	144
	Total	68	109	177

These totals aid in determining what values would be expected if alcohol consumption and relapse to smoking were not related.

As is usual in statistics, assume at first that there is no relationship between these variables. If this assumption is true, then the proportion of people in the “Yes” and “No” columns should be equal for each level of the alcohol consumption variable. If there was no effect for

consumption of alcohol, then we expect these values to be the same except for random variation. To determine the *expected value* for each cell, compute

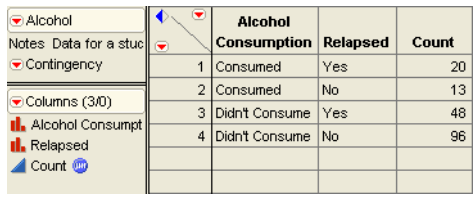
$$\frac{\text{Row total} \times \text{Column Total}}{\text{Table Total}}$$

for each cell. Instead of computing it by hand, let's enter the data into JMP to perform the calculations.

Entering Two-Way Data into JMP

Before two-way table data can be analyzed, it needs to be *flattened* or *stacked* so that it is arranged in two data columns for the variables, and one data column for frequency counts. These steps can be completed as follows:

- Select **File > New > Data Table** to create a new data table.
- Right-click on the title of Column 1 and select **Column Info.**
- Make the name of the column Alcohol Consumption and its data type **Character**. JMP automatically changes the modeling type to **Nominal**.
- Select **Columns > New Column** to create a second column in the data table. Repeat the above process to make this a character column named Relapsed.
- Create a third column named Count to hold the cell counts from the two-way table. Since this column will hold numbers, make sure its data type is **Numeric** and its modeling type is **Continuous**.
- Select **Rows > Add Rows** and add four rows to the table—one for each cell in the two-way table.
- Enter the data so that the data table looks like the one shown to the right.



	Alcohol Consumption	Relapsed	Count
1	Consumed	Yes	20
2	Consumed	No	13
3	Didn't Consume	Yes	48
4	Didn't Consume	No	96

These steps have been completed, and the resulting table is included in the sample data as Alcohol.jmp.

Testing for Independence

One explanatory note is in order at this point. Although the computations in this situation use the counts of the data, the statistical test deals with proportions. The *independence* we are concerned with is the independence of the probabilities associated with each cell in the table. Specifically, let

$$\rho_i = \frac{n_i}{n} \text{ and } \rho_j = \frac{n_j}{n}$$

where ρ_i and ρ_j are, respectively, the probabilities associated with each of the i rows and j columns. Now, let ρ_{ij} be the probability associated with the *cell* located at the ith row and jth column. The null hypothesis of independence is that $\rho_{ij}=\rho_i\rho_j$. Although the computations we present use counts, do not forget that the essence of the null hypothesis is about probabilities.

The test for independence is the X^2 statistic, whose formula is

$$\sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

To compute this statistic in JMP:

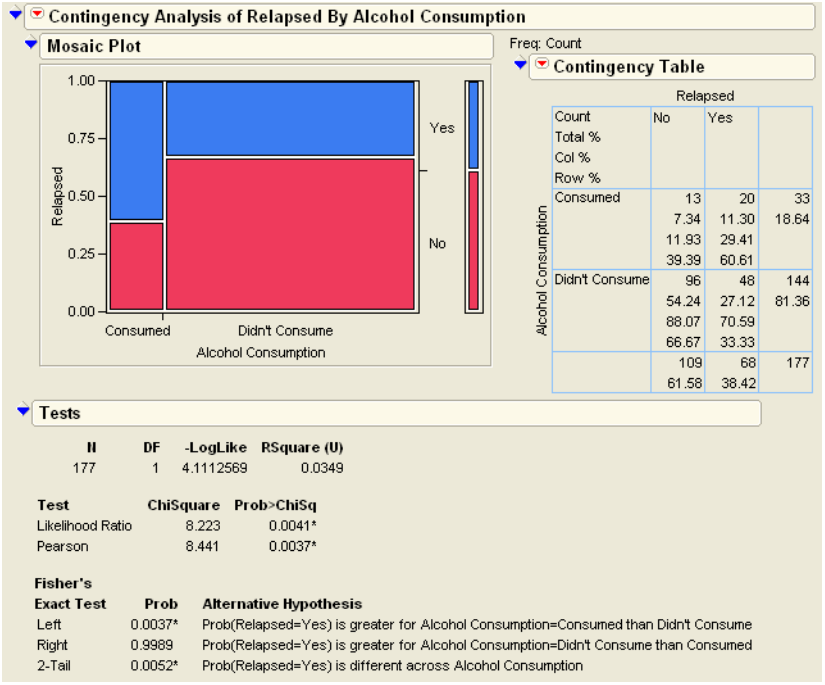
- Select **Analyze > Fit Y By X.** Set Alcohol Consumption as the **X**, Relapsed as the **Y**, and Count as the **Freq**.

This produces a report that contains the contingency table of counts, which should agree with the two-way table used as the source of the data. To see the information relevant to the computation of the X^2 statistic,

- Right-click inside the contingency table and uncheck **Row%**, **Col%** and **Total%**.
- Again, right-click inside the contingency table and make sure that **Count**, **Expected, Deviation**, and **Cell Chi-Square** are checked.

The Tests table (**Figure 12.4**) shows the Likelihood Ratio and Pearson Chi-square statistics. This represents the chi square for the table (the sum of all the cell chi-square values). The *p*-values for the chi-square tests are less than 0.05, so the null hypothesis is rejected. Alcohol consumption seems to be associated with whether the patient relapsed into smoking.

Figure 12.4 Contingency Report



The composition of the Pearson X^2 statistic can be seen cell by cell. The cell for "Yes" and "Consumed" in the lower left has an actual count of 20 and an expected count of 12.678. Their difference (deviation) is 7.322. This cell's contribution to the overall chi-square is

$$\frac{(20 - 12.678)^2}{12.678}$$

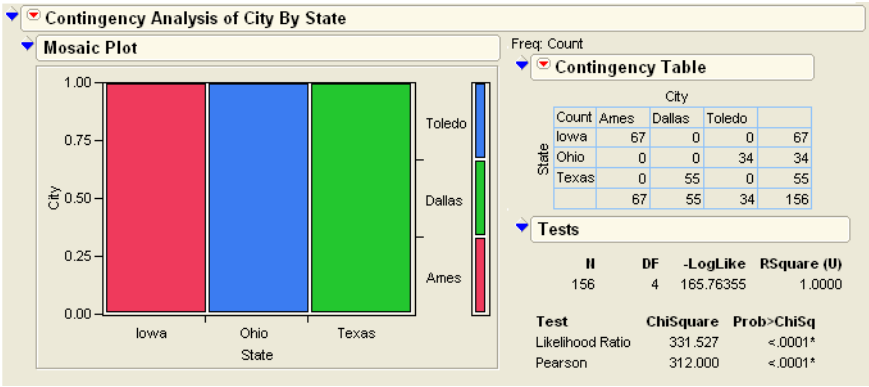
which is 4.22. Repeating this procedure for each cell shows the chi-square as 2.63 + 0.60 + 4.22 + 0.97, which is 8.441.

If You Have a Perfect Fit

If a fit is perfect, every response's category is predicted with probability 1. The response is completely determined by which sample it is in. In the other extreme, if the fit contributes nothing, then each distribution of the response in each sample subgroup is the same.

As an example, consider collecting information for 156 people on what city and state they live in. It's likely that one would think that there is a perfect fit between the city and the state of a person's residence. If the city is known, then the state is almost surely known. **Figure 12.5** shows what this perfect fit looks like.

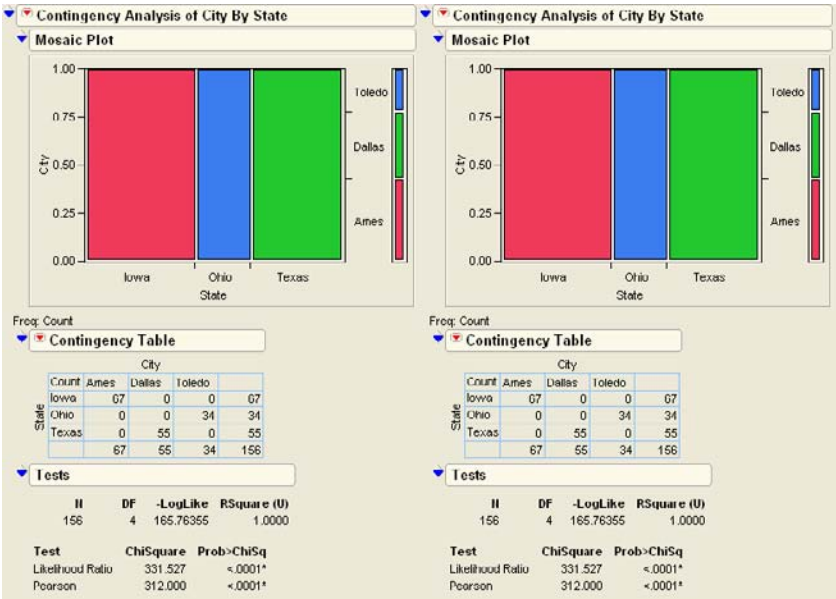
Figure 12.5 Mosaic Plot, Crosstabs, and Tests Table for City by State



Now suppose the analysis includes people from Austin, a second city in Texas. City still predicts state perfectly, but not the other way around (state does not predict city). Conducting these two analyses shows that the chi-squares are the same. They are invariant if you switch the Y and X variables. However, the mosaic plot, the attribution of the log-likelihood and R^2 are different (**Figure 12.6)**.

What happens if the response rates are the same in each cell as shown in **Figure 12.5**? Examine the artificial data for this situation and notice the mosaic levels line up perfectly and the chi-squares are zero.

Figure 12.6 Comparison of Plots, Tables, and Tests When X and Y Are Switched



Special Topic: Correspondence Analysis—Looking at Data with Many Levels

Correspondence analysis is a graphical technique that shows which rows or columns of a frequency table have similar patterns of counts. Correspondence analysis is particularly valuable when you have many levels, because it is difficult to find patterns in tables or mosaic plots with many levels.

The data table Mbtied.jmp has counts of personality types by educational level (Educ) and gender (Myers and McCaulley). The values of educational level are D for dropout, HS for high school graduate, and C for college graduate. Gender and Educ are concatenated to form the variable GenderEd. The goal is to determine the relationships between GenderEd and personality type. Remember, there is no implication of any cause-and-effect relationship because there is no way to tell whether personality affects education or education affects personality. The data can, however, show trends. The following example shows how correspondence analysis can help identify trends in categorical data:

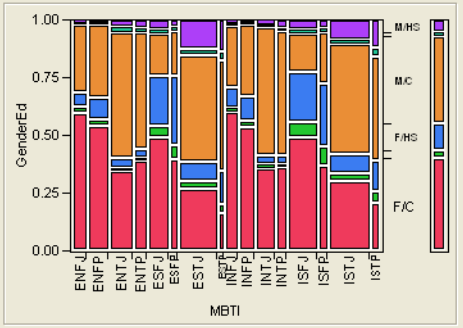
- Open the data table called Mbtied.jmp and choose **Analyze** > **Fit Y by X** with MBTI as X and GenderEd as Y, and Count as the Freq variable.

Now try to make sense out of the resulting mosaic plot and contingency table shown in **Figure 12.7**. It has with 96 cells—too big to understand at a glance. A correspondence analysis will clarify some patterns.

Figure 12.7 Mosaic Plot and Table for MBTI by GenderEd

Contingency Analysis of GenderEd By MBTI

Mosaic Plot



Contingency Table

GenderEd

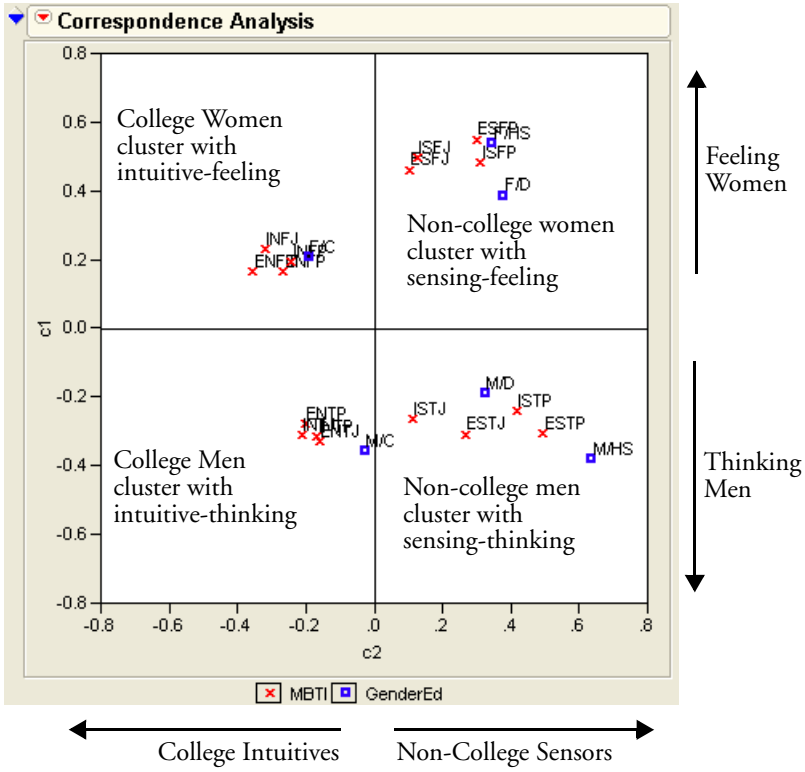
MBTI	F/C	F/D	F/HS	M/C	M/D	M/HS	
ENFJ	376	15	39	185	5	8	628
ENFP	453	20	81	264	7	9	834
ENTJ	321	13	39	508	23	26	930
ENTP	193	4	19	251	8	17	492
ESFJ	423	41	189	158	11	38	860
ESFP	142	23	108	69	6	12	360
ESTJ	400	47	126	683	38	186	1480
ESTP	40	10	35	117	10	33	245
INFJ	325	10	48	145	6	7	541
INFP	351	18	71	203	7	5	655
INTJ	274	12	32	425	8	14	765
INTP	168	7	18	248	6	17	464
ISFJ	575	78	256	200	20	46	1175
ISFP	144	32	106	85	8	13	388
ISTJ	485	49	129	761	30	137	1591
ISTP	66	17	41	144	12	38	318
	4736	396	1337	4446	205	606	11726

- Select **Correspondence Analysis** from the popup menu next to the Contingency Analysis title to see the plot in **Figure 12.8**.

The Correspondence Analysis plot organizes the row and column profiles in a two-dimensional space so that the X values that have similar Y profiles tend to cluster together, and the Y values that have similar X profiles tend to cluster together. In this case, you want to see how the GenderEd groups are associated with the personality groups.

This plot shows patterns more clearly. Gender and the Feeling(F)/Thinking(T) component form a cluster, and education clusters with the Intuition(N)/Sensing(S) personality indicator. The Extrovert(E)/Introvert(I) and Judging(J)/Perceiving(P) types do not separate much. The most separation among these is the Judging(J)/Perceiving(P) separation among the Sensing(S)/Thinking (T) types (mostly non-college men).

Figure 12.8 Correspondence Analysis Plot



The correspondence analysis indicates that the Extrovert/Introvert and Judging/Perceiving do not separate well for education and gender.

Continuous Factors with Categorical Responses: Logistic Regression

Suppose that a response is categorical, but the probabilities for the response change as a function of a continuous predictor. In other words, you are presented with a problem with a continuous X and a categorical Y. Some situations like this are the following:

- Whether you bought a car this year (categorical) as a function of your disposable income (continuous).
- The kind of car you bought (categorical) as a function of your age (continuous).

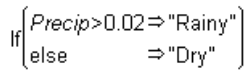
- The probability of whether a patient lived or died (categorical) as a function of blood pressure (continuous).

Problems like these call for *logistic regression*. Logistic regression provides a method to estimate the probability of choosing one of the response levels as a smooth function of the factor. It is called logistic regression because the S-shaped curve it uses to fit the probabilities is called the logistic function.

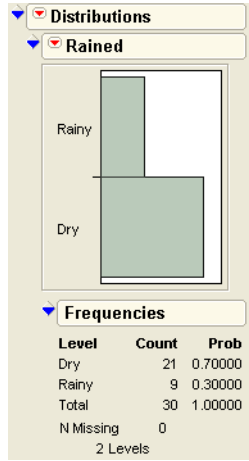
Fitting a Logistic Model

The Spring.jmp sample is a weather record for the month of April. The variable Precip measures rainfall.

- Open Spring.jmp.
- Add a variable named Rained to categorize rainfall using the formula shown to the right.



- Choose **Analyze > Distribution** to generate a histogram and frequency table of the Rained variable.



Out of the 30 days in April, there were 9 rainy days. Therefore, with no other information, you predict a 9/30 = 30% chance of rain for every day.

Suppose you want to increase your probability of correct predictions by including other variables. You may use morning temperature or barometric pressure to help make more informed predictions. Let's examine these cases.

In each case, the thing being modeled is the probability of getting one of several responses. The probabilities are constrained to add to 1. In the simplest situation, like this rain example, the response has two levels (a *binary* response). Remember that statisticians like to take logs of probabilities. In this case, what they fit is the difference in logs of the two probabilities as a linear function of the factor variable.

If p denotes the probability for the first response level, then $1-p$ is the probability of the second, and the linear model is written

$$\log(p) - \log(1-p) = b0 + b1*X \text{ or } \log(p/(1-p)) = b0 + b1*X$$

where $\log(p/(1-p))$ is called the *logit* of p or the *log odds-ratio*.

There is no error term here because the predicted value is not a response level; it is a probability distribution for a response level. For example, if the weatherman predicts a 90% chance of rain, you don't say he erred if it didn't rain.

The accounting is done by summing the negative logarithms of the probabilities attributed by the model to the events that actually did occur. So if p is the precipitation probability from the weather model, then the score is $-\log(p)$ if it rains, and $-\log(1-p)$ if it doesn't. A weather reporter that is a perfect predictor comes up with a p of 1 when it rains ($-\log(p)$ is zero if p is 1) and a p of zero when it doesn't rain ($-\log(1-p)=0$ if $p=0$). The perfect score is zero. No surprise $-\log(p) = 0$ means perfect predictions. If you attributed a probability of zero to an event that occurred, then the –log-likelihood would be infinity, a pretty bad score for a forecaster.

So the inverse logit of the model $b0+b1*X$ expresses the probability for each response level, and the estimates are found so as to maximize the likelihood. That is the same as minimizing the negative sum of logs of the probabilities attributed to the response levels that actually occurred for each observation.

You can graph the probability function as shown in **Figure 12.9**. The curve is just solving for p in the expression

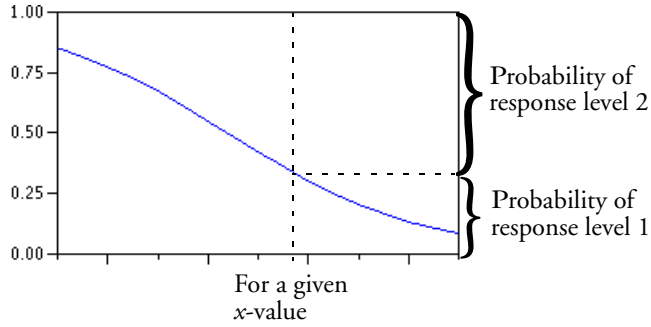
$$\log(p/(1-p)) = b0+b1*X$$

which is

$$p = 1/(1+\exp(-(b0+b1*X))$$

For a given value of X, this expression evaluates the probability of getting the first response. The probability for the second response is the remaining probability, $1-p$, because they must sum to 1.

Figure 12.9 Logistic Regression Fits Probabilities of a Response Level



To fit the rain column by temperature and barometric pressure for the spring rain data:

- Choose **Analyze > Fit Y by X** specifying the nominal column Rained as Y, and the continuous columns Temp and Pressure as the X variables.

The Fit Y by X platform produces a separate logistic regression for each predictor variable.

The cumulative probability plot on the left in **Figure 12.10** shows that the relationship with temperature is very weak. As the temperature ranges from 35 to 75, the probability of dry weather only changes from 0.73 to 0.66. The line of fit partitions the whole probability into the response categories. In this case, you read the probability of Dry directly on the vertical axis. The probability of Rained is the distance from the line to the top of the graph, which is 1 minus the axis reading. The weak relationship is evidenced by the very flat line of fit; the precipitation probability doesn't change much over the temperature range.

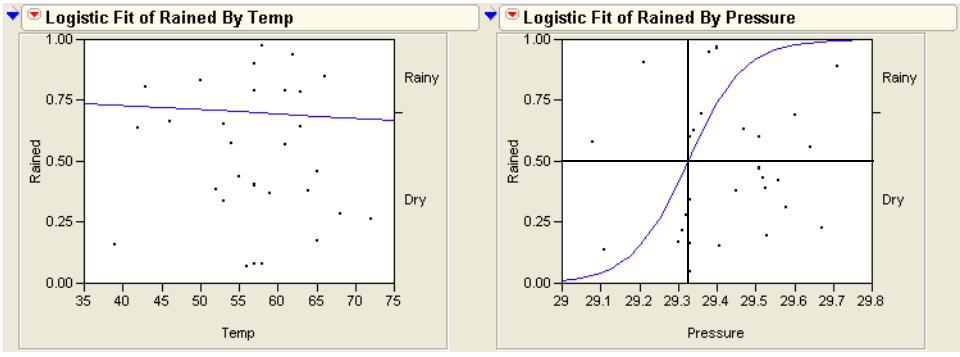
The plot on the right in **Figure 12.10** indicates a much stronger relationship with barometric pressure. When the pressure is 29.0 inches, the fitted probability of rain is near 100% (0 probability for Dry at the left of the graph). The curve crosses the 50% level at 29.32. (You can use the crosshair tool to see this.) At 29.8, the probability of rain drops to nearly zero (therefore, nearly 1.0 for Dry).

You can also add reference lines at the known X and Y values.

- Double-click on the Rained (Y) axis to bring up an axis modification dialog. Enter 0.5 as a reference line.
- Double-click on the Pressure (X) axis and enter 29.32 in the axis modification dialog.

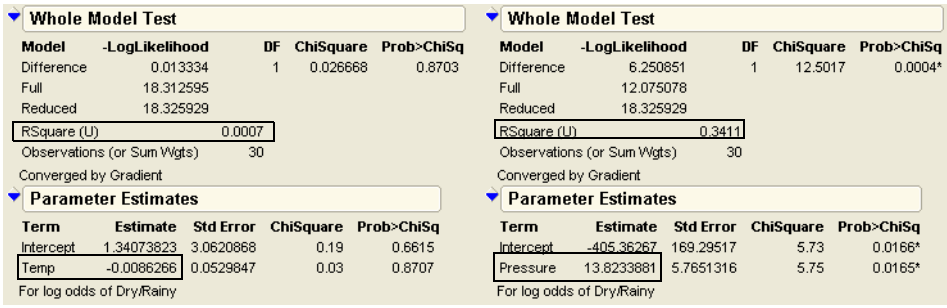
When both reference lines appear, they intersect on the logistic curve as shown in the plot on the right in **Figure 12.10**.

Figure 12.10 Cumulative Probability Plot for Discrete Rain Data



The Whole-Model Test table and the Parameter Estimates table reinforce the plot. The R^2 measure of fit, which can be interpreted on a scale of 0 to 100%, is only 0.07% (see **Figure 12.11**). A 100% R^2 would indicate a model that predicted outcomes with certainty. The likelihood ratio chi-square is not at all significant. The coefficient on temperature is a very small -0.008. The parameter estimates can be unstable because they have high standard errors with respect to the estimates.

Figure 12.11 Logistic Regression for Discrete Rain Data



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.013334	1	0.026668	0.8703
Full	18.312595			
Reduced	18.325929			
RSquare (U)	0.0007			
Observations (or Sum Wgts)	30			

Converged by Gradient

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1.34073823	3.0620868	0.19	0.6615
Temp	-0.0086266	0.0529847	0.03	0.8707

For log odds of Dry/Rainy

Rained by Temp

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	6.250851	1	12.5017	0.0004*
Full	12.075078			
Reduced	18.325929			
RSquare (U)	0.3411			
Observations (or Sum Wgts)	30			

Converged by Gradient

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-405.36267	169.29517	5.73	0.0166*
Pressure	13.8233881	5.7651316	5.75	0.0165*

For log odds of Dry/Rainy

Rained by Pressure

In contrast, the overall R^2 measure of fit with barometric pressure is 34%. The likelihood ratio chi-square is highly significant and the parameter coefficient for Pressure increased to 13.8 (**Figure 12.11**).

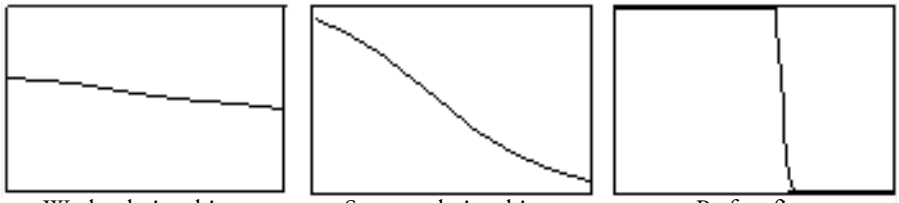
The conclusion is that if you want to predict whether the weather will be rainy, it doesn't help to know the temperature, but it does help to know the barometric pressure.

Degrees of Fit

The illustrations in **Figure 12.12** summarize the degree of fit as shown by the cumulative logistic probability plot.

When the fit is weak, the parameter for the slope term (X factor) in the model is small, which gives a small slope to the line in the range of the data. A perfect fit means that before a certain value of X, all the responses are one level, and after that value of X, all the responses are another level. A strong model can bet almost all of its probability on one event happening. A weak model has to bet conservatively with the background probability, less affected by the X factor's values.

Figure 12.12 Strength of Fit in Logistic Regression



Note that when the fit is perfect, as shown on the rightmost graph of **Figure 12.12**, the slope of the logistic line approaches infinity. This means that the parameter estimates are also infinite. In practice, the estimates are allowed to inflate only until the likelihood converges and are marked as unstable by the computer program. You can still test hypotheses, because they are handled through the likelihood, rather than using the estimate's (theoretically infinite) values.

A Discriminant Alternative

There is another way to think of the situation where the response is categorical and factor is continuous. You can reverse the roles of the Y and X and treat this problem as one of finding the distribution of temperature and pressure on rainy and dry days. Then, work backwards to obtain prediction probabilities. This technique is called *discriminant analysis*. (Discriminant analysis is discussed in detail in Chapter 18).

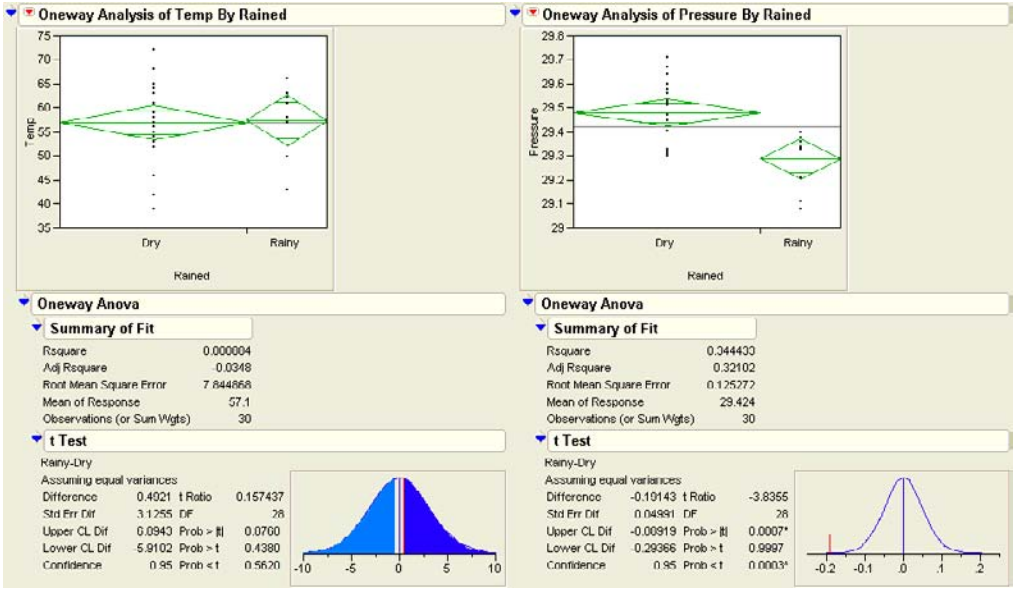
- For this example, open (or make active) the Spring.jmp data table. (Note: If you open it from scratch, you need to add the Rained variable as detailed on page 298.)
- Choose **Analyze > Fit Y by X** specifying Temp and Pressure as the Y variables and Rained as X.
- Select **Means/Anova/Pooled t** from the Display popup menu showing beneath the plots to see the results in **Figure 12.13**.

You can quickly see that the difference between the relationships of temperature and pressure to raininess. However, the discriminant approach is a somewhat strange way to go about this example and has some problems:

- The standard analysis of variance assumes that the factor distributions are Normal.
- Discriminant analysis works backwards: First, in the weather example you are trying to predict rain. But the ANOVA approach designates Rained as the independent variable, from which you can say something about the predictability of temperature and pressure.

Then, you have to reverse-engineer your thinking to infer raininess from temperature and pressure.

Figure 12.13 Temperature and Pressure as a Function of (Discrete) Rain Variable



Inverse Prediction

If you want to know what value of the *X* regressor yields a certain probability, you can solve the equation, $\log(p/(1-p)) = b0 + b1*X$, for *X*, given *p*. This is often done for toxicology situations, where the *X*-value for *p*=50% is called an *LD50* (Lethal Dose for 50%). Confidence intervals for these inverse predictions (called *fiducial confidence intervals*) can be obtained.

The Fit Model platform has an inverse prediction facility. Let's find the LD50 for pressure in the rain data—that is, the value of pressure that gives a 50% chance of rain.

- Choose **Analyze > Fit Model**.

When the Fit Model dialog appears,

- Select Rained in the variable selection list and assign it as Y. Select Pressure and assign it as a Model Effect.
- Click **Run Model**.

- When the platform appears, select **Inverse Prediction** from the popup menu on the analysis title bar.

This displays the dialog at the left in **Figure 12.14**.

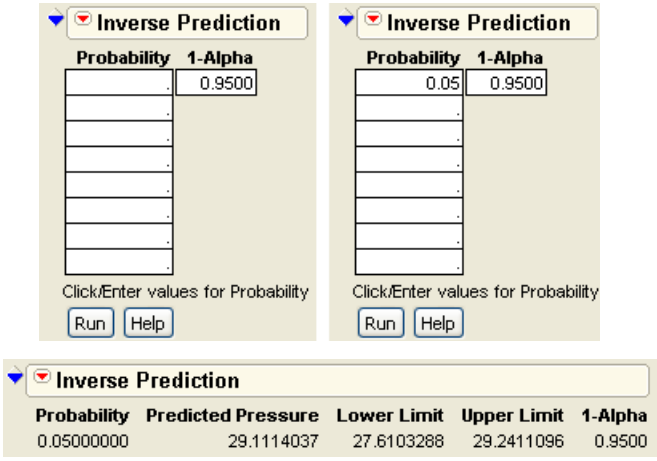
The Probability and 1–Alpha fields are editable. You can fill the dialog with any values of interest. The result is an inverse probability for each probability request value you entered, at the specified alpha level.

- For this example, enter 0.5 as the first entry in the Probability column.
- Click **Run**.

The Inverse Prediction table shown in **Figure 12.14** appears appended to the platform tables.

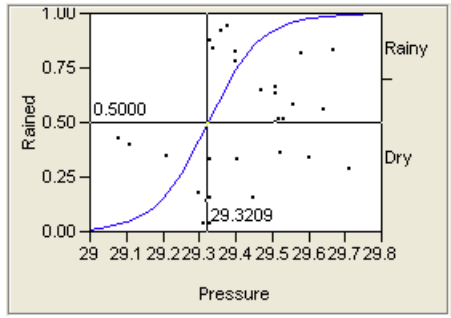
The inverse prediction computations say that there is a 50% chance of rain when the barometric pressure is 29.32.

Figure 12.14 Inverse Prediction Dialog



To see this on the graph,

- Get the crosshair tool from the Tools menu or toolbar. Click and drag the crosshair tool on the logistic plot until the horizontal line is at the 0.50 value of Rained.
- Hold the crosshair at that value and drag to the logistic curve. You can then read the Pressure value, slightly more than 29.3, on the *x*-axis.

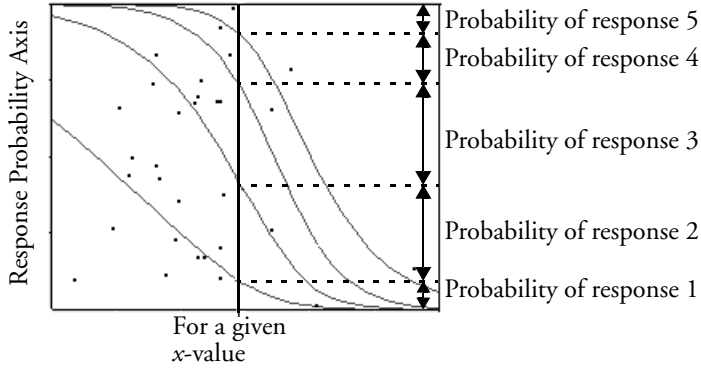


Polytomous Responses: More Than Two Levels

If there are more than two response categories, the response is said to be *polytomous* and a generalized logistic model is used. For the curves to be very flexible, you have to fit a set of linear model parameters for each of $r - 1$ response levels. The logistic curves are accumulated in such a way as to form a smooth partition of the probability space as a function of the regression model. In **Figure 12.15**, the probabilities are the distances between the curves, which add up to 1.

Where the curves are close together, the model is saying the probability of a response level is very low. Where the curves separate widely, the fitted probabilities are large.

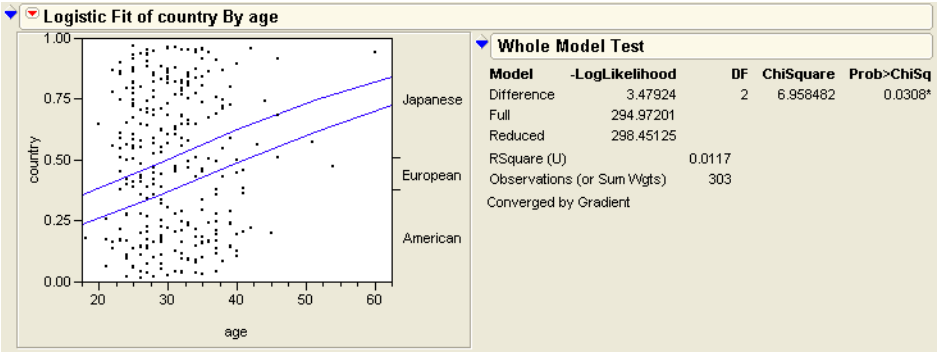
Figure 12.15 Polytomous Logistic Regression with Five Response Levels



For example, consider fitting the probabilities for country with the carpoll.jmp data as a smooth function of age. The result (**Figure 12.16**) shows the relationship where younger

individuals tend to buy more Japanese cars and older individuals tend to buy more American cars. Note the double set of estimates (two curves) needed to describe three responses.

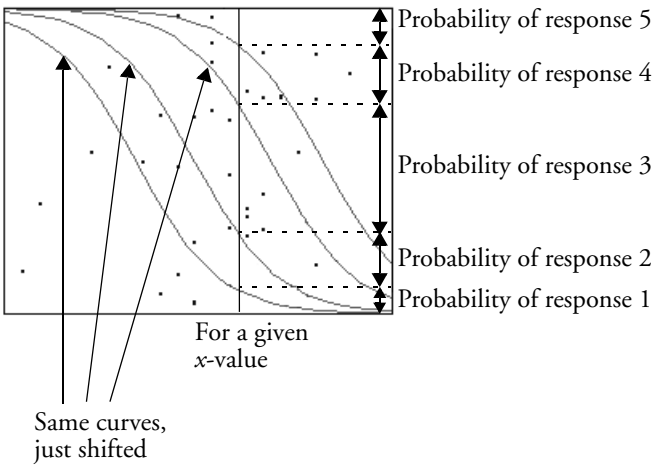
Figure 12.16 Cumulative Probability Plot and Logistic Regression for Country by Age



Ordinal Responses: Cumulative Ordinal Logistic Regression

In some cases, you don't need the full generality of multiple linear model parameter fits for the $r - 1$ cases, but can assume that the logistic curves are the same, only shifted by a different amount over the response levels. This means that there is only one set of regression parameters on the factor, but $r - 1$ intercepts for the r responses.

Figure 12.17 Ordinal Logistic Regression Cumulative Probability Plot



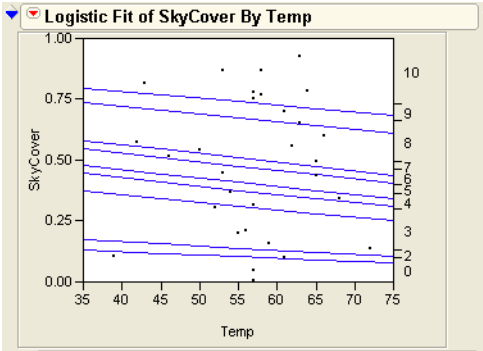
The logistic curve is actually fitting the sum of the probabilities for the responses at or below it, so it is called a *cumulative* ordinal logistic regression. In the Spring data table, there is a column called SkyCover with values 1 to 10.

First, note that you don't need to treat the response as nominal because the data have a natural order. Also, in this example, there is not enough data to support the large number of parameters needed by a 10-response level nominal model. Instead, use a logistic model that fits SkyCover as an ordinal variable with the continuous variables Temp and Humid1: PM.

- Change the modeling type of the SkyCover column to Ordinal by clicking the icon next to the column name in the Columns Panel, located to the left of the data grid.
- Chose **Analyze > Fit Y by X** and specify SkyCover as Y, and columns Temp and Humid1:PM as the X variables.

Figure 12.18 indicates that the relationship of SkyCover to Temp is very weak, with an R^2 of 0.09%, fairly flat lines, and a nonsignificant chi-square. The direction of the relation is that the higher sky covers are more likely with the higher temperatures.

Figure 12.18 Ordinal Logistic Regression for Ordinal Sky Cover with Temperature



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.058689	1	0.117378	0.7319
Full	61.792601			
Reduced	61.851290			
RSquare (U)	0.0009			
Observations (or Sum Wgts)	30			

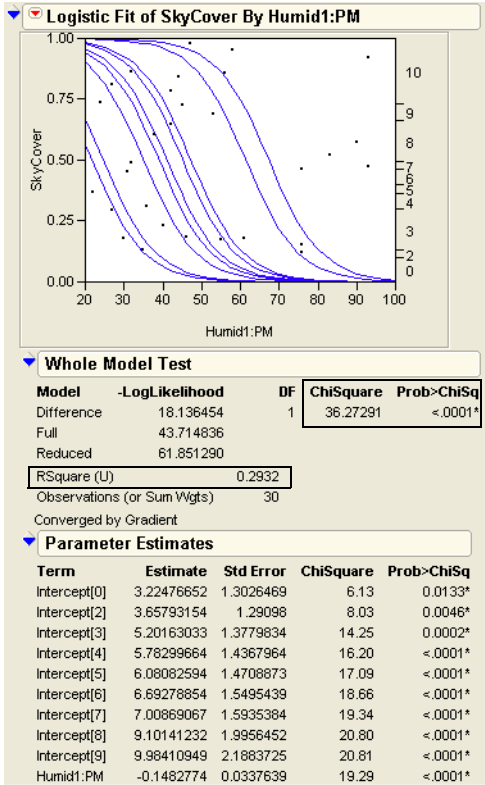
Converged by Objective

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[0]	-1.3689394	2.4768831	0.31	0.5805
Intercept[2]	-1.041078	2.4619505	0.18	0.6724
Intercept[3]	-0.0145838	2.444006	0.00	0.9952
Intercept[4]	0.2864849	2.4440495	0.01	0.9067
Intercept[5]	0.42731708	2.4446479	0.03	0.8612
Intercept[6]	0.69713892	2.446689	0.08	0.7757
Intercept[7]	0.83021599	2.4480727	0.12	0.7345
Intercept[8]	1.5258711	2.4590811	0.39	0.5349
Intercept[9]	1.84452034	2.4660911	0.56	0.4545
Temp	-0.0144402	0.0423949	0.12	0.7334

Figure 12.19 indicates that the relationship with humidity is quite strong. As the humidity rises to 70%, it predicts a 50% probability of a sky cover of 10. At 100% humidity, the sky cover will be almost certainly be 10. The R^2 is 29% and the likelihood ratio chi-square is highly significant.

Figure 12.19 Logistic Regression for Ordinal Sky Cover with Humidity



Note that the curves have bigger shifts in the popular 30% and 80% categories. Also, no data occurs for SkyCover = 2, so that is not even in the model.

There is a useful alternative interpretation to this ordinal model. Suppose you assume there is some continuous response with a random error component that the linear model is really fitting. But, for some reason, you can't observe the response directly. You are given a number that indicates which of r ordered intervals contains the actual response, but you don't know how the intervals are defined. You assume that the error term follows a logistic distribution, which has a similar shape to a Normal distribution. This case is identical to the ordinal cumulative logistic model, and the intercept terms are estimating the threshold points that define the intervals corresponding to the response categories.

Unlike the nominal logistic model, the ordinal cumulative logistic model is efficient to fit for hundreds of response levels. It can be used effectively for continuous responses when there are n unique response levels for n observations. In such a situation, there are $n - 1$ intercept parameters constrained to be in order, and there is one parameter for each regressor.

Surprise: Simpson's Paradox: Aggregate Data versus Grouped Data

Several statisticians have studied the “hot hand” phenomenon in basketball. The idea is that basketball players seem to have hot streaks, when they make most of their shots, alternating with cold streaks when they shoot poorly. The Hothand.jmp table contains the free throw shooting records for two Boston Celtics players (Larry Bird and Rick Robey) over the 1980-81 and 1981-82 seasons (Tversky and Gilovich, 1989).

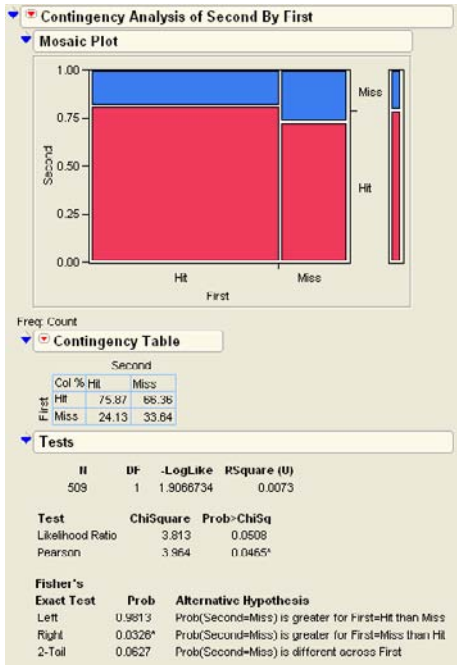
The null hypothesis is that two sequential free throw shots are independent. There are two directions in which they could be non-independent, the positive relationship (hot hand) and a negative relationship (cold hand).

The Hothand.jmp sample data have the columns First and Second (first shot and second shot) for the two players and a count variable. There are 4 possible shooting combinations: hit-hit, hit-miss, miss-hit, and miss-miss.

- Open Hothand.jmp.
- Choose **Analyze > Fit Y by X.** Select Second as Y, First as X, and Count as the Freq variable, then click **OK**.
- When the report appears, right-click in the contingency table and deselect all displayed numbers except **Col%.**

The results in **Figure 12.20** show that if the first shot is made, then the probability of making the second is 75.8%; if the first shot is missed the probability of making the second is 24.1%. This tends to support the hot hand hypothesis. The two chi-square statistics are on the border of 0.05 significance.

Figure 12.20 Crosstabs and Tests for Hothand Basketball Data

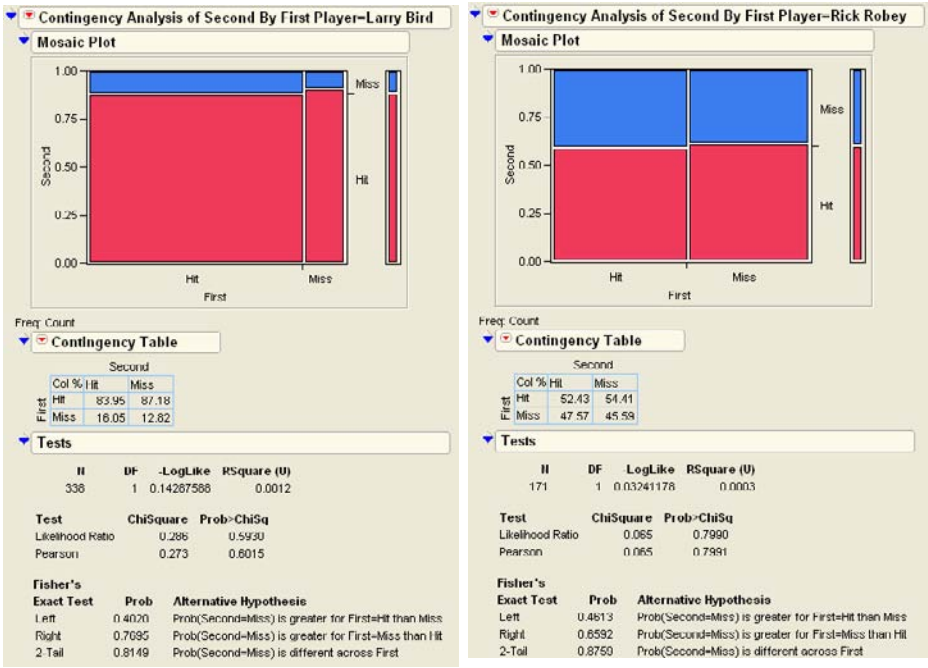


Does this analysis really confirm the hot hand phenomenon? A researcher (Wardrop 1995), looked at contingency tables for each player. You can do this using the By-groups.

- Again, Choose **Analyze > Fit Y by X** with Second as Y, First as X, and Count as the Freq variable.
- This time, assign Player as the **By** variable in the Launch dialog.

The results for the two players are shown in **Figure 12.21**.

Figure 12.21 Crosstabs and Tests for Grouped High-End Basketball Data



Contrary to the first result, both players shot better the second time after a miss than after a hit. So how can this be when the aggregate table gives the opposite results from both individual tables? This is an example of a phenomenon called *Simpson's paradox* (Simpson, 1951; Yule, 1903).

In this example, it is not hard to understand what happens if you think how the aggregated table works. If you see a hit on the first throw, the player is probably Larry Bird, and since he is usually more accurate he will likely hit the second basket. If you see a miss on the first throw, the player is likely Rick Robey, so the second throw will be less likely to hit. The hot hand relationship is an artifact that the players are much different in scoring percentages generally and populate the aggregate unequally.

A better way to summarize the aggregate data, taking into account these background relationships, is to use a blocking technique called the *Cochran-Mantel-Haenszel* test.

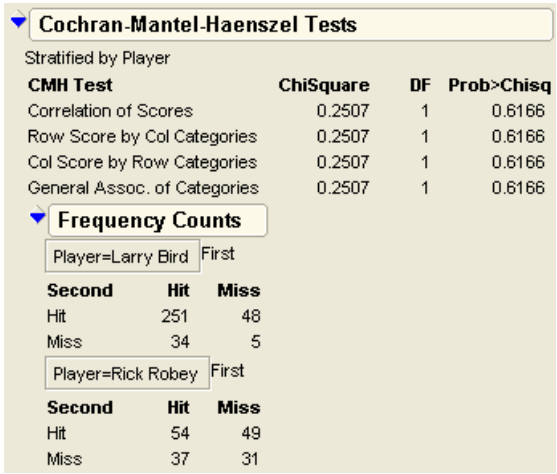
- Click the aggregate Fit Y by X report to make it active and choose **Cochran Mantel Haenszel** from the popup menu on the title bar.

A grouping dialog appears that lists the variables in the data table.

- Select Player as the grouping variable in this dialog and click **OK** to see the table in **Figure 12.22**.

These results are more accurate because they are based on the grouping variable instead of the ungrouped data. Based on these *p*-values, the null hypothesis of independence is not rejected.

Figure 12.22 Crosstabs and Tests for Grouped Hothand Basketball Data



Cochran-Mantel-Haenszel Tests

Stratified by Player

CMH Test	ChiSquare	DF	Prob>Chisq
Correlation of Scores	0.2507	1	0.6166
Row Score by Col Categories	0.2507	1	0.6166
Col Score by Row Categories	0.2507	1	0.6166
General Assoc. of Categories	0.2507	1	0.6166

Frequency Counts

Player=Larry Bird First

Second	Hit	Miss
Hit	251	48
Miss	34	5

Player=Rick Robey First

Second	Hit	Miss
Hit	54	49
Miss	37	31

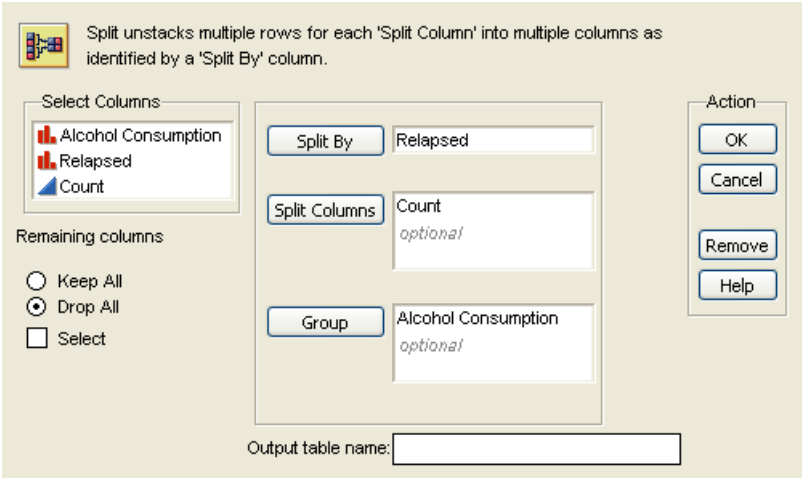
Generalized Linear Models

In recent years, *generalized linear models* have emerged as an alternative (and usually equivalent) approach to logistic models. There are two different formulations for generalized linear models that apply most often to categorical response situations: the *binomial* model and the *Poisson* model. To demonstrate both formulations, we use the Alcohol.jmp data table. **Figure 12.4** on page 293 shows that relapses into smoking are not independent of alcohol consumption, supported by a chi-square test ($G^2 = 8.22$, $p = 0.0041$).

The binomial approach is always applicable when there are only two response categories. To use it, we must first reorganize the data.

- Select Tables > Split.
- Assign variables as shown in **Figure 12.23**.

Figure 12.23 Split Dialog for Alcohol Data



- Add a new column that computes the total of the columns No + Yes, as shown in **Figure 12.24**.

Figure 12.24 Final Data Table

	Alcohol Consumption	No	Yes	Total
1	Consumed	13	20	33
2	Didn't Consume	96	48	144

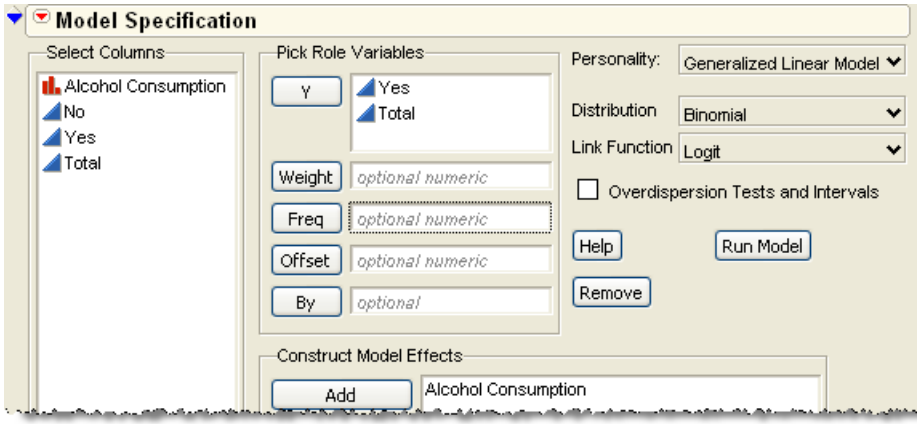
No+Yes

We're now ready to fit a model.

- Select **Analyze > Fit Model**.
- Change the **Personality** to **Generalized Linear Model**.
- Choose the **Binomial** distribution.
- Assign Yes and Total as **Y**.
- Remove No from the **Freq** column.
- Assign Alcohol Consumption as the effect in the model.

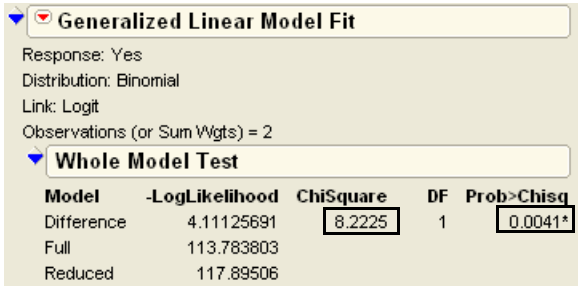
The Fit Model dialog should appear like the one in **Figure 12.25**.

Figure 12.25 Binomial Fit Model Dialog



Click **Run Model**.

In the resulting report, notice that the chi-square test and the *p*-value agree with our previous findings.



The other approach, the Poisson model, uses the original data. In this formulation, the count is portrayed as the response.

Select **Analyze > Fit Model**.

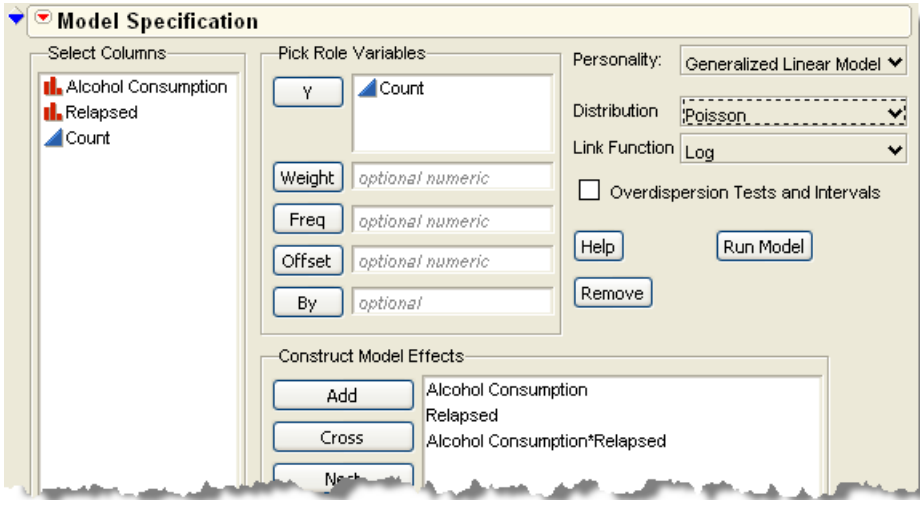
Choose the **Generalized Linear Model** personality.

Choose the **Poisson** distribution.

Assign Count as **Y**.

- Remove Count as a **Freq** variable. (It was automatically assigned since it has a Freq role in the data table, but that's not how we're using it here.)
- Select Alcohol Consumption and Relapsed in the original data table, then select **Macros > Full Factorial**.

This adds both main effects and their crossed term.



- Click **Run Model**.

Examine the Effects Tests section of the report. There are three tests reported, but we ignore the main effect tests because they don't involve a response category. The interaction between the two variables is the one we care about, and again it is identical with previous results.

Effect Tests

Source	DF	ChiSquare	Prob>Chisq
Alcohol Consumption	1	65.9387	<.0001*
Relapsed	1	0.4299	0.5121
Alcohol Consumption*Relapsed	1	8.2225	0.0041*

Since all these methods produce equivalent results (in fact, identical test statistics), they are interchangeable. The choice of method can then be a matter of convenience, generalizability, or comfort of the researcher.

Exercises

1. M.A. Chase and G.M. Dummer conducted a study in 1992 to determine what traits children regarded as important to popularity. Their data is represented in the file Children's Popularity.jmp. Demographic information was recorded, as well as the rating given to four traits assessing their importance to popularity: Grades, Sports, Looks, and Money.

   (a) Is there a difference based on gender on the importance given to making good grades?

   (b) Is there a difference based on gender on the importance of excelling in sports?

   (c) Is there a difference based on gender on the importance of good looks or on having money?

   (d) Is there a difference between Rural, Suburban, and Urban students on rating these four traits?

2. One of the concerns of textile manufacturers is the absorbency of materials that clothes are made out of. Clothes that can quickly absorb sweat (such as cotton) are often thought of as more comfortable than those that cannot (such as polyester). To increase absorbency, material is often treated with chemicals. In this fictional experiment, several swatches of denim were treated with two acids to increase their absorbency. They were then assessed to determine if their absorbency had increased or not. The investigator wanted to determine if there is a difference in absorbency change for the two acids under consideration. The results are presented in the following table:

		Acid	
		A	**B**
Absorbency	**Increased**	54	40
	Did Not Increase	25	40

   Does the researcher have evidence to say that there is a difference in absorbency between the two acids?

3. The taste of cheese can be affected by the additives that it contains. McCullagh and Nelder (1983) report a study (conducted by Dr. Graeme Newell) to determine the effects of four different additives on the taste of a cheese. The tasters responded by rating the taste of each cheese on a scale of 1 to 9. The results are in the data table Cheese.jmp.

   (a) Produce a mosaic plot to examine the difference of taste among the four cheese additives.

   (b) Do the statistical tests say that the difference amongst the additives is significant?

(c) Conduct a correspondence analysis to determine which of the four additives results in the best-tasting cheese.

4. The file Titanic.jmp contains information on the Passengers of the RMS Titanic. The four variables represent the class (first, second, third, and crew), age, sex, and survival status (yes or no) for each passenger. Use JMP to answer the following questions:

   (a) How many passengers were in each class?

   (b) How many passengers were children?

   (c) How many passengers were on the boat? How many survived?

   (d) Test the hypothesis that there is no difference in the survival rate among classes.

   (e) Test the hypothesis that there is no difference in the survival rate between males and females.

5. Do dolphins alter their behavior based on the time of day? To study this phenomenon, a marine biologist in Denmark gathered the data presented in Dolphins.jmp (Rasmussen, 1998). The variables represent different activities observed in groups of dolphins, with the Groups variable showing the number of groups observed.

   (a) Does this data show evidence that dolphins exhibit different behaviors during different times of day?

   (b) There is a caution displayed with the chi-square statistic. Should you reject the results of this analysis based on the warning?

13 Multiple Regression

Overview

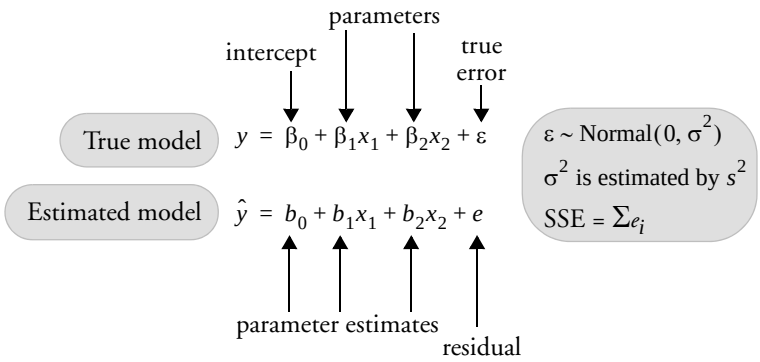
Multiple regression is the technique of fitting or predicting a response variable from a linear combination of several other variables. The fitting principle is least squares, the same as with simple linear regression.

Many regression concepts were introduced in previous chapters, so this chapter concentrates on showing some new concepts not encountered in simple regression: the point-by-point picture of a hypothesis test with the leverage plot, collinearity (the situation in which one regressor variable is closely related to another), and the case of exact linear dependencies.

Parts of a Regression Model

Linear regression models are the sum of the products of coefficient parameters and factors. In addition, linear models for continuous responses are usually specified with a Normally-distributed error term. The parameters are chosen such that their values minimize the sum of squared residuals. This technique is called estimation by *least squares*.

Figure 13.1 Parts of a Linear Model



Note in **Figure 13.1** the differences in notation between the assumed true model with unknown parameters and the estimated model.

response, Y

The *response* (or *dependent*) *variable* is the one you want to predict. Its estimates are the dependent variable, $\hat{y}$, in the regression model.

regressors, X's

The *regressors* (x) in the regression model are also called *independent variables*, *factors*, *explanatory variables*, and other discipline-specific terms. The regression model uses a linear combination of these effects to fit the response value.

coefficients, parameters

The fitting technique produces estimates of the parameters, which are the coefficients for the linear combination that defines the regression model.

intercept term

Most models have intercept terms to fit a constant in a linear equation. This is equivalent to having an x-variable that always has the value 1. The intercept is meaningful by itself only if it is meaningful to know the predicted value where all the regressors are zero. However, the

intercept plays a strong role in testing the rest of the model, because it represents the mean if all the other terms are zero.

error, residual

If the fit isn't perfect, then there is error left over. *Error* is the difference between an actual value and its predicted value. When speaking of true parameters, this difference is called *error*. When using estimated parameters, this difference is called a *residual*.

A Multiple Regression Example

Aerobic fitness can be evaluated using a special test that measures the oxygen uptake of a person running on a treadmill for a prescribed distance. However, it would be more economical to evaluate fitness with a formula that predicts oxygen uptake using simple measurements, such as running time and pulse measurements.

To develop such a formula, run time and pulse measurements were taken for 31 participants who each ran 1.5 miles. Their oxygen uptake, pulses, times, and other descriptive information was recorded. (Rawlings 1988, data courtesy of A.C. Linnerud). **Figure 13.2** shows a partial listing of the data, with variables Age, Weight, O2 Uptake (the response measure), Run Time, Rest Pulse, Run Pulse, and Max Pulse.

Figure 13.2 The Oxygen Uptake Data Table

Linnerud
Notes Linneruds Fitne
Columns (7/0)
Age
Weight
O2 Uptake
Run Time
Rest Pulse
Run Pulse
Max Pulse
Rows
All rows 31

	Age	Weight	O2 Uptake	Run Time	Rest Pulse	Run Pulse	Max Pulse
1	38	81.87	60.055	8.63	48	170	186
2	38	89.02	49.874	9.22	55	178	180
3	40	75.07	45.313	10.07	62	185	185
4	40	75.98	45.681	11.95	70	176	180
5	42	68.15	59.571	8.17	40	166	172
6	44	85.84	54.297	8.65	45	156	184
7	43	81.19	49.091	10.85	64	162	170
8	44	73.03	50.541	10.13	45	168	168
9	44	89.47	44.609	11.37	62	178	182
10	44	81.42	39.442	13.08	63	174	176
11	45	66.45	44.754	11.12	51	176	176

Now, investigate Run Time and Run Pulse as predictors of oxygen uptake (O2 Uptake):

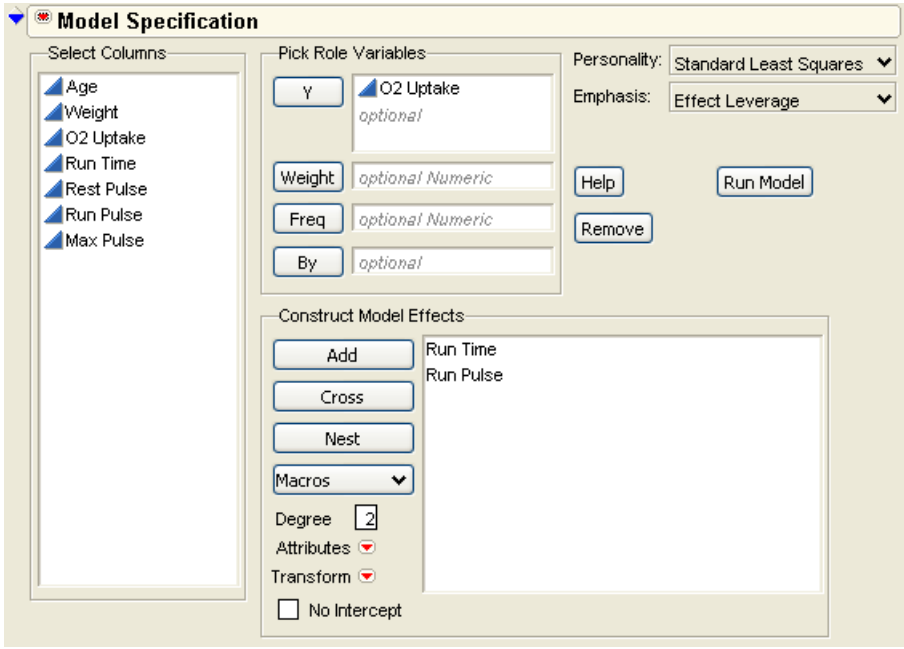
- Open the Linnerud.jmp sample data table. Choose **Analyze > Fit Model** to see the Fit Model dialog.

- Select the O2 Uptake column and click **Y** to make it the response (Y) variable. Select Run Time and Run Pulse, and click **Add** to make them the Effects in Model.

Your dialog should look like the one in **Figure 13.3**.

- Click **Run Model** to launch the platform.

Figure 13.3 Fit Model Dialog for Multiple Regression

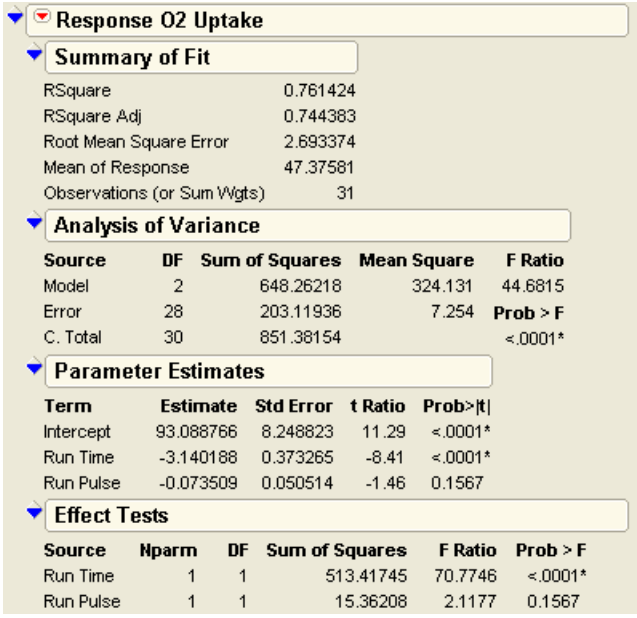


Now you have tables, shown in **Figure 13.4**, that report on the regression fit:

- The Summary of Fit table shows that the model accounted for 76% of the variation around the mean (R^2). The remaining residual error is estimated to have a standard deviation of 2.69 (Root Mean Square Error).
- The Parameter Estimates table shows Run Time to be highly significant ($p < 0.0001$), but Run Pulse is not significant ($p = 0.16$). Using these parameter estimates, the prediction equation is

  O2 Uptake = 93.089 – 3.14 Run Time – 0.0735 Run Pulse
- The Effect Test table shows details of how each regressor contributes to the fit.

Figure 13.4 Statistical Tables for Multiple Regression Example



Response O2 Uptake

Summary of Fit

RSquare	0.761424
RSquare Adj	0.744383
Root Mean Square Error	2.693374
Mean of Response	47.37581
Observations (or Sum Wgts)	31

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	648.26218	324.131	44.6815
Error	28	203.11936	7.254	Prob > F
C. Total	30	851.38154		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	93.088766	8.248823	11.29	<.0001*
Run Time	-3.140188	0.373265	-8.41	<.0001*
Run Pulse	-0.073509	0.050514	-1.46	0.1567

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Run Time	1	1	513.41745	70.7746	<.0001*
Run Pulse	1	1	15.36208	2.1177	0.1567

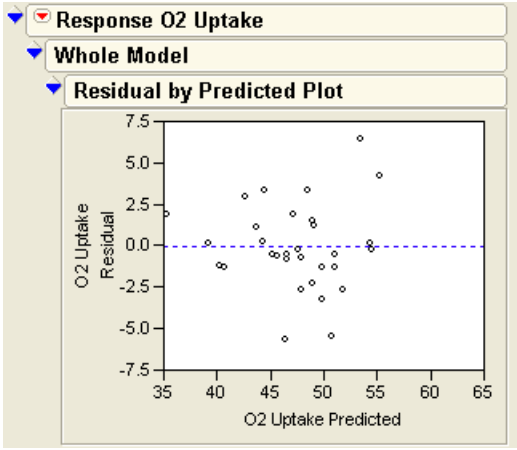
Residuals and Predicted Values

The residual is the difference between the actual response and the response predicted by the model. The residuals represent the error in the model. Points that don't fit the model have large residuals. It is helpful to look at a plot of the residuals and the predicted values, so JMP automatically appends a residual plot to the bottom of the Whole Model report, as shown in **Figure 13.5**.

Note: The points in **Figure 13.5** show as a medium-sized circle marker instead of the default dot. To change the marker or marker size,

- Select all of the rows in the data table (Ctrl+A or ⌘+A).
- Right-click anywhere in the plot frame and choose **Row Markers** from the menu that appears.
- Select the marker you want to use from the Markers palette.

Figure 13.5 Residual Plot for Multiple Regression Example

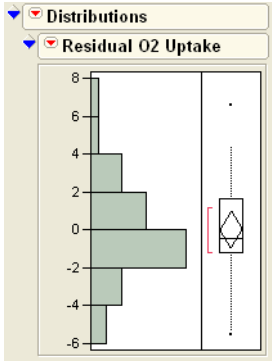


You can save these residuals as a column in the data table:

- Select **Save Columns > Residuals** from the triangle popup menu found on the title of the report.

The result is a new column in the data table called Residual O2 Uptake, which contains the residual for each response point. To examine them in more detail:

- Choose **Analyze > Distribution** and select the column of residuals to view the distribution of the residuals, as shown to the right.



Many researchers do this routinely to verify that the residuals are not too non-Normal to warrant concern about violating Normality assumptions.

You might also want to store the prediction formula from the multiple regression:

- Select **Save Columns > Prediction Formula** from the popup menu to create a new column in the data table called Predicted O2 Uptake. Its values are the calculated predicted values for the model.

- To see the formula used to generate the values in the column, right-click at the top of the Predicted O2 Uptake column and choose **Formula** from the menu that appears. The Formula Editor window opens and displays the formula

93.0087761+–3.1410876•Run Time+–0.0735095•Run Pulse

This formula defines a plane for O2 Uptake as a function of Run Time and Run Pulse. The formula stays in the column and is evaluated whenever new rows are added, or when variables used in the expression change their values. You can cut-and-paste or drag this formula into other JMP data tables.

The Analysis of Variance Table

The Whole-Model report consists of several tables that compare the full model fit to the simple mean model fit.

The Analysis of Variance table (shown to the right) lists the sums of squares and degrees of freedom used to form the whole model test:

Response O2 Uptake

Whole Model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	648.26218	324.131	44.6815
Error	28	203.11936	7.254	Prob > F
C. Total	30	851.38154		<.0001*

- The **Error Sum of Squares** (SSE) is 203.1. It is the sum of squared residuals after fitting the full model.
- The **C. Total Sum of Squares** is 851.4. It is the sum of squared residuals if you removed all the regression effects except for the intercept and, therefore, fit only the mean.
- The **Model Sum of Squares** is 648.3. It is the sum of squares caused by the regression effects, which measures how much variation is accounted for by the regressors. It is the difference between the Total Sum of Squares and the Error Sum of Squares.

The Error, C. Total, and Model sums of squares are the ingredients needed to test the whole-model null hypothesis that all the parameters in the model are zero except for the intercept (the simple mean model).

The Whole Model *F*-Test

To form an *F*-test,

1. Divide the Model Sum of Squares (648.3 in this example) by the number of terms (effects) in the model excluding the intercept. That divisor (2 in this case) is found in the column labeled DF (Degrees of Freedom). The result is the *Mean Square for the model*.
2. Divide the Error Sum of Squares (208.119 in this example) by the its associated degrees of freedom, 28, giving the *Mean Square for Error*.
3. Compute the *F*-ratio as the Model Mean Square divided by the Mean Square for Error.

The significance level, or *p*-value, for this ratio is then calculated for the proper degrees of freedom (2 used in the numerator and 28 used in the denominator). The *F*-ratio, 44.6815, in

the analysis of variance table shown above, is highly significant (p<0.0001), which lets us reject the null hypothesis and indicates that the model does fit better than simply the mean.

Whole-Model Leverage Plot

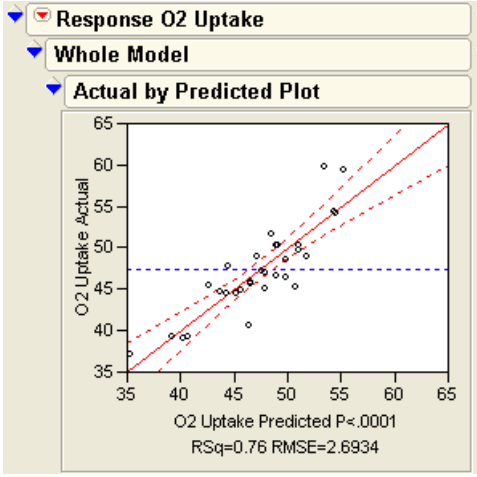
There is a good way to view this whole-model hypothesis graphically using a scatterplot of actual response values against the predicted values. The plot below shows the actual values versus predicted values for this aerobic exercise example.

A 45° line from the origin shows where the actual response and predicted response are equal. The vertical distance from a point to the 45° line of fit is the difference of the actual and the predicted values—the *residual error*. The mean is shown by the horizontal dashed line. The distance from a point to the horizontal line at the mean is what the residual would be if you removed all the effects from the model.

The portrayal of a plot that compares residuals from the two models in this way is called a *leverage plot*. The idea is to get a feel for how much better the sloped line fits than the horizontal line.

Superimposed on the plot are the confidence curves representing the 0.05-level whole-model hypothesis. If the confidence curves do not contain the horizontal line, the whole-model *F*-test is significant.

The leverage plot shown to the right is for the whole model, which includes both Run Time and Run Pulse.



Details on Effect Tests

You can explore the significance of an effect in a model by looking at the distribution of the estimate, or by looking at the contribution of the effect to the model:

- To look at the distribution of the estimate, compute its standard error. The standard error can be used either to construct confidence intervals for the parameter or to perform a *t*-test that the parameter is equal to some value (usually zero). The *t*-tests are given in the Parameter Estimates table. Confidence intervals can also be requested.
- If you take an effect out of the model, then the error sum of squared increases. That difference in sums of squares (with the effect included and excluded) can be used to

construct an *F*-test on whether the contribution of the effect to the model is significant. The *F*-tests are given in the Effect Tests table.

It turns out that *F*-tests and *t*-tests are equivalent. The square of the *t*-value in the Parameter Estimates table is the same as the *F*-statistic in the Effect Test table. For example, the square of the *t*-ratio (8.41) for Run Time is 70.77, which is the *F*-ratio for Run Time.

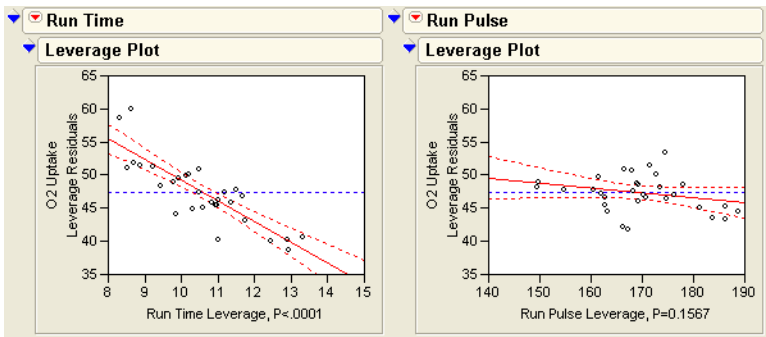
Effect Leverage Plots

Scroll to the right on the regression report to see plots detailing how each effect contributes to the model fit. The plots for the effect tests are *also* called leverage plots, although they are not the same as the leverage plots encountered in the whole-model test. The *effect leverage* plots (see **Figure 13.6**) show how each effect contributes to the fit after all the other effects have been included in the model. A leverage plot for a hypothesis test (an effect) is any plot with horizontal and sloped reference lines and points laid out having the following two properties:

- The distance from each point to the sloped line measures the residual for the full model. The sums of the squares of these residuals form the error sum of squares (SSE).
- The distance from each point to the horizontal line measures the residual for the restricted model without the effect. The sums of the squares of these residuals form the SSE for the constrained model (the model without the effect).

In this way, it is easy to see point by point how the sum of squares for the effect is formed. The difference in sums of squares of the two residual distances forms the numerator for the *F*-test for the effect.

Figure 13.6 Leverage Plots for Significant Effect and Nonsignificant Effect



Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Run Time	1	1	513.41745	70.7746	<.0001*
Run Pulse	1	1	15.36208	2.1177	0.1567

The leverage plot for an effect is interpreted in the same way as a simple regression plot. In fact, JMP superimposes a kind of 95% confidence curve on the sloped line that represents the full model. If the line is sloped significantly away from horizontal, then the confidence curves don't surround the horizontal line that represents the constrained model, and the effect is significant. Alternatively, when the confidence curves enclose the horizontal line, the effect is not significant at the 0.05 level.

The leverage plots in **Figure 13.6** show that Run Time is significant and Run Pulse is not. You can see the significance by how the points support (or don't support) the line of fit in the plot and by whether the confidence curves for the line cross the horizontal line.

There is a leverage plot for any kind of effect or set of effects in a model, or for any linear hypothesis. Leverage plots in the special case of single regressors are also known by the terms *partial plot*, *partial regression leverage plot*, and *added variable plot*.

Collinearity

Sometimes with a regression analysis, there is a close linear relationship between two or more effects. These two regressors are said to have a *collinearity* problem. It is a problem because the regression points do not occupy all the directions of the regression space very well. The fitting plane is not well supported in certain directions. The fit is weak in those directions, and the estimates become unstable, which means they are sensitive to small changes in the data.

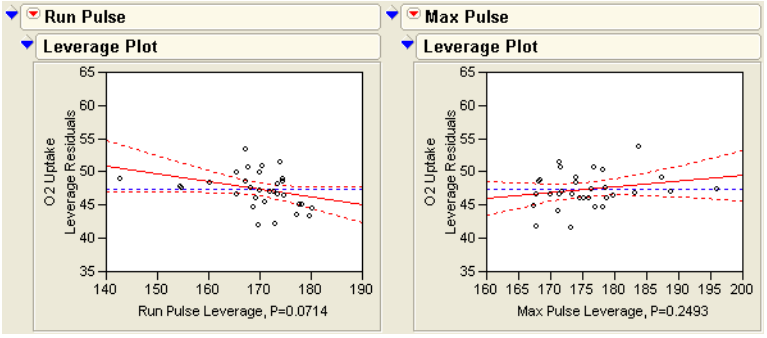
In the statistical results, this phenomenon translates into large standard errors for the parameter estimates and potentially large values for the parameter estimates themselves. This occurs because a small random change in the narrow direction can have a huge effect on the slope of the corresponding fitting plane. An indication of collinearity in leverage plots is when the points tend to collapse horizontally toward the center of the plot.

To see an example of collinearity, consider the aerobic exercise example with the correlated effects Max Pulse and Run Pulse:

- With the Linnerud exercise table active, choose **Analyze > Fit Model** (or click on the existing Fit Model dialog if it is still open).
- Complete the Fit Model dialog by adding Max Pulse as an effect in the model after Run Time and Run Pulse.
- Click **Run Model**.

When the new analysis report appears, scroll to the Run Pulse and Max Pulse leverage plots. Note in **Figure 13.7** that Run Pulse is very near the boundary of 0.05 significance, and thus the confidence curves almost line up along the horizontal line, without actually crossing it.

Figure 13.7 Leverage Plots for Effects in Model



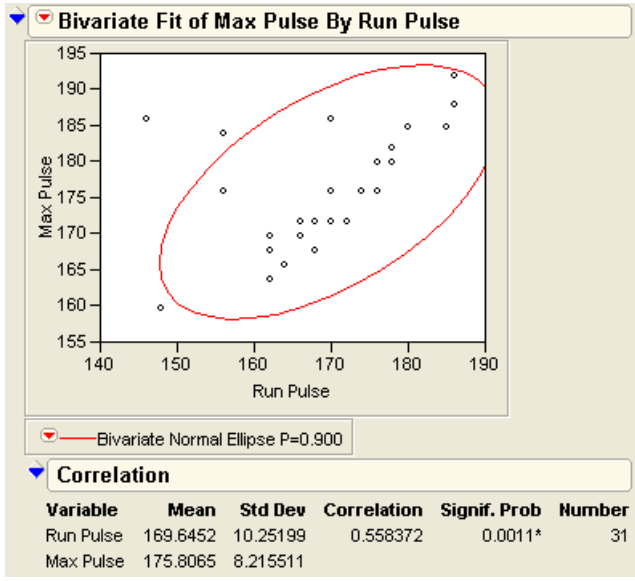
Now, as an example, let's change the relationship between these two effects by changing a few values to cause collinearity.

- Choose **Analyze > Fit Y by X**, selecting Max Pulse as the Y variable and Run Pulse as the X variable.

This produces a scatterplot showing the bivariate relationship.

- Select **Density Ellipse > .90** from the popup menu on the title bar.

You should now see the scatterplot with ellipse shown to the right. The **Density Ellipse** command also generates the Correlation table beneath the plot, which shows the current correlation to be 0.56.



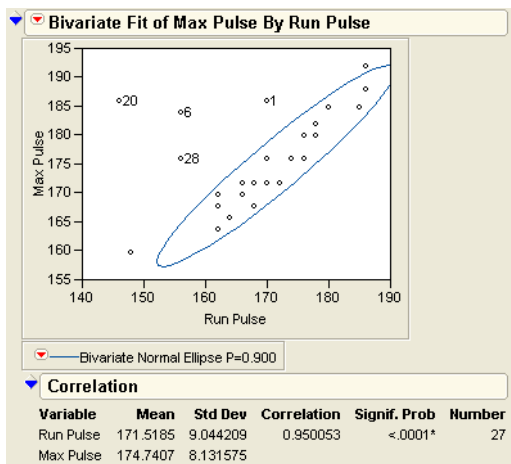
The variables don't appear to be collinear, since points are scattered in all directions. However, it appears that if four points are excluded, the correlation would increase dramatically. To see this, exclude these points and rerun the analysis.

- Highlight the points shown in the scatterplot below. You can do this by highlighting rows in the data table, or you can Shift-click the points in the scatterplot. With these points highlighted, choose **Rows > Label/Unlabel** to identify them.
- Choose **Rows > Exclude** while the rows are highlighted.

Notice in the spreadsheet that these points are now marked with a label tag and the not symbol(⊘).

- Again select a 0.90 **Density Ellipse** from the popup menu on the analysis title bar.

Now the ellipse and the Correlation table shows the relationship without the excluded points. The new correlation is 0.95, and the ellipse is much narrower, as shown in the plot to the right.



Now, run the regression model again to see the effect of excluding these points that created collinearity:

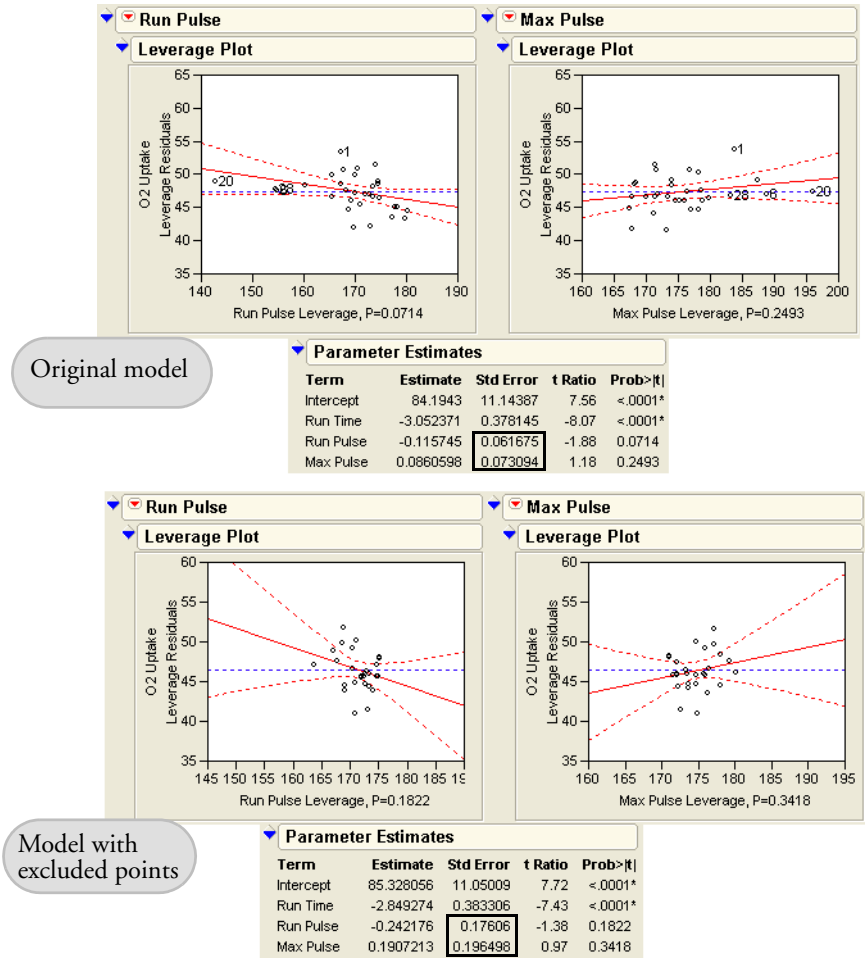
- Click **Run Model** again in the Fit Model dialog (with the same model as before).

Examine both the Parameter Estimates table and the leverage plots for Run Pulse and Max Pulse, comparing them with the previous report (see **Figure 13.8**).

The parameter estimates and standard errors for the last two regressors have more than doubled in size.

The leverage plots now have confidence curves that flare out because the points themselves collapse towards the middle. When a regressor suffers collinearity, the other variables have already absorbed much of that variable's variation, and there is less left to help predict the response. Another way of thinking about this is that there is less leverage of the points on the hypothesis. Points that are far out horizontally are said to have high leverage for the hypothesis test; points in the center have little leverage.

Figure 13.8 Comparison of Model Fits



Exact Collinearity, Singularity, Linear Dependency

Here we construct a variable to show what happens when there is an exact linear relationship, the extreme of collinearity, among the effects:

- Click on the Linnerud data table (or re-open it if it was accidentally closed).
- Choose **Columns > New Column**, to add a new variable (call it Run-Rest) to the data table.
- Use the Formula Editor to create a formula that computes the difference between Run Pulse and Rest Pulse.

Now run a model of O2 Uptake against all the response variables, including the new variable Run-Rest.

The report in **Figure 13.9** shows the signs of trouble. In the parameter estimates table, there are notations on Rest Pulse and Run Pulse that the estimates are biased, and on Run-Rest that it is zeroed. With exact linear dependency, the least squares solution is no longer unique, so JMP chooses the solution that zeroes out the parameter estimate for variables that are linearly dependent on previous variables. The Singularity Details report shows what the exact relationship is, in this case expressed in terms of Rest Pulse. The *t*-tests for the parameter estimates must now be interpreted in a conditional sense. JMP refuses to make tests for the non-estimable hypotheses for Rest Pulse, Run Pulse, and Run-Rest, and shows them with no degrees of freedom.

Figure 13.9 Report When There is a Linear Dependency

Response O2 Uptake

Singularity Details

Rest Pulse = Run Pulse - Run-Rest

Whole Model

Summary of Fit

RSquare	0.785309
RSquare Adj	0.720901
Root Mean Square Error	2.461207
Mean of Response	46.42937
Observations (or Sum Wgts)	27

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	443.15097	73.8585	12.1928
Error	20	121.15077	6.0575	Prob > F
C. Total	26	564.30174		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		105.03367	14.34429	7.32	<.0001*
Age		-0.219157	0.112002	-1.96	0.0645
Weight		-0.05691	0.059584	-0.96	0.3509
Run Time		-2.613413	0.420172	-6.22	<.0001*
Rest Pulse	Biased	-0.015947	0.07543	-0.21	0.8347
Run Pulse	Biased	-0.273798	0.174174	-1.57	0.1316
Max Pulse		0.1845045	0.191153	0.97	0.3460
Run-Rest	Zeroed	0	0	.	.

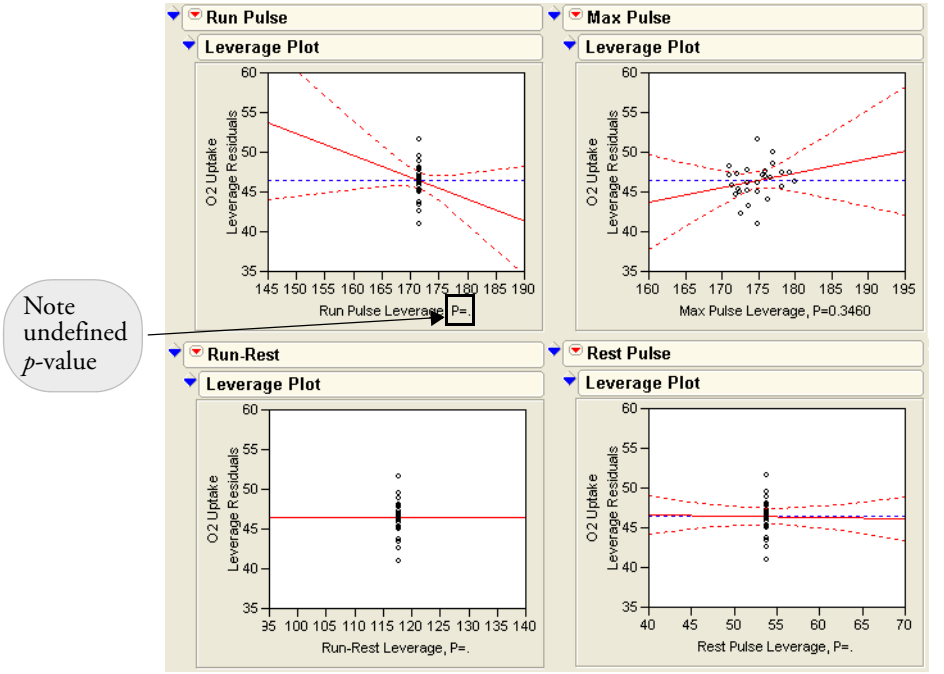
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
Age	1	1	23.19306	3.8288	0.0645	
Weight	1	1	5.52602	0.9123	0.3509	
Run Time	1	1	234.34659	38.6868	<.0001*	
Rest Pulse	1	0	0.00000	.	.	LostDFs
Run Pulse	1	0	0.00000	.	.	LostDFs
Max Pulse	1	1	5.64347	0.9316	0.3460	
Run-Rest	1	0	0.00000	.	.	LostDFs

You can see in the leverage plots for the three variables involved in the exact dependency, Rest Pulse, Run Pulse, and Run-Rest, that the points have completely collapsed horizontally—

nothing has any leverage for these effects (see **Figure 13.10**). However, you can still test the unaffected regressors, like Max Pulse, and make good predictions.

Figure 13.10 Leverage Plots When There is a Linear Dependency



The Longley Data: An Example of Collinearity

The Longley data set is famous, and is run routinely on most statistical packages to test accuracy of calculations. Why is it a challenge? Look at the data:

- Open the Longley.jmp data table.
- Choose **Analyze > Fit Model.**
- Enter Y as Y, and all the X columns as the model effects.
- Make sure **Effect Leverage** is selected as the Fit Personality.
- Click **Run Model** to see results shown in **Figure 13.11**.

Figure 13.11 Multiple Regression Report for Model with Collinearity

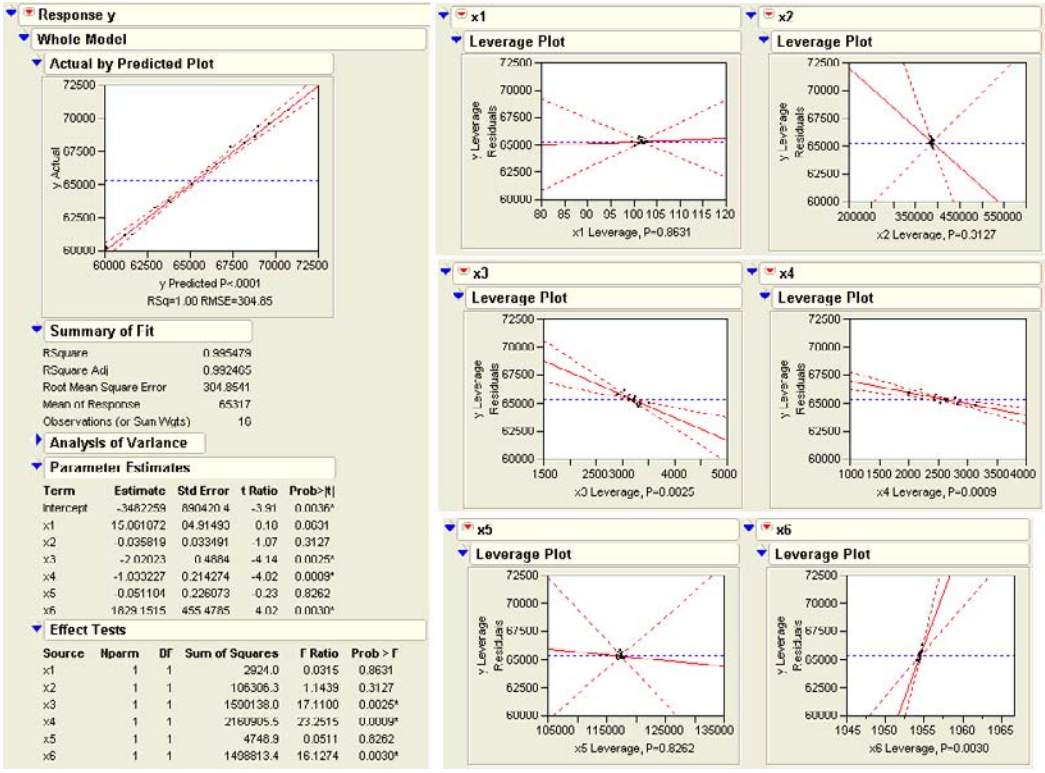


Figure 13.11 shows the whole-model regression analysis. Looking at this overall picture doesn't give information on which (if any) of the six regressors are affected by collinearity. It's not obvious what the problems are in this regression until you look at leverage plots for the effects, which show that x1, x2, x5, and x6 have collinearity problems. Their leverage plots appear very unstable with the points clustered at the center of the plot and the confidence lines showing no confidence at all.

The Case of the Hidden Leverage Point

Data were collected in a production setting where the yield of a process was related to three variables called Aperture, Ranging, and Cadence. Suppose you want to find out which of these effects are important, and in what direction:

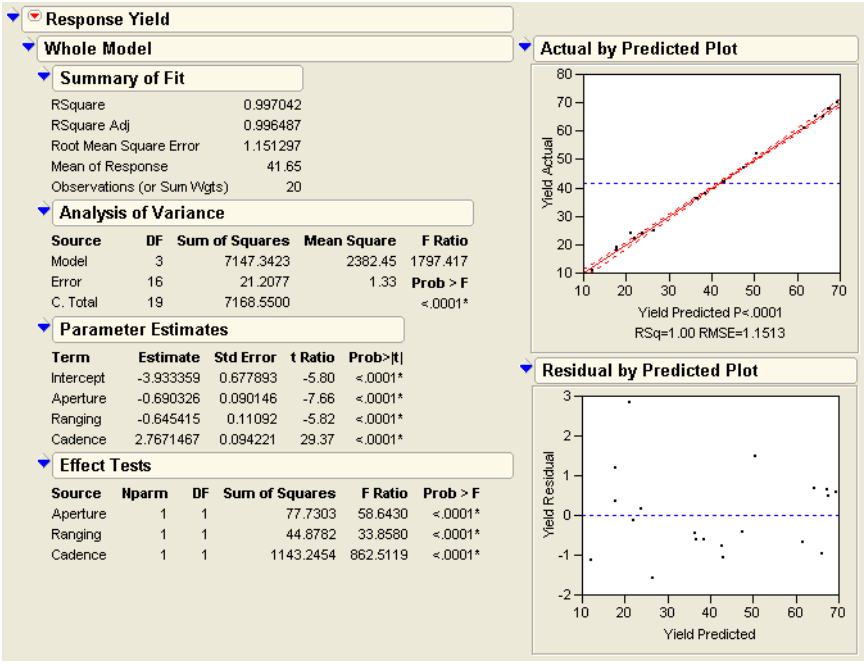
- Open the Ro.jmp data table and choose **Analyze > Fit Model**.

- Enter Yield as Y, and Aperture, Ranging, and Cadence as the Effects in Model.
- Click **Run Model**.

Everything looks fine in the tables shown in **Figure 13.12**. The Summary of Fit table shows an R^2 of 99.7%, which makes the regression model look like a great fit. All *t*-statistics are highly significant—but don't stop there.

JMP automatically creates the residual plot on the right in **Figure 13.12**.

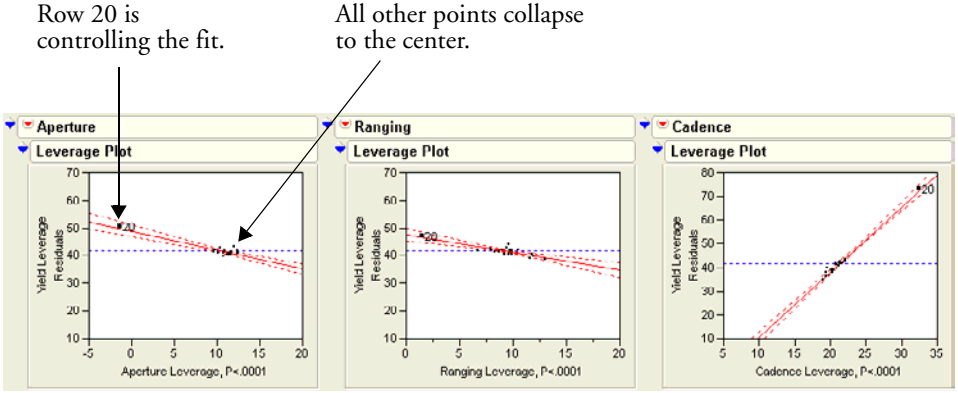
Figure 13.12 Tables and Plots for Model with Collinearity



JMP produces all the standard regression results, and many more graphics. For each regression effect, there is a leverage plot showing what the residuals would be without that effect in the model. Note in **Figure 13.13** that row 20, which appeared unremarkable in the whole-model leverage and residual plots, is far out into the extremes of the effect leverage plots.

It turns out that row 20 has monopolistic control of the estimates on all the parameters. All the other points appear wimpy because they track the same part of the shrunken regression space.

Figure 13.13 Leverage Plots That Detect Unusual Points



In a real analysis, row 20 would warrant special attention. Suppose, for example, that row 20 had an error, and was really 32 instead of 65:

- Change the value of Yield in row 20 from 65 to 32 and run the model again.

The Parameter Estimates for both the corrected table and incorrect table are shown to the right. The top table shows the parameter estimates computed from the data with an incorrect point. The bottom table has the corrected estimates. In high response ranges, the first prediction equation would give very different results than the second equation. The R^2 is again high and the parameter estimates are all significant—but every estimate is completely different even though only one point changed!

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-3.933359	0.677893	-5.80	<.0001*
Aperture	-0.690326	0.090146	-7.66	<.0001*
Ranging	-0.645415	0.11092	-5.82	<.0001*
Cadence	2.7671467	0.094221	29.37	<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.035267	0.246935	-0.14	0.8882
Aperture	1.7097579	0.032837	52.07	<.0001*
Ranging	1.7318952	0.040405	42.86	<.0001*
Cadence	0.2853072	0.034322	8.31	<.0001*

Mining Data with Stepwise Regression

Let's try a regression analysis on the O2 Uptake variable with a set of 30 randomly-generated columns as regressors. It seems like all results should be nonsignificant with random regressors, but that's not always the case.

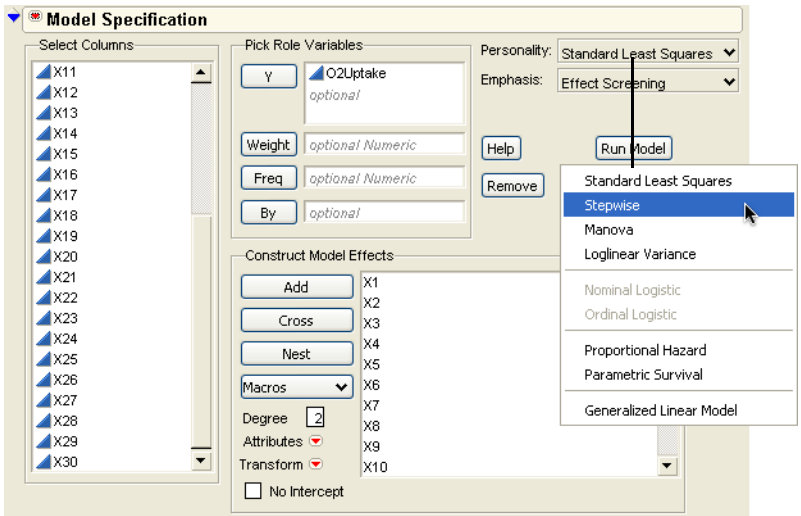
- Open the Linnrand.jmp data table.

The data table has 30 columns named X1 to X30, each of which has a uniform random number generator stored as a column formula. In other words, the data table is filled with random variables.

- Choose **Analyze > Fit Model** and use O2 Uptake as Y. Select X1 through X30 as the Effects in Model.
- Select **Stepwise** from the fitting personality popup menu, as shown in **Figure 13.14**, then click **Run Model**.

This stepwise approach launches a different regression platform, geared to playing around with different combinations of effects.

Figure 13.14 Fit Model Dialog for Stepwise Regression



To run a stepwise regression, use the control panel that appears after you run the model from the Fit Model dialog (see **Figure 13.15**).

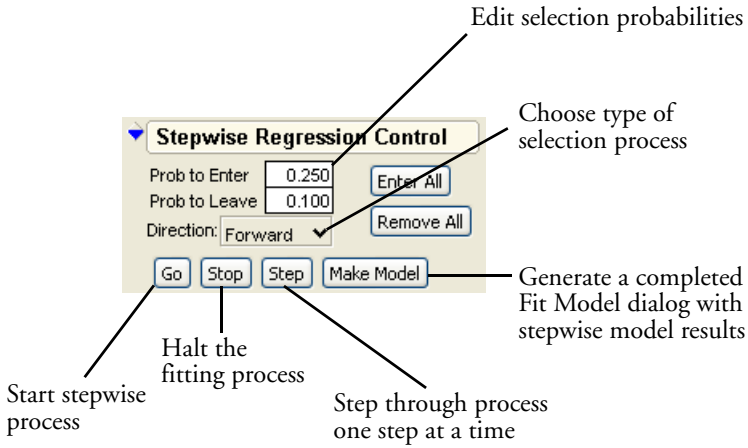
- Click **Go** in the Stepwise Regression Control panel to begin the stepwise variable selection process.

By default, stepwise runs a forward selection process. At each step, it adds the variable to the regression that is most significant. You can also select **Backward** or **Mixed** as the stepwise direction from the control panel popup menu.

The Forward process that we chose selects variables that are significant at the level specified in the **Prob to Enter** field on the control panel. (If we had used Backward of Mixed methods,

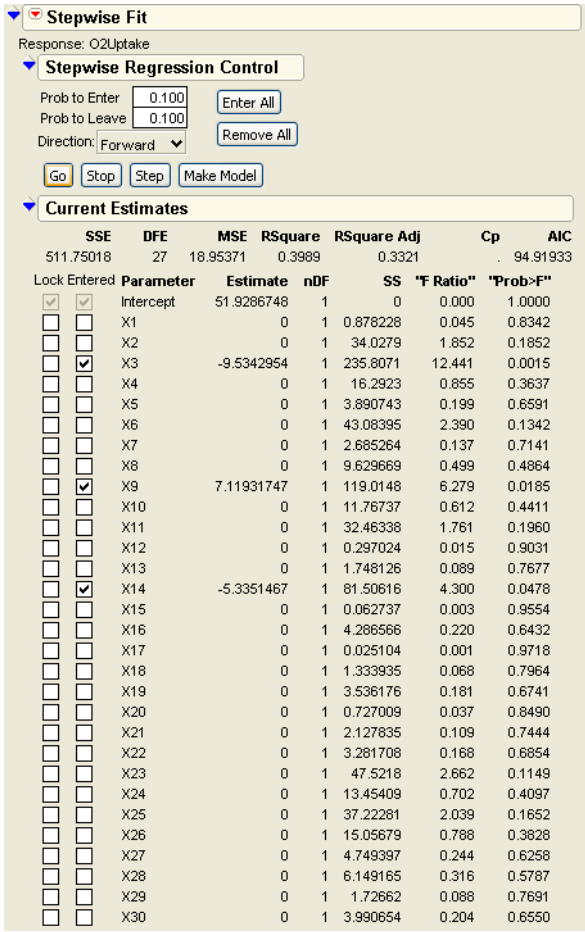
which add and remove variables, we'd specify something in the **Prob to Leave** field.) The process stops when no more variables are significant, and displays the Current Estimates table, shown in **Figure 13.16**. You also see an Iteration table (not shown here) that lists the order in which the variables entered the model.

Figure 13.15 Stepwise Regression Control Panel



Note: The example in **Figure 13.16** was run with **Forward** direction to enter variables with 0.1 **Prob to Enter**. If you run this problem, the results may differ because the X variables are generated by random number functions.

Figure 13.16 Current Estimates Table Showing Selected Variables



Stepwise Fit

Response: O2Uptake

Stepwise Regression Control

Prob to Enter 0.100 Enter All
Prob to Leave 0.100
Direction: Forward Remove All

Go Stop Step Make Model

Current Estimates

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
511.75018	27	18.95371	0.3989	0.3321	.	94.91933

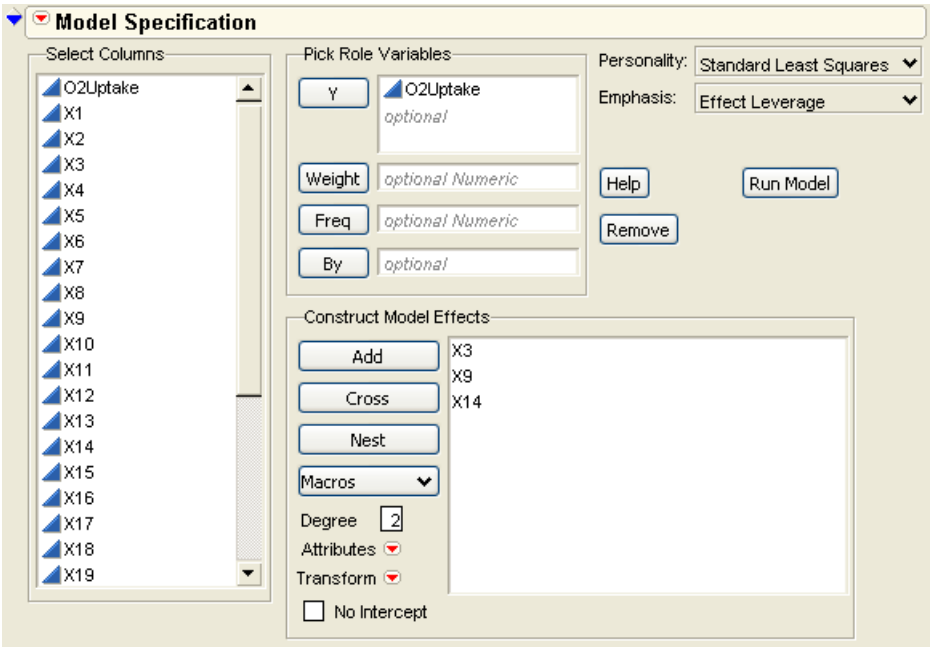
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	51.9286748	1	0	0.000	1.0000
☐	☐	X1	0	1	0.878228	0.045	0.8342
☐	☐	X2	0	1	34.0279	1.852	0.1852
☐	☑	X3	-9.5342954	1	235.8071	12.441	0.0015
☐	☐	X4	0	1	16.2923	0.855	0.3637
☐	☐	X5	0	1	3.890743	0.199	0.6591
☐	☐	X6	0	1	43.08395	2.390	0.1342
☐	☐	X7	0	1	2.685264	0.137	0.7141
☐	☐	X8	0	1	9.629669	0.499	0.4864
☐	☑	X9	7.11931747	1	119.0148	6.279	0.0185
☐	☐	X10	0	1	11.76737	0.612	0.4411
☐	☐	X11	0	1	32.46338	1.761	0.1960
☐	☐	X12	0	1	0.297024	0.015	0.9031
☐	☐	X13	0	1	1.748126	0.089	0.7677
☐	☑	X14	-5.3351467	1	81.50616	4.300	0.0478
☐	☐	X15	0	1	0.062737	0.003	0.9554
☐	☐	X16	0	1	4.286566	0.220	0.6432
☐	☐	X17	0	1	0.025104	0.001	0.9718
☐	☐	X18	0	1	1.333935	0.068	0.7964
☐	☐	X19	0	1	3.536176	0.181	0.6741
☐	☐	X20	0	1	0.727009	0.037	0.8490
☐	☐	X21	0	1	2.127835	0.109	0.7444
☐	☐	X22	0	1	3.281708	0.168	0.6854
☐	☐	X23	0	1	47.5218	2.662	0.1149
☐	☐	X24	0	1	13.45409	0.702	0.4097
☐	☐	X25	0	1	37.22281	2.039	0.1652
☐	☐	X26	0	1	15.05679	0.788	0.3828
☐	☐	X27	0	1	4.749397	0.244	0.6258
☐	☐	X28	0	1	6.149165	0.316	0.5787
☐	☐	X29	0	1	1.72662	0.088	0.7691
☐	☐	X30	0	1	3.990654	0.204	0.6550

- After the stepwise selection finishes selecting variables, click **Make Model** on the control panel.

The Fit Model dialog shown in **Figure 13.17** then appears, and you can run a standard least squares regression with the effects that were selected by the stepwise process as most active.

- Select **Run Model** on the Fit Model Dialog.

Figure 13.17 Fit Model Dialog Generated by the Make Model Option



When you run the model, you get the standard regression reports, shown to the right. The Parameter Estimates table shows all effects significant at the 0.05 level.

Summary of Fit

RSquare	0.398918
RSquare Adj	0.332131
Root Mean Square Error	4.353586
Mean of Response	47.37581
Observations (or Sum Wgts)	31

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	51.928675	2.366656	21.94	<.0001*
X3	-9.534295	2.70307	-3.53	0.0015*
X9	7.1193175	2.841091	2.51	0.0185*
X14	-5.335147	2.572754	-2.07	0.0478*

However, what has just happened is that we created enough data to generate a number of coincidences, and then gathered those coincidences into one analysis and ignored the rest of the variables. This is like gambling all night in a casino, but exchanging money only for those hands where you win. When you mine data to the extreme, you get results that are too good to be true.

Exercises

1. The file Grandfather Clocks.jmp (Smyth, 2000) contains data on grandfather clocks sold at open auction. Included are the selling price, age of the clock, and number of bidders on the clock. You are interested in predicting price based on the other two variables.

(a) Use the Fit Model platform to construct a model using age as the only predictor of price. What is the R^2 for this single predictor model?

(b) Add the number of bidders to the model. Does the R^2 increase markedly?

2. In *Gulliver's Travels*, the Lilliputians make an entire set of clothes for the (giant) Gulliver by only taking a few measurements from his body:

> *"The seamstresses took my measure as I lay on the ground, one standing at my neck, and another at my mid-leg, with a strong cord extended, that each held by the end, while a third measured the length of the cord with a rule of an inch long. Then they measured my right thumb, and desired no more; for by a mathematical computation, that twice round the thumb is once round the wrist, and so on to the neck and the waist, and by the help of my old shirt, which I displayed on the ground before them for a pattern, they fitted me exactly."* (Swift, 1735)

Is there a relationship among the different parts of the body? Body Measurements.jmp (Larner, 1996) contains measurements collected as part of a statistics project in Australia from 22 male subjects. In this exercise, you will construct a model to predict the mass of a person based on other characteristics.

(a) Using the Fit Model platform with personality **Standard Least Squares**, construct a model with mass as Y, and all other variables as effects in the model.

(b) Examine the resulting report and determine the effect that has the least significance to the model. In other words, find the effect with the largest Prob >F. Remove this effect from the model and re-run the analysis.

(c) Repeat part (b) until all effects have significance at the 0.05 level.

(d) Now, use the Fit Model platform with personality **Stepwise** to produce a similar model. Enter all effects into a Backward stepwise model with Prob to Leave set at 0.05. Compare this model to the one you generated in part (c).

3. Biologists are interested in determining factors that will predict the amount of time an animal will sleep during the day. To investigate the possibilities, Allison and Ciccetti (1976) gathered information on 62 different mammals. Their data is presented in the file Sleeping Animals.jmp.The variables describe body weight, brain weight, total time spent sleeping in two different states ("dreaming" and "non-dreaming"), life span, and gestation time. The researchers also calculated indices to represent predation (1 meaning unlikely to be preyed upon, 5 meaning likely to be preyed upon), exposure (1 meaning the animal sleeps in a well-protected den, 5 meaning most exposure), and an overall danger index, based upon predation, exposure, and other factors (1 meaning least danger from other animals, 5 meaning most danger).

(a) Use the Fit Y By X platform to examine the single-variable relationships between TotalSleep and the other variables. Which two variables look like they have the highest correlation with TotalSleep?

(b) If you remove NonDreaming from consideration, which two variables appear to be most correlated with TotalSleep?

(c) Construct a model using the two explanatory variables you found in part (b) and note its effectiveness.

(d) Construct a model using forward stepwise regression (still omitting Non-Dreaming), with 0.10 as the probability to enter and leave the model. Compare this model to the one you constructed in part (c).

(e) Construct two models using mixed stepwise regressions and compare it to the other models you have found. Which is the most effective at predicting total amount of sleep?

(f) Comment on the generalizability of this model. Would it be safe to use it to predict sleep times for a llama? Or a gecko? Explain your reasoning.

(g) Explore models that predict sleep in the dreaming and non-dreaming stages. Do the same predictors appear to be valid?

4. The file Cities.jmp contains a collection of pollution data for 52 cities around the country.

(a) Use the techniques of this chapter to build a model predicting Ozone for the cities listed. Use any of the continuous variables as effects.

(b) After you are satisfied with your model, determine whether there is an additional effect caused by the region of the country the city is in.

(c) In your model, interpret the coefficients of the significant effects.

(d) Comment on the generalizability of your model to other cities.

14 Fitting Linear Models

Overview

Several techniques, of increasing complexity, have been covered in this book. From fitting single means, to fitting multiple means, to fitting situations where the regressor is a continuous function, specific techniques have been demonstrated to address a wide variety of statistical situations. This chapter introduces a new approach involving *general linear models*, which will encompass all the models covered so far and extend to many more situations. They are all unified under the technique of least squares, fitting parameters to minimize the sum of squared residuals.

The techniques can be generalized even further to cover categorical response models, and other more specialized applications.

The General Linear Model

Linear models are the sum of the products of coefficient parameters and factor columns. The linear model is rich enough to encompass most statistical work. By using a coding system, you can map categorical factors to regressor columns. You can also form interactions and nested effects from products of coded terms. **Table 14.1** lists many of the situations handled by the general linear model approach. To read the model notation in **Table 14.1**, suppose that factors A, B, and C are categorical factors, and that X1, X2, and so forth, are continuous factors.

Table 14.1. Different Linear Models

Situation	Model Notation	Comments
One-way ANOVA	Y = A	Add a different value for each level.
Two-way ANOVA no interaction	Y = A, B	Additive model with terms for A and B.
Two-way ANOVA with interaction	Y = A, B, A*B	Each combination of A and B has a unique add-factor.
Three-way factorial	Y = A, B, A*B, C, A*C, B*C, A*B*C	For *k*-way factorial, 2^k–1 terms. The higher order terms are often dropped.
Nested model	Y = A, B[A]	The levels of B are only meaningful within the context of A levels, e.g., City[State], pronounced "city within state".
Simple regression	Y = X1	An intercept plus a slope coefficient times the regressor.
Multiple regression	Y = X1, X2, X3, X4,...	There can be dozens of regressors.
Polynomial regression	Y = X1, $X1^2$, $X1^3$, $X1^4$	Linear, quadratic, cubic, quartic,...
Quadratic response surface model	Y = X1, X2, X3 $X1^2$, $X2^2$, $X3^2$, X1*X2, X1*X3, X2*X3	All the squares and cross products of effects define a quadratic surface with a unique critical value where the slope is zero, which can be a minimum or a maximum, or a saddle point.

Situation	Model Notation	Comments
Analysis of covariance	Y = A, X1	Main effect (A), adjusting for the covariate (X1).
Analysis of covariance with different slopes	Y = A, X1, A*X1	Tests that the covariate slopes are different in different A groups.
Nested slopes	Y = A, X1[A]	Separate slopes for separate groups.
Multivariate regression	Y1, Y2 = X1, X2...	The same regressors affect several responses.
MANOVA	Y1, Y2, Y3 = A	A categorical variable affects several responses.
Multivariate repeated measures	sum and contrasts of (Y1 Y2 Y3)= A and so on.	The responses are repeated measurements over time on each subject.

Kinds of Effects in Linear Models

The richness of the general linear model results from the kind of effects you can include as columns in a coded model. Special sets of columns can be constructed to support various effects.

Intercept term

Most models have an intercept term to fit the constant in the linear equation. This is equivalent to having a regressor variable whose value is always 1. If this is the only term in the model, its purpose is to estimate the mean. This part of the model is so automatic that it becomes part of the background, the submodel to which the rest of the model is compared.

The only case in which intercepts are not used is the one in which the surface of fit must go through the origin. This happens in mixture models, for example. If you suppress the intercept term, then certain statistics (such as the whole-model *F*-test and the R^2) do not apply because the question is no longer of just fitting a grand mean submodel against a full model.

In some cases (like mixture models) the intercept is suppressed but there is a hidden intercept in the factors. This case is detected and the R^2 and *F* are reported as usual.

Continuous effects

These values are direct regressor terms, used in the model without modification. If all the effects are continuous, then the linear model is called multiple regression (See Chapter 13 for an extended discussion of multiple regression).

Categorical effects

The model must fit a separate constant for each level of a categorical effect. These effects are coded columns through an internal coding scheme, which is described in the next section. These are also called *main effects* when contrasted with compound effects, such as interactions.

Interactions

These are crossings of categorical effects, where you fit a different constant for each combination of levels of the interaction terms. Interactions are often written with an asterisk between terms, such as Age*Sex. If continuous effects are crossed, they are multiplied together.

Nested effects

Nested effects occur when a term is only meaningful in the context of another term, and thus is kind of a combined main effect and interaction with the term within which it is nested. For example, city is nested within state, because if you know you're in Chicago, you also know you're in Illinois. If you specify a city name alone, like Trenton, then Trenton, New Jersey could be confused with Trenton, Michigan. Nested effects are written with the upper level term in parentheses or brackets, like City[State].

It is also possible to have combinations of continuous and categorical effects, and to combine interactions and nested effects.

Coding Scheme to Fit a One-Way ANOVA as a Linear Model

When you include categorical variables in a model, JMP converts the categorical values (levels) into internal columns of numbers and analyzes the data as a linear model. The rules to make these columns are the *coding scheme*. These columns are sometimes called *dummy* or *indicator* variables. They make up an internal design matrix used to fit the linear model.

Coding determines how the parameter estimates are interpreted. However, note that the interpretation of parameters is different than the construction of the coded columns. In JMP, the categorical variables in a model are such that:

- There is an indicator column for each level of the categorical variable except the last level. An indicator variable is 1 for a row that has the value represented by that indicator, is –1 for rows that have the last categorical level (for which there is no indicator variable), and zero otherwise.
- A parameter is interpreted as the comparison of that level with the average effect across all levels. The effect of the last level is the negative of the sum of all the parameters for that effect. That is why this coding scheme is often called *sum-to-zero coding*.

Different coding schemes are the reason why different answers are reported in different software packages. The coding scheme doesn't matter in many simple cases, but it does matter in more complex cases, because it affects the hypotheses that are tested.

It's best to start learning the new approach covered in this chapter by looking at a familiar model. In Chapter 9, “Comparing Many Means: One-Way Analysis of Variance,” you saw the Drug.jmp data, comparing three drugs (Snedecor and Cochran, 1967). Let's return to this data table and see how the general linear model handles a one-way ANOVA.

The sample table called Drug.jmp (**Figure 14.1**) contains the results of a study that measured the response of 30 subjects after treatment by one of three drugs. First, look at the one-way analysis of variance given previously by the Fit Y by X continuous-by-nominal platform.

Figure 14.1 Drug Data Table with One-Way Analysis of Variance Report

	Drug	x	y
1	a	11	6
2	a	8	0
3	a	5	2
4	a	14	8
5	a	19	11
6	a	6	4

Oneway Analysis of y By Drug

Oneway Anova

Summary of Fit

Rsquare	0.227826
Adj Rsquare	0.170628
Root Mean Square Error	6.070878
Mean of Response	7.9
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Drug	2	293.6000	146.800	3.9831	0.0305*
Error	27	995.1000	36.856		
C. Total	29	1288.7000			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
a	10	5.3000	1.9198	1.3609	9.239
d	10	6.1000	1.9198	2.1609	10.039
f	10	12.3000	1.9198	8.3609	16.239

Std Error uses a pooled estimate of error variance

Now, do the same analysis using the Fit Model command.

- Open Drug.jmp, which has variables called Drug, y, and, x.

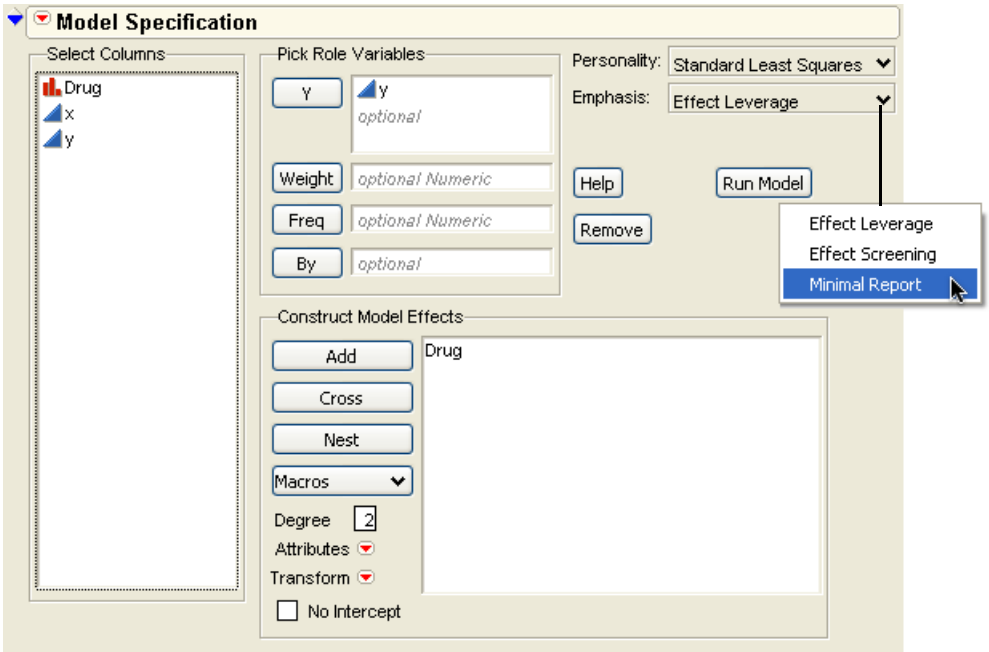
Drug has values “a”, “d”, and “f”, where “f” is really a placebo treatment. y is bacteria count after treatment, and x is a baseline count.

- Choose **Analyze > Fit Model**.
- When the Fit Model dialog appears, select y as the Y (response) variable. Select Drug and click **Add** to see it included as a Model Effect**.**
- Select **Minimal Report** from the Emphasis drop-down menu.

The completed dialog should look like the one in **Figure 14.2**.

- Click **Run Model** to see the analysis result in **Figure 14.3**.

Figure 14.2 Fit Model Dialog for Simple One-Way ANOVA



Now compare the reports from this new analysis (**Figure 14.3**) with the one-way ANOVA reports in **Figure 14.1**. Note that the statistical results are the same: the same R^2, ANOVA F-test, means, and standard errors on the means. (The ANOVA F-test is in both the Whole-Model Analysis of Variance table and in the Effect Test table because there is only one effect in the model.)

Although the two platforms produce the same results, the way the analyses were run internally was not the same. The Fit Model analysis ran as a regression on an intercept and two regressor variables constructed from the levels of the model main effect. The next section describes how this is done.

Figure 14.3 ANOVA Results Given by the Fit Model Platform

Response y

Summary of Fit

RSquare	0.227826
RSquare Adj	0.170628
Root Mean Square Error	6.070878
Mean of Response	7.9
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	293.6000	146.800	3.9831
Error	27	995.1000	36.856	Prob > F
C. Total	29	1288.7000		0.0305*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.9	1.108386	7.13	<.0001*
Drug[a]	-2.6	1.567494	-1.66	0.1088
Drug[d]	-1.8	1.567494	-1.15	0.2609

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Drug	2	2	293.60000	3.9831	0.0305*

Effect Details

[

Regressor Construction

The terms in the Parameter Estimates table are named according to what level each is associated with.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.9	1.108386	7.13	<.0001*
Drug[a]	-2.6	1.567494	-1.66	0.1088
Drug[d]	-1.8	1.567494	-1.15	0.2609

The terms are called Drug[a] and Drug[d]. Drug[a] means that the regressor variable is coded as 1 when the level is “a”, –1 when the level is “f”, and 0 otherwise. Drug[d] means that the variable is 1 when the level is “d”, –1 when the level is “f”, and 0 otherwise. You can write the notation for Drug[a] as ([Drug=a]-[Drug=f]), where [Drug=a] is a one-or-zero indicator of whether the drug is “a” or not. The regression equation then looks like this:

$$y = b_0 + b_1*((\text{Drug=a})-(\text{Drug=f})) + b_2*((\text{Drug=d})-(\text{Drug=f})) + \text{error}$$

So far, the parameters associated with the regressor columns in the equation are represented by the names b_0, b_1 and so forth.

Interpretation of Parameters

What is the interpretation of the parameters for the two regressors, now named in the equation as b_1, and b_2? The equation can be rewritten as

$$y = b_0 + b_1*[\text{Drug=a}] + b_2*[\text{Drug=d}] + (-b_1-b_2)*[\text{Drug=f}] + \text{error}$$

The sum of the coefficients (b_1, b_2, and $-b_1-b_2$) on the three indicators is always zero (again, sum-to-zero coding). The advantage of this coding is that the regression parameter tells you immediately how its level differs from the average response across all the levels.

Predictions Are the Means

To verify that the coding system works, calculate the means, which are the predicted values for the levels "a", "d", and "f", by substituting the parameter estimates shown previously into the regression equation

$$\text{Pred } y = b_0 + b_1*([\text{Drug=a}]-[\text{Drug=f}]) + b_2*([\text{Drug=d}]-[\text{Drug=f})$$

For the "a" level,

$$\text{Pred } y = 7.9 + -2.6*(1-0) + -1.8*(0-0) = 5.3, \text{ which is the mean } y \text{ for "a".}$$

For the "d" level,

$$\text{Pred } y = 7.9 + -2.6*(0-0) + -1.8*(1-0) = 6.1, \text{ which is the mean } y \text{ for "d".}$$

For the "f" level,

$$\text{Pred } y = 7.9 + -2.6*(0-1) + -1.8*(0-1) = 12.3, \text{ which is the mean } y \text{ for "f".}$$

Parameters and Means

Now, substitute the means symbolically and solve for the parameters as functions of these means. First, write the equations for the predicted values for the three levels, called A for "a", D for "d" and P for "f".

$$\text{MeanA} = b_0 + b_1*1 + b_2*0$$

$$\text{MeanD} = b_0 + b_1*0 + b_2*1$$

$$\text{MeanP} = b_0 + b_1*(-1) + b_2*(-1)$$

After solving for the *b*'s, the following coefficients result:

$$b_1 = \text{MeanA} - (\text{MeanA}+\text{MeanD}+\text{MeanP})/3$$

$$b_2 = \text{MeanD} - (\text{MeanA}+\text{MeanD}+\text{MeanP})/3$$

$$(-b_1-b_2) = \text{MeanP} - (\text{MeanA}+\text{MeanD}+\text{MeanP})/3$$

Each level's parameter is interpreted as how different the mean for that group is from the mean of the means for each level.

In the next sections you will meet the generalization of this and other coding schemes, with each coding scheme having a different interpretation of the parameters.

Keep in mind that the coding of the regressors does not necessarily follow the same rule as the interpretation of the parameters. (This is a result from linear algebra, resting on the fact that the inverse of a matrix is its transpose only if the matrix is orthogonal).

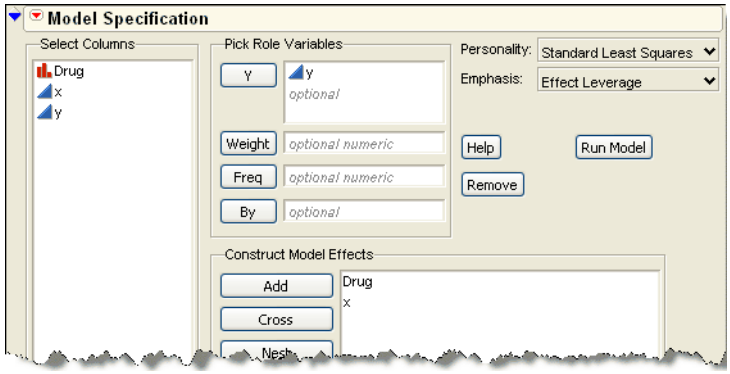
Overall, analysis using this coding technique is a way to convert estimating group means into an equivalent regression model. It's all the same least squares results using a different approach.

Analysis of Covariance: Putting Continuous and Classification Terms into the Same Model

Now take the previous drug example that had one main effect (Drug), and now add the other term (x) to the model. x is a regular regressor, meaning it is a continuous effect, and is called a *covariate*.

- In the Fit Model dialog window used earlier, click x and then **Add**. Now both Drug and x are effects, as shown in **Figure 14.4**.
- Click **Run Model** to see the results in **Figure 14.5**.

Figure 14.4 Fit Model Dialog for Analysis of Covariance

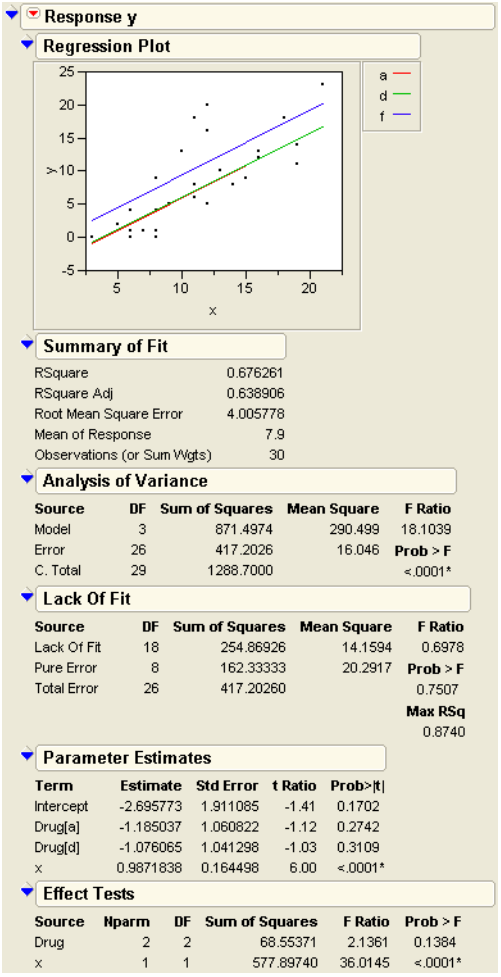


This new model is a hybrid between the ANOVA models with nominal effects and the regression models with continuous effects. Because the analysis method uses a coding scheme, the categorical term can be put into the model with the regressor.

The new results show that adding the covariate x to the model raises the R^2 from 22.78% (from **Figure 14.3**) to 67.62%. The parameter estimate for x is 0.987, which is not unexpected because the response is the pre-treatment bacteria count, and x is the baseline count before treatment. With a coefficient of nearly 1 for x, the model is really fitting the difference in bacteria counts. The difference in counts has a smaller variation than the absolute counts.

The *t*-test for x is highly significant. Because the Drug effect uses two parameters, refer to the *F*-tests to see if Drug is significant. The *F*-test has a null hypothesis that both parameters are zero. The surprising *p*-value for Drug is now 0.1384.

Figure 14.5 Analysis of Covariance Results Given by the Fit Model Platform

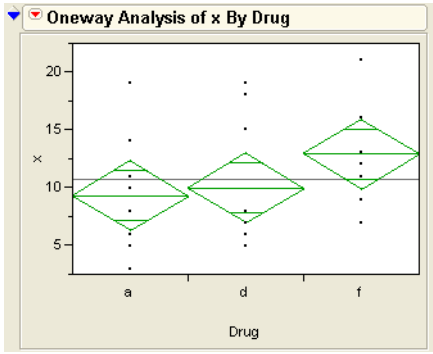


Drug, which was significant in the previous model, is no longer significant! How could this be? The error in the model has been reduced, so it should be easier for differences to be detected.

One possible explanation is that there might be a relationship between x and Drug.

- Choose **Analyze > Fit Y by X**, selecting x as Y and Drug as X.
- When the one-way platform appears, select the **Means/Anova** command from the popup menu on the title of the scatterplot.

Look at the results, shown to the right, to examine the relationship of the covariate x to Drug.



It appears that the drugs have not been randomly assigned—or, if they were, they drew an unlikely unbalanced distribution. The toughest cases (with the most bacteria) tended to be given the inert drug "f." This gave the "a" and "d" drugs a head start at reducing the bacteria count until x was brought into the model.

When fitting models where you don't control all the factors, you may find that the factors are interrelated, and the significance of one depends on what else is in the model.

The Prediction Equation

The prediction equation generated by JMP can be stored as a formula in its own column.

- Close the Fit Y by X window and return to the Fit Model results.
- Select **Save Columns > Prediction Formula** command from the red triangle popup menu on the uppermost title bar of the report.

This command creates a new column in the data table called Pred Formula Y. To see this formula (the prediction formula),

- Right-click in the column heading area and select **Formula**.

This opens a calculator window with the following formula for the prediction equation:

$$-2.695772906127 + \text{Match}(Drug, \begin{bmatrix} \text{"a"} \Rightarrow -1.1850365373806 \\ \text{"d"} \Rightarrow -1.0760652051714 \\ \text{"f"} \Rightarrow 2.261101742552 \\ \text{else} \Rightarrow . \end{bmatrix}) + 0.98718381112985 * x$$

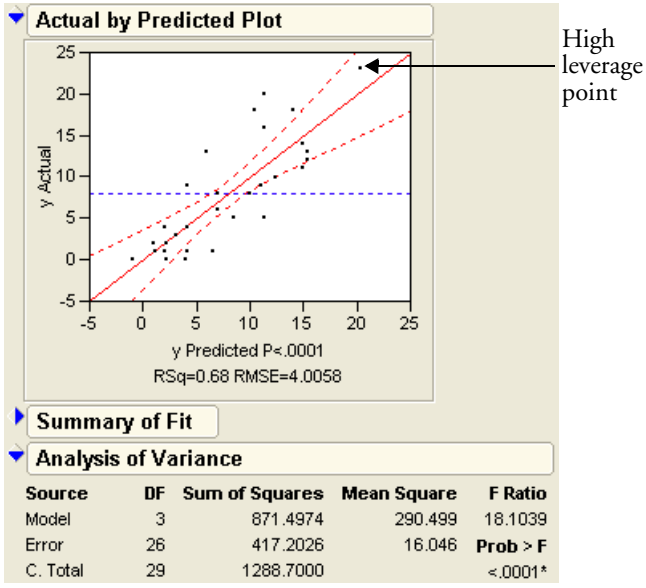
The Whole-Model Test and Leverage Plot

The whole-model test shows how the model fits as a whole compared with the null hypothesis of fitting only the mean. This is equivalent to testing the null hypothesis that all the parameters in the linear model are zero except for the intercept. This fit has three degrees of

freedom, 2 from Drug and 1 from the covariate x. The *F* of 18.1 is highly significant (see **Figure 14.6**).

The whole-model leverage plot (produced when **Effect Leverage** is selected rather than **Minimal Report** in the Fit Model dialog) is a plot of the actual value versus its predicted value. The residual is the distance from each point to the 45° line of fit (where the actual is equal to the predicted).

Figure 14.6 Whole-Model Test and Leverage Plot for Analysis of Covariance



The leverage plot in **Figure 14.6** shows the hypothesis test point by point. The points that are far out horizontally (like point 25) tend to contribute more to the test because the predicted values from the two models differ more there. Points like this are called *high-leverage points*.

Effect Tests and Leverage Plots

Now look at the effect leverage plots in **Figure 14.7** to examine the details for testing each effect in the model. Each effect test is computed from a difference in the residual sums of squares that compare the fitted model to the model without that effect.

For example, the sum of squares (SS) for x can be calculated by noting that the SS(error) for the full model is 417.2, but the SS(error) was 995.1 for the model that had only the Drug main effect (see **Figure 14.3**). The SS(error) for the model that includes the covariate is 417.2. The reduction in sum of squares is the difference, 995.1 – 417.2 = 577.9, as you can

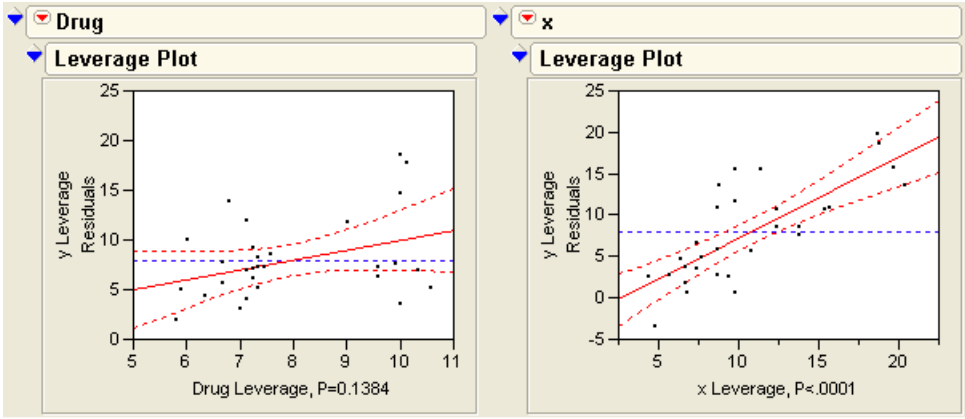
see in the Effect Test table (**Figure 14.7**). Similarly, if you remove Drug from the model, the SS(Error) grows from 417.2 to 485.8, a difference of 68.6 from the full model.

The leverage plot shows the composition of these sums of squares point by point. The Drug leverage plot in **Figure 14.7** shows the effect on the residuals that would result from removing Drug from the model. The distance from each point to the sloped line is its residual. The distance from each point to the horizontal line is what its residual would be if Drug were removed from the model. The difference in the sum of squares for these two sets of residuals is the sum of squares for the effect, which is the numerator of the *F*-test for the Drug effect.

You can evaluate leverage plots in a way that is similar to evaluating a plot for a simple regression that has a mean line and confidence curves on it. The effect is significant if the points are able to pull the sloped line significantly away from the horizontal line. The confidence curves are placed around the sloped line to show the 0.05-level test. The curves cross the horizontal line if the effect is significant (as on the right in **Figure 14.7**), but they encompass the horizontal line if the effect is not significant (as on the left).

Click on points to see which ones are high-leverage—away from the middle on the horizontal axis. Note whether they seem to support the test or not. If the points support the test, they are on the side trying to pull the line toward a higher slope.

Figure 14.7 Effect Tests and Effect Leverage Plots for Analysis of Covariance



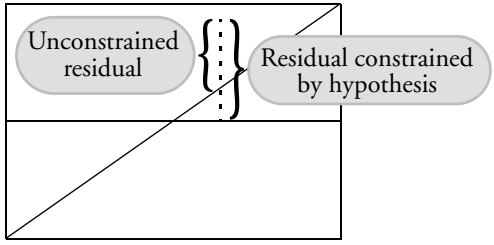
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Drug	2	2	68.55371	2.1361	0.1384
x	1	1	577.89740	36.0145	<.0001*

Figure 14.8 summarizes the elements of a leverage plot. A leverage plot for a specific hypothesis is any plot with the following properties:

- There is a sloped line representing the full model and a horizontal line representing a model constrained by an hypothesis.
- The distance from each point to the sloped line is the residual from the full model.
- The distance from each point to the horizontal line is the residual from the constrained model.

Figure 14.8 Schematic Defining a Leverage Plot



Least Squares Means

It might not be fair to make comparisons between raw cell means in data that you fit to a linear model. Raw cell means do not compensate for different covariate values and other factors in the model. Instead, construct predicted values that are the expected value of a typical observation from some level of a categorical factor when all the other factors have been set to neutral values. These predicted values are called *least squares means*. There are other terms used for this idea: *marginal means*, *adjusted means*, and *marginal predicted values*.

The role of these adjusted or least squares means is that they allow comparisons of levels with the control of other factors being held fixed.

In the drug example, the least squares means are the predicted values expected for each of the three values of Drug, given that the covariate x is held at some constant value. The constant value is chosen for convenience to be the mean of the covariate, which is 10.7333. The prediction equation gives the least squares means as follows:

fit equation:

$$-2.695 - 1.185\ \text{drug[a-f]} - 1.0760\ \text{drug[d-f]} + 0.98718\ \text{x}$$

for a:

$$-2.695 - 1.185\ (1) - 1.0760\ (0) + 0.98718\ (10.7333) = 6.71$$

for d:

$$-2.695 - 1.185\ (0) - 1.0760\ (1) + 0.98718\ (10.7333) = 6.82$$

for f:

$$-2.695 - 1.185\ (-1) - 1.0760\ (-1) + 0.98718\ (10.7333) = 10.16$$

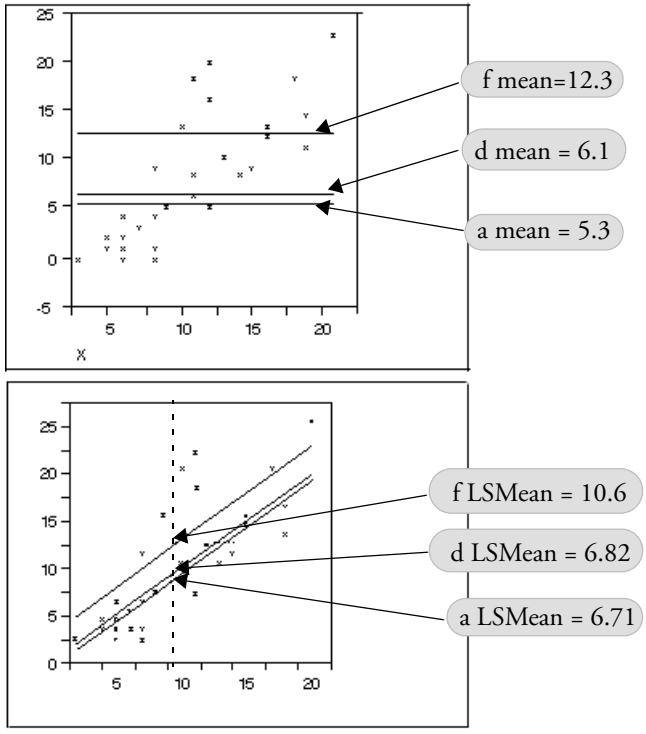
To verify these results, return to the Fit Model platform and open the Least Squares Means table for the Drug effect (shown to the right).

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
a	6.714963	1.2884943	5.3000
d	6.823935	1.2724690	6.1000
f	10.161102	1.3159234	12.3000

In the diagram shown at the top in **Figure 14.9**, the ordinary means are taken with different values of the covariate, so it is not fair to compare them. In the diagram at the bottom in **Figure 14.9**, the least squares means for this model are the intersections of the lines of fit for each level with the X value of 10.733. With this data, the least squares means are less separated than the raw means.

Figure 14.9 Diagram of Ordinary Means (top) and Least Squares Means (bottom)



Lack of Fit

The lack-of-fit test is the opposite of the whole-model test. Where the whole-model tests whether anything in the model is significant, the lack-of-fit tests whether anything left out of the model is significant. Unlike all other tests, it is usually desirable for the lack-of-fit test to be nonsignificant. That is, the null hypothesis is that the model does not need higher-order effects. A significant lack-of-fit test is advice to add more effects to the model using higher orders of terms already in the model.

But how is it possible to test effects that haven't been put in the model? All tests in linear models compare a model with a constrained or reduced version of that model. To test all the terms that could be in the model but are not—now *that* would be amazing!

Lack-of-fit compares the fitted model with a saturated model using the same terms. A *saturated* model is one that has a parameter for each combination of factor values that exists in the data. For example, a one-way analysis of variance is already saturated because it has a parameter for each level of the single factor. A complete factorial with all higher order interactions is completely saturated. For simple regression, saturation would be like having a separate coefficient to estimate for each value of the regressor.

If the lack-of-fit test is significant, there is some significant effect that has been left out of the model, and that effect is a function of the factors already in the model. It could be a higher-order power of a regressor variable, or some form of interaction among classification variables. If a model is already saturated, there is no lack-of-fit test possible.

The other requirement for a lack-of-fit test in continuous responses is that there be some exact replications of factor combinations in the data table. These exact duplicate rows (except for responses) allow the test to estimate the variation to use as a denominator in the lack-of-fit *F*-test. The error variance estimate from exact replicates is called *pure error* because it is independent of whether the model is right or wrong (assuming that it includes all the right factors).

In the drug model with covariate, the observations shown in **Table 14.2** form exact replications of data for Drug and x. The sum of squares around the mean in each replicate group reveals the contributions to pure error.

This pure error represents the best that can be done in fitting these terms to the model for this data. Whatever is done to the model involving Drug and x, these replicates and this error always exists. Pure error exists in the model regardless of the form of the model.

Table 14.2. Lack-of-Fit Analysis

Replicate Rows	Drug	x	y	Pure Error DF	Contribution to Pure Error
6 8	a a	6 6	4 1	1	$4.5=(4-2.5)^2+(1-2.5)^2$
1 9	a a	11 11	6 8	1	$2.0 = (6-7)^2+(6-8)^2$
11 12	d d	6 6	0 2	1	2.0
14 16 18	d d d	8 8 8	1 4 9	2	32.667
27 28 30	f f f	12 12 12	5 16 20	2	120.667
21 26	f f	16 16	13 12	1	0.5
Total				8	162.333

Pure error can reveal how complete the model is. If the error variance estimate from the model is much greater than the pure error, then adding higher order effects of terms already in the model improves the fit.

The Lack-of-Fit table for this example is shown to the right. The difference between the total error from the fitted model and pure error is called Lack-of-Fit error. It represents all the terms that might have been added to the model, but were not. The ratio of the lack-of-fit mean square to the pure error mean square is the *F*-test for lack-of-fit. For the covariate model, the lack-of-fit error is not significant, which is good because it is an indication that the model is adequate with respect to the terms included in the model.

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	18	254.86926	14.1594	0.6978
Pure Error	8	162.33333	20.2917	**Prob > F**
Total Error	26	417.20260		0.7507
				Max RSq
				0.8740

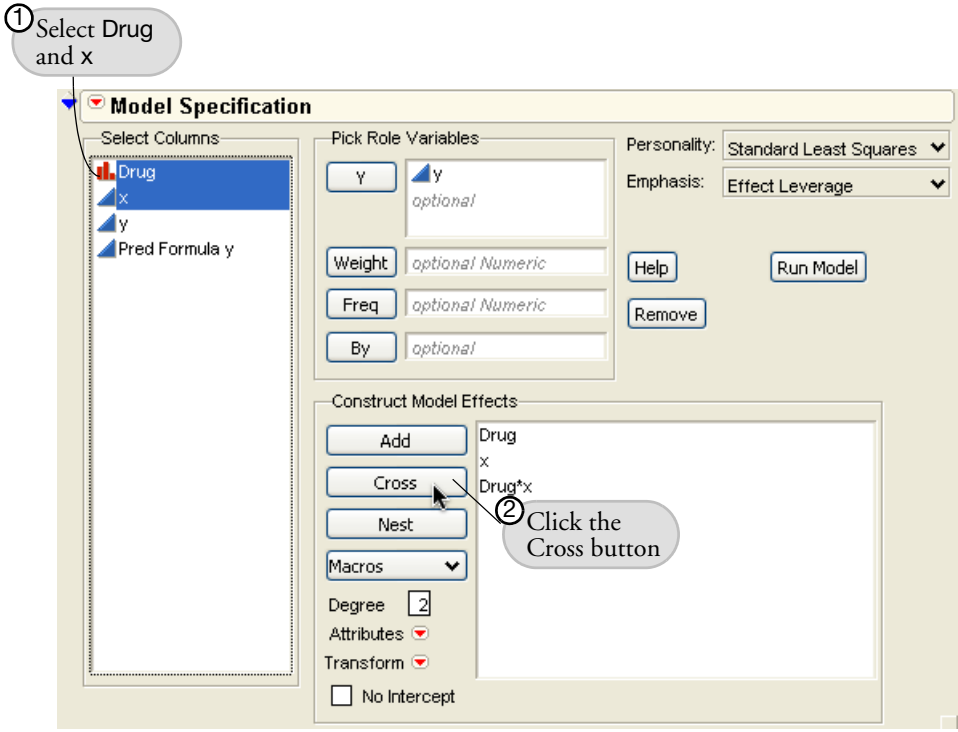
Separate Slopes: When the Covariate Interacts with the Classification Effect

When a covariate model includes a main effect and a covariate regressor, the analysis uses a separate intercept for the covariate regressor for each level of the main effect.

If the intercepts are different, could the slopes of the lines also be different? To find out, a method to capture the interaction of the regression slope with the main effect is needed. This is accomplished by introducing a crossed term, the interaction of Drug and x, into the model:

- Return to the Fit Model dialog, which already has Drug and x as effects in the model.
- Shift-click on Drug and x in the column selector list so that both columns are highlighted as shown in **Figure 14.10**.
- Click the **Cross** button. This gives an effect in the model called Drug*x.
- Click **Run Model** to see the results (**Figure 14.11**).

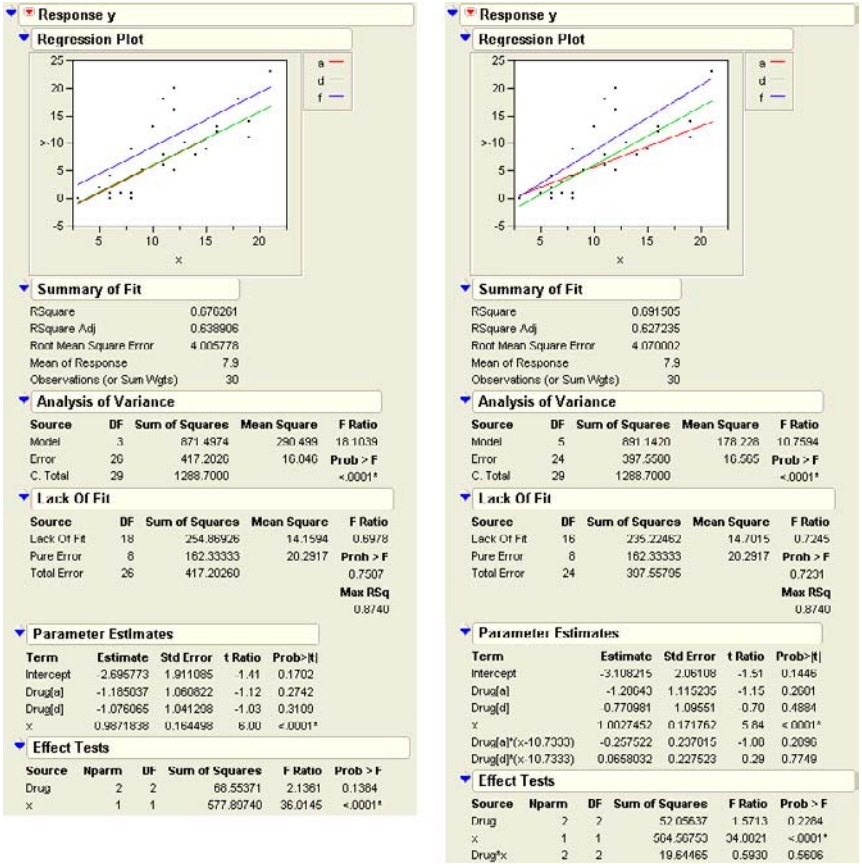
Figure 14.10 Fit Model Dialog for Analysis of Covariance with Separate Slopes



This specification adds two parameters to the linear model that allow the slopes for the covariate to be different for each Drug level. The new variable is the product of the dummy variables for Drug and the covariate values.

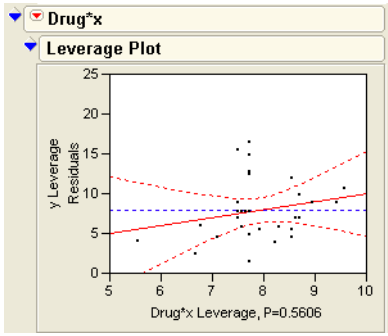
The Summary of Fit tables in **Figure 14.11** compare this separate slopes fit to the same slopes fit, showing an increase in R^2 from 67.63% to 69.15%.

Figure 14.11 Analysis of Covariance with Same Slopes (left) and Separate Slopes (right)



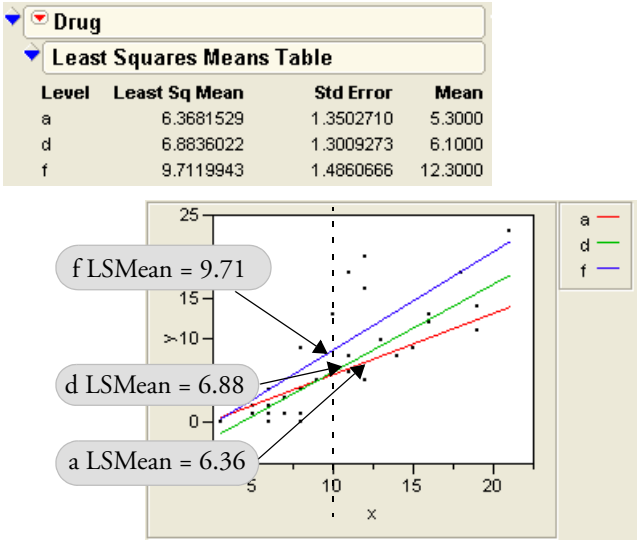
The separate slopes model shifts two degrees of freedom from the lack-of-fit error to the model, increasing the model degrees of freedom from 3 to 5. The pure error seen in both Lack-of-Fit tables is the same because there are no new variables in the separate slopes covariance model. The new effect in the separate slopes model is constructed from terms already in the original analysis of covariance model.

The Effect Test table in the report on the right in **Figure 14.11** shows that the test for the new term Drug*x for separate slopes is not significant; the *p*-value is 0.56 (shown below the plot on the right). The confidence curves on the leverage plot for the Effect Test do not cross the horizontal mean line, showing that the interaction term doesn't significantly contribute to the model. The least squares means for the separate slopes model have a more dubious value now.



Previously, with the same slopes on x as shown in **Figure 14.11**, the least squares means changed with whatever value of x was used, but the separation between them did not. Now, with separate slopes as shown in **Figure 14.12**, the separation of the least squares means is also a function of x. The least squares means are more or less significantly different depending on whatever value of x is used. JMP uses the overall mean, but this does not represent any magic standard base. Notice that a and d cross near the mean x, so their least squares means happen to be the same only because of where the x covariate was set.

Figure 14.12 Illustration of Covariance with Separate Slopes



Interaction effects always have the potential to cloud the main effect, as will be seen again with the two-way model in the next section.

Two-Way Analysis of Variance and Interactions

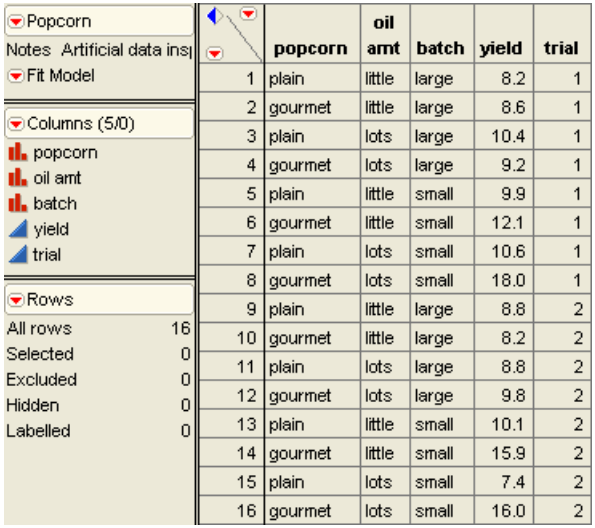
This section shows how to analyze a model in which there are two nominal or ordinal classification variables—a two-way model instead of a one-way model.

For example, a popcorn experiment was run, varying three factors and measuring the popped volume yield per volume of kernels. The goal was to see which factors gave the greatest volume of popped corn. **Figure 14.13** shows a listing of the popcorn data.

- To see the data, open the sample table Popcorn.jmp.

It has variables popcorn with values “plain” and “gourmet”; batch, which designates whether the popcorn was popped in a large or small batch; and oil amt with values “lots” or “little.” Start with two of the three factors, popcorn and batch.

Figure 14.13 Listing of the Popcorn Data Table

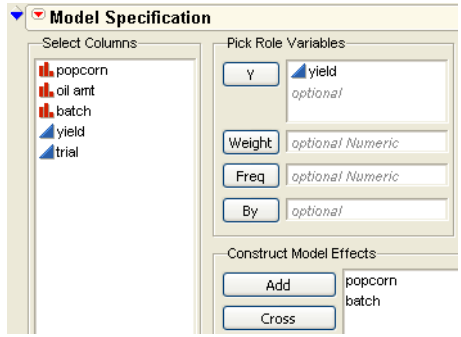


	popcorn	oil amt	batch	yield	trial
1	plain	little	large	8.2	1
2	gourmet	little	large	8.6	1
3	plain	lots	large	10.4	1
4	gourmet	lots	large	9.2	1
5	plain	little	small	9.9	1
6	gourmet	little	small	12.1	1
7	plain	lots	small	10.6	1
8	gourmet	lots	small	18.0	1
9	plain	little	large	8.8	2
10	gourmet	little	large	8.2	2
11	plain	lots	large	8.8	2
12	gourmet	lots	large	9.8	2
13	plain	little	small	10.1	2
14	gourmet	little	small	15.9	2
15	plain	lots	small	7.4	2
16	gourmet	lots	small	16.0	2

- Choose **Analyze > Fit Model**.

- When the Fit Model dialog appears, select yield as the Y (response) variable. Select popcorn and batch and click **Add** to use them as the Model Effects.

The Fit Model dialog should look like the one shown to the right.



- Click **Run Model** to see the analysis.

Figure 14.14 shows the analysis tables for the two-factor analysis of variance:

- The model explains 56% of the variation in yield (the R^2).
- The remaining variation has a standard error of 2.248 (Root Mean Square Error).
- The significant lack-of-fit test (p-value of 0.0019) says that there is something in the two factors that is not being captured by the model. The factors are affecting the response in a more complex way than is shown by main effects alone. The model needs an interaction term.
- Each of the two effects has two levels, so they each have a single parameter. Thus the t-test results are identical to the F-test results. Both factors are significant.

Figure 14.14 Two-Factor Analysis of Variance for Popcorn Experiment

Response yield

Summary of Fit

RSquare	0.560527
RSquare Adj	0.492916
Root Mean Square Error	2.248247
Mean of Response	10.75
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	83.81000	41.9050	8.2904
Error	13	65.71000	5.0546	Prob > F
C. Total	15	149.52000		0.0048*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	1	37.210000	37.2100	15.6674
Pure Error	12	28.500000	2.3750	Prob > F
Total Error	13	65.710000		0.0019*
				Max RSq
				0.8094

Parameter Estimates

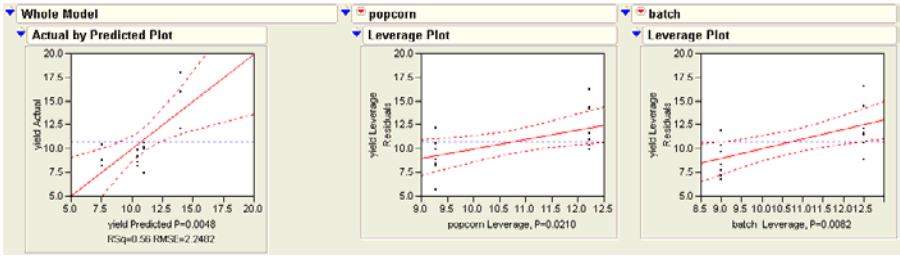
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.75	0.562062	19.13	<.0001*
popcorn[gourmet]	1.475	0.562062	2.62	0.0210*
batch [large]	-1.75	0.562062	-3.11	0.0082*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
popcorn	1	1	34.810000	6.8868	0.0210*
batch	1	1	49.000000	9.6941	0.0082*

The leverage plots in **Figure 14.15** show the point-by-point detail for the fit as a whole and the fit as it is carried by each factor partially. Because this is a balanced design, all the points have the same leverage. Therefore. they are spaced out horizontally the same in the leverage plot for each effect.

Figure 14.15 Leverage Plots for Two-Factor Popcorn Experiment with Interaction



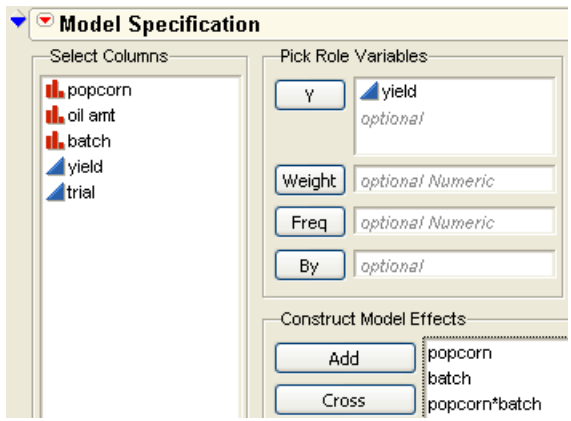
The lack-of-fit test shown in **Figure 14.14** is significant. We reject the null hypothesis that the model is adequate, and decide to add an interaction term, also called a *crossed effect*. An interaction means that the response is not simply the sum of a separate function for each term. In addition, each term affects the response differently depending on the level of the other term in the model.

The popcorn by batch interaction is added to the model as follows:

- Return to the Fit Model dialog, which already has the popcorn and batch terms in the model.

We want to select both popcorn and batch in the **Select Columns** list. To do so,

- Click on popcorn and Control-click (⌘-click on the Macintosh) on batch to extend the selection.
- Click the **Cross** button to see the popcorn*batch interaction effect in the Fit Model dialog as shown to the right.
- Click **Run Model** to see the tables in **Figure 14.16**.



Including the interaction term increased the R^2 from 56% to 81%. The standard error of the residual (Root Mean Square Error) has gone down from 2.2 to 1.54.

Figure 14.16 Statistical Analysis of Two-Factor Experiment with Interaction

Response yield

Summary of Fit

RSquare	0.80939
RSquare Adj	0.761738
Root Mean Square Error	1.541104
Mean of Response	10.75
Observations (or Sum Wgts)	16

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.75	0.385276	27.90	<.0001*
popcorn[gourmet]	1.475	0.385276	3.83	0.0024*
batch [large]	-1.75	0.385276	-4.54	0.0007*
popcorn[gourmet]*batch [large]	-1.525	0.385276	-3.96	0.0019*

Effect Tests

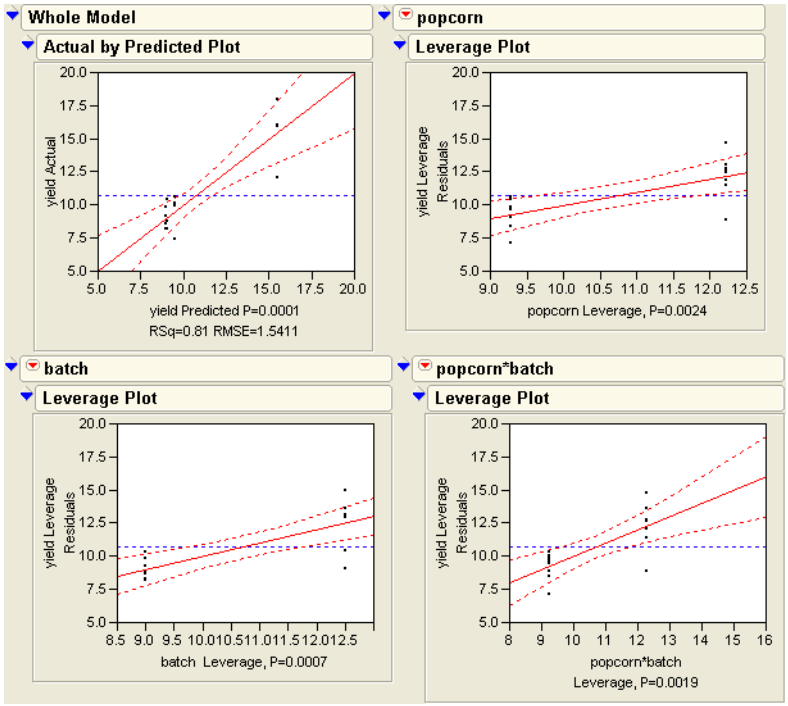
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
popcorn	1	1	34.810000	14.6568	0.0024*
batch	1	1	49.000000	20.6316	0.0007*
popcorn*batch	1	1	37.210000	15.6674	0.0019*

The Effect Test table shows that all effects are significant. The popcorn*batch effect has a *p*-value of 0.0019, highly significant. The number of parameters (and degrees of freedom) of an interaction are the product of the number of parameters of each term in the interaction. The popcorn*batch interaction has one parameter (and one degree of freedom) because the popcorn and batch terms each have only one parameter.

An interesting phenomenon, which is true only in balanced designs, is that the parameter estimates and sums of squares for the main effects are the same as in the previous fit without interaction. The *F*-tests are different only because the error variance (Mean Square Error) is smaller in the interaction model. The interaction effect test is identical to the lack-of-fit test in the previous model.

Again, the leverage plots (**Figure 14.17**) show the tests in point-by-point detail. The confidence curves clearly cross the horizontal. The effects tests (shown in **Figure 14.6**) confirm that the model and all effects are highly significant.

Figure 14.17 Leverage Plots for Two-Factor Experiment with Interaction



You can see some details of the means in the least squares means table, but in a balanced design (equal numbers in each level and no covariate regressors), they are equal to the raw cell means.

To see profile plots for each effect, select the **LSMeans Plot** command in the popup menu at the top of each leverage plot.

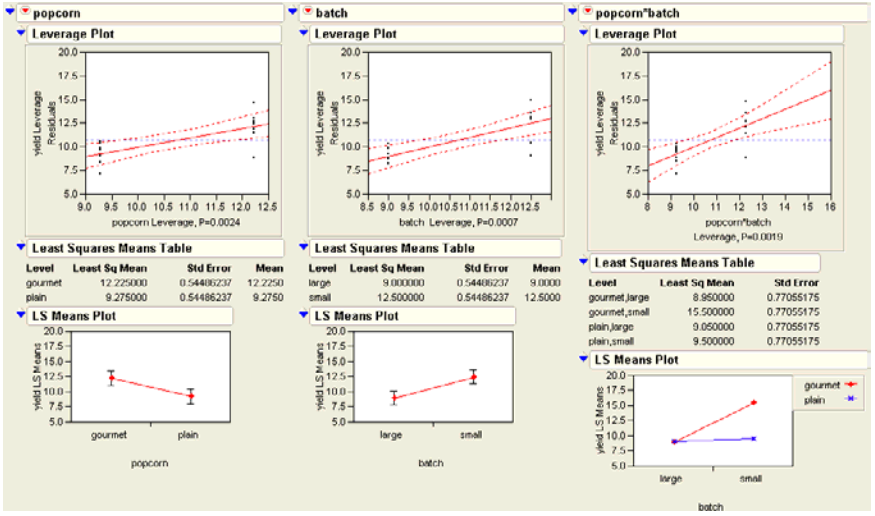
The result is a series of profile plots below each effect's report. Profile plots are a graphical form of the values in the Least Squares Means table.

- The leftmost plot in **Figure 14.18** is the profile plot for the popcorn main effect. The "gourmet" popcorn seems to have a higher yield.
- The middle plot is the profile plot for the batch main effect. It looks like small batches have higher yields.
- The rightmost plot is the profile plot for the popcorn by batch interaction effect.

Looking at the effects together in an interaction plot shows that the popcorn type matters for small batches but not for big ones. Said another way, the batch size matters for gourmet

popcorn, but not for plain popcorn. In an interaction profile plot, one interaction term is on the *x*-axis and the other term forms the different lines.

Figure 14.18 Interaction Plots and Least Squares Means



Optional Topic: Random Effects and Nested Effects

This section talks about nested effects, repeated measures, and random effects mixed models. This is a large collection of topics to cover in a few pages, so hopefully this overview will be an inspiration to look to other textbooks and study these topics more completely.

As an example, consider the following situation. Six animals from two species were tracked, and the diameter of the area that each animal wandered was recorded. Each animal was measured four times, once per season.

Figure 14.19 shows a listing of the Animals data.

Figure 14.19 Listing of the Animals Data Table

Animals
Notes Example Data-
Repeated Measures

Columns (4/0)
species
subject
miles
season

Rows
All rows 24
Selected 0
Excluded 0
Hidden 0
Labelled 0

	species	subject	miles	season
1	FOX	1	0	fall
2	FOX	1	0	winter
3	FOX	1	5	spring
4	FOX	1	3	summer
5	FOX	2	3	fall
6	FOX	2	1	winter
7	FOX	2	5	spring
8	FOX	2	4	summer
9	FOX	3	4	fall
10	FOX	3	3	winter
11	FOX	3	6	spring
12	FOX	3	2	summer
13	COYOTE	1	4	fall
14	COYOTE	1	2	winter
15	COYOTE	1	7	spring
16	COYOTE	1	8	summer
17	COYOTE	2	5	fall
18	COYOTE	2	4	winter
19	COYOTE	2	6	spring
20	COYOTE	2	6	summer
21	COYOTE	3	7	fall
22	COYOTE	3	5	winter
23	COYOTE	3	8	spring
24	COYOTE	3	9	summer

Nesting

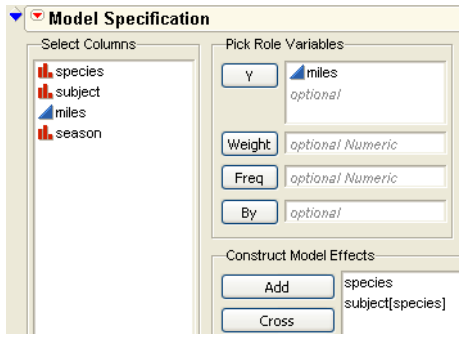
One feature of the data is that the labeling for each subject animal is nested within species. The observations for subject 1 for species Fox are not for the same animal as subject 1 for Coyote. The way to express this in a model is to always write the subject effect as subject[species], which is read as “subject nested within species” or “subject within species.” The rule about nesting is that whenever you refer to a subject with a given level of factor, if that implies what another factor's level is, then the factor should only appear in nested form.

When the linear model machinery in JMP sees a nested effect such as “B within A”, denoted B[A], it computes a new set of A parameters for each level of B. The Fit Model dialog allows for nested effects to be specified as in the following example.

- Open the Animals.jmp table.
- Choose **Analyze > Fit Model**.
- Use miles as the Y variable.
- Add species to the Model Effects list.
- Add subject to the Model Effects list.

- Select species in the Select Columns list and select subject in the Effects in Model list.
- Click the **Nest** button.

This adds the nested effect subject[species] shown to the right.



- Add season to the Model Effects and click **Run Model** to see the results in **Figure 14.20**.

This model runs fine, but it has something wrong with it. The *F*-tests for all the effects in the model use the residual error in the denominator. The reason that this is an error, and the solution to this problem, are presented below.

Figure 14.20 Results for Animal Data Analysis

Response miles

Summary of Fit

RSquare	0.494413
RSquare Adj	0.353972
Root Mean Square Error	1.968502
Mean of Response	4.458333
Observations (or Sum Wgts)	24

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	68.20833	13.6417	3.5204
Error	18	69.75000	3.8750	Prob > F
C. Total	23	137.95833		0.0215*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.4583333	0.401819	11.10	<.0001*
species[COYOTE]	1.4583333	0.401819	3.63	0.0019*
species[COYOTE]:subject[1]	-0.666667	0.803638	-0.83	0.4177
species[COYOTE]:subject[2]	-0.666667	0.803638	-0.83	0.4177
species[FOX]:subject[1]	-1	0.803638	-1.24	0.2293
species[FOX]:subject[2]	0.25	0.803638	0.31	0.7593

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
species	1	1	51.041667	13.1720	0.0019*
subject[species]	4	4	17.166667	1.1075	0.3834

Note in **Figure 14.20** the treatment of nested effects in the model. There is one parameter for the two levels of species ("Fox" and "Coyote"). Subject is nested in species, so there is a separate set of two parameters (for three levels of subject) for subject within each level of species, giving a total of four parameters for subject. Season, with four levels, has three parameters. The total parameters for the model (not including the intercept) is 1 for species + 4 for subject + 3 for season = 8.

Use the **Save Prediction Formula** command and look at the Formula Editor window for the saved formula to see the following prediction equation using the parameter estimates:

$$
\begin{aligned}
&4.45833333333333 \\
&+\text{Match}\left(\textit{species}\right)\begin{Bmatrix}\text{"COYOTE"} \Rightarrow 1.45833333333333\\ \text{"FOX"} \Rightarrow -1.4583333333333\\ \text{else} \Rightarrow .\end{Bmatrix} \\
&+\text{Match}\left(\textit{species}\right)\begin{Bmatrix}\text{"COYOTE"} \Rightarrow \text{Match}\left(\textit{subject}\right)\begin{Bmatrix}1 \Rightarrow -0.6666666666667\\ 2 \Rightarrow -0.6666666666667\\ 3 \Rightarrow 1.33333333333333\\ \text{else} \Rightarrow .\end{Bmatrix}\\ \text{"FOX"} \Rightarrow \text{Match}\left(\textit{subject}\right)\begin{Bmatrix}1 \Rightarrow -1\\ 2 \Rightarrow 0.25\\ 3 \Rightarrow 0.75\\ \text{else} \Rightarrow .\end{Bmatrix}\\ \text{else} \Rightarrow .\end{Bmatrix}
\end{aligned}
$$

Repeated Measures

As mentioned above, the previous analysis has a problem—the *F*-test used to test the species effect is constructed using the model residual in the denominator, which isn't appropriate for this situation. The following sections explain this problem and outline solutions.

There are three ways to understand this problem, which correspond to three different (but equivalent) resolutions:

- The effects can be declared as random, causing JMP to synthesize special *F*-tests.
- The observations can be made to correspond to the experimental unit.
- The analysis can be viewed as a multivariate problem.

The key in each method is to focus on only one of the effects in the model. In the animals example, the effect is species—how does the wandering radius differ between "Fox" and "Coyote"? The species effect is incorrectly tested in the previous example so species is the effect that needs attention.

Method 1: Random Effects-Mixed Model

The subject effect is what is called a *random effect*. The animals were selected randomly from a large population, and the variability from animal to animal is from some unknown distribution. To generalize to the whole population, study the species effect with respect to the variability.

It turns out that if the design is balanced, it is possible to use an appropriate random term in the model as an error term instead of using the residual error to get an appropriate test. In this case subject[species], the nested effect acts as an error term for the species main effect.

To construct the appropriate *F*-test, do a hand calculation using the results from Effect Test table shown previously in **Figure 14.20**. Divide the mean square for species by the mean square for subject[species] as shown by the following formula.

$$F = \frac{51.041667}{\frac{17.166667}{4}}$$

This *F*-test has 1 numerator degree of freedom and 4 denominator degrees of freedom, and evaluates to 11.89.

Random effects in general statistics texts, often described in connection with *split plots* or *repeated measures designs*, describe which mean squares need to be used to test each model effect.

Now, let's have JMP do this calculation instead of doing it by hand. JMP will give the correct tests even if the design is not balanced.

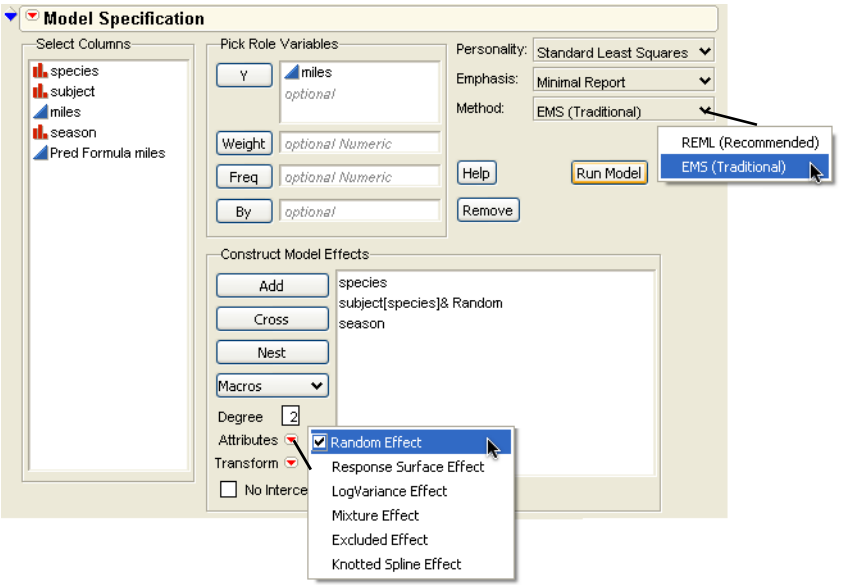
First, specify subject[species] as a random effect.

- Click the Fit Model dialog window for the Animals data to make it active.
- Click to highlight subject[species] showing in the Model Effects list.
- Select the **Random Effect** attribute found in the **Attributes** popup menu.

The subject[species] effect then appears with &Random appended to it as shown in **Figure 14.21**.

- Select **EMS (Traditional)** from the Method popup menu as shown.
- Click **Run Model** to see the results.

Figure 14.21 Fit Model Dialog Using a Random Effect



JMP constructs tests for random effects using the following steps:

1. First, the expected mean squares are found. These are coefficients that relate the mean square to the variances for the random effects.

Expected Mean Squares

The Mean Square per row by the Variance Component per column

EMS

	Intercept	species	subject[species]&Random	season
Intercept	0	0	0	0
species	0	12	4	0
subject[species]&Random	0	0	4	0
season	0	0	0	6

plus 1.0 times Residual Error Variance

2. Next, the variance component estimates are found using the mean square values for the random effects and their coefficients. It is possible (but rare) for a variance component estimate to be negative.

Variance Component Estimates

Component	Var Comp Est	Percent of Total
subject[species]&Random	0.701389	32.063
Residual	1.486111	67.937
Total	2.1875	100.000

These estimates based on equating Mean Squares to Expected Value.

3. For each effect in the model, JMP then determines what linear combination of other mean squares would make an appropriate denominator for an *F*-test. This denominator is the linear combination of mean squares that has the same expectation as the mean square of the effect (numerator) under the null hypothesis.

Test Denominator Synthesis

Source	MS Den	DF Den	Denom MS Synthesis
species	4.29167	4	subject[species]&Random
subject[species]&Random	3.875	18	Residual

4. *F*-tests are now constructed using the denominators synthesized from other mean squares. If an effect is prominent, then it will have a much larger mean square than expected under the null hypothesis for that effect.

Tests wrt Random Effects

Source	SS	MS Num	DF Num	F Ratio	Prob > F
species	51.0417	51.0417	1	11.8932	0.0261*
subject[species]&Random	17.1667	4.29167	4	2.8879	0.0588
season	47.4583	15.8194	3	10.6449	0.0005*

Again, the *F*-statistic for species is 11.89 with a *p*-value of 0.026. The tests for the other factors use the residual error mean square, which are the same as the tests done in the first model.

What about the test for season? Because the experimental unit for season corresponds to each row of the original table, the residual error in the first model is appropriate. The *F* of 10.64 with a *p*-value 0.0005 means that the miles (the response) does vary across season. If an interaction between species and season is included in the model, it is also be correctly tested using the residual mean square.

Note 1: There is an alternate way to define the random effects that produces slightly different expected mean squares, variance component estimates, and *F*-tests. The argument over which method of parameterization is more informative has been ongoing for 40 years.

Note 2: With random effects, it is not only the *F*-tests that need to be refigured. Standard deviations and contrasts on least squares means may need to be adjusted, depending on details of the situation. Consult an expert if you need to delve into the details of an analysis.

Note 3: Another method of estimating models with random effects, called REML, is available. Though it is the recommended method, most textbooks describe the EMS approach, which is why it is described here. Eventually, textbooks will shift to REML.

Method 2: Reduction to the Experimental Unit

There are only 6 animals, but are there 24 rows in the data table because each animal is measured 4 times. However, taking 4 measurements on each animal doesn't make it legal to count each measurement as an observation. Measuring each animal millions of times and throwing all the data into a computer would yield an extremely powerful—and incorrect—test.

The experimental unit that is relevant to species is the individual animal, not each measurement of that animal. When testing effects that only vary subject to subject, the experimental unit should be a single response per subject, instead of the repeated measurements.

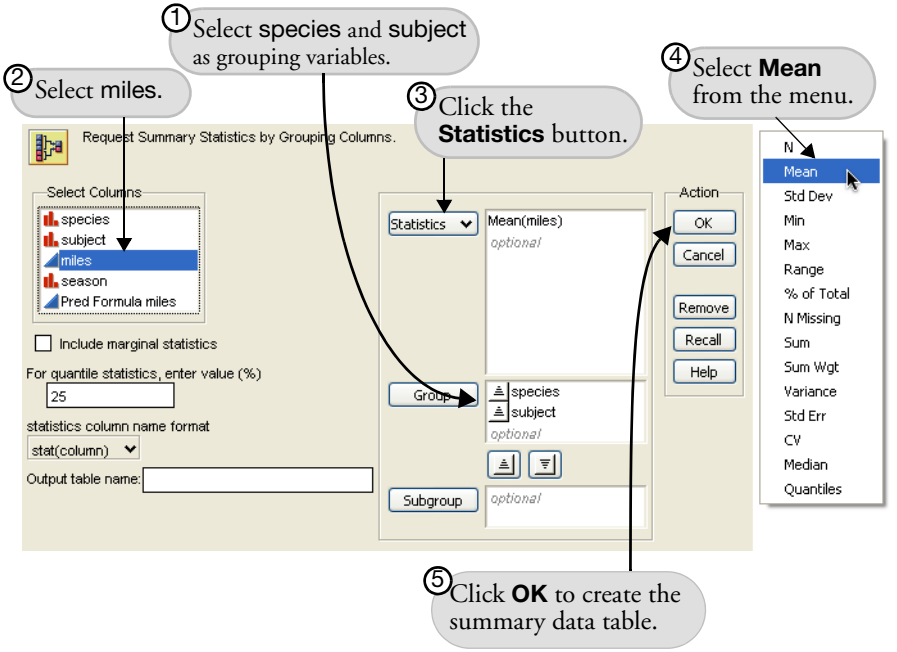
One way to handle this situation is to group the data to find the means of the repeated measures, and analyze these means instead of the individual values.

- With the Animals table active, choose **Tables > Summary**, which displays the dialog shown in **Figure 14.22**.
- Pick both species and subject as grouping variables.
- Highlight the miles variable as shown and select **Mean** from the **Stats** popup menu on the Summary dialog to see Mean(Miles) in the dialog.

This notation indicates you want to see the mean (average) miles for each subject within each species.

- Click **OK** to see the summary table shown at the bottom of **Figure 14.22**.

Figure 14.22 Summary Dialog and Summary Table for Animals Data



Animals By (speci
Source

Columns (4/0)
species
subject
N Rows
Mean(miles)

	species	subject	N Rows	Mean(miles)
1	COYOTE	1	4	5.25
2	COYOTE	2	4	5.25
3	COYOTE	3	4	7.25
4	FOX	1	4	2
5	FOX	2	4	3.25
6	FOX	3	4	3.75

Now fit a model to the summarized data.

- With the Animals.jmp by species subject table (the Summary table) active, choose **Analyze > Fit Model**.
- Use Mean(miles) as Y, and species as the Model Effects variable.
- Click **Run Model** to see the proper *F*-test of 11.89 for species, with a *p*-value of 0.0261.

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
species	1	1	12.760417	11.8932	0.0261*

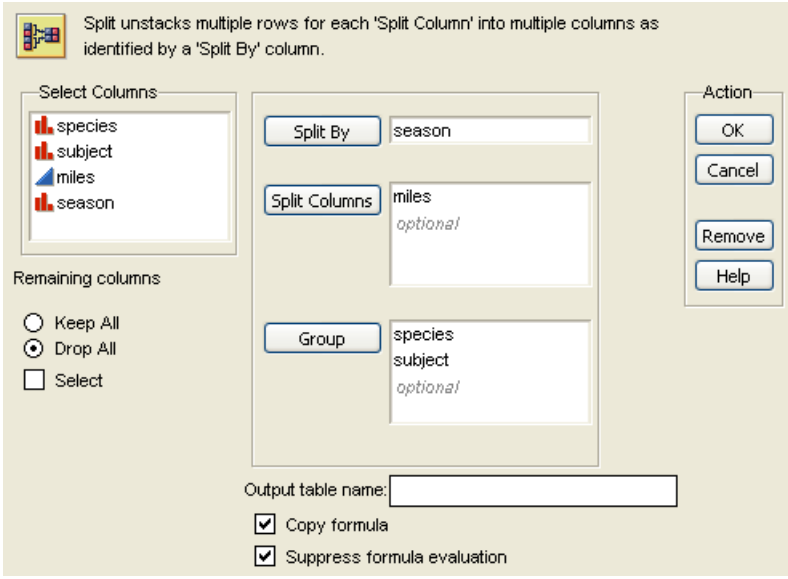
Note that this is the same result as the calculation shown in the previous section.

Method 3: Correlated Measurements-Multivariate Model

In the animal example there were multiple (four) measurements for the same animal. These measurements are likely to be correlated in the same sense as two measurements are correlated in a paired *t*-test. This situation of multiple measurements of the same subject is called *repeated measures*, or a *longitudinal* situation. This kind of experimental situation can be looked at as a multivariate problem.

To use a multivariate approach, the data table must be rearranged so that there is only one row for each individual animal, with the four measurements on each animal in four columns:

- To rearrange the Animals table, choose **Tables > Split**, which splits columns.
- Complete the Split Columns dialog by assigning miles as the **Split Columns** variable, season as the **Split By** variable (its values become the new column names), and species and subject as **Group** variables.



- Click **OK** to see a new untitled table like the one shown in **Figure 14.23**.

Figure 14.23 Rearrangement of the Animals Data

Untitled 3
Source

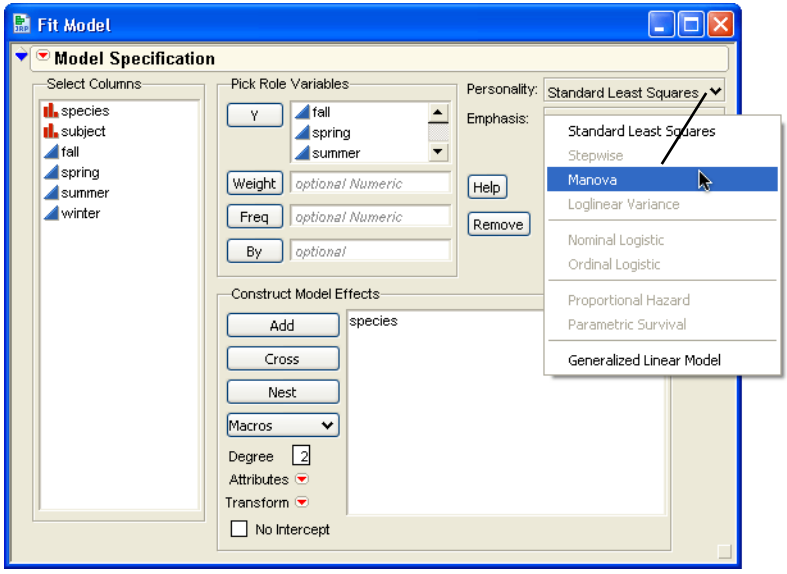
Columns (6/0)
species
subject
fall
spring
summer
winter

	species	subject	fall	spring	summer	winter
1	COYOTE	1	4	7	8	2
2	COYOTE	2	5	6	6	4
3	COYOTE	3	7	8	9	5
4	FOX	1	0	5	3	0
5	FOX	2	3	5	4	1
6	FOX	3	4	6	2	3

Then, fit a multivariate model with four Y variables and a single response:

- Choose **Analyze > Fit Model** and select fall, spring, summer, and winter as Y variables.
- Select species as the Model Effect.
- Select **Manova** from the fitting personality popup menu as shown in **Figure 14.24**.

Figure 14.24 Model for Manova

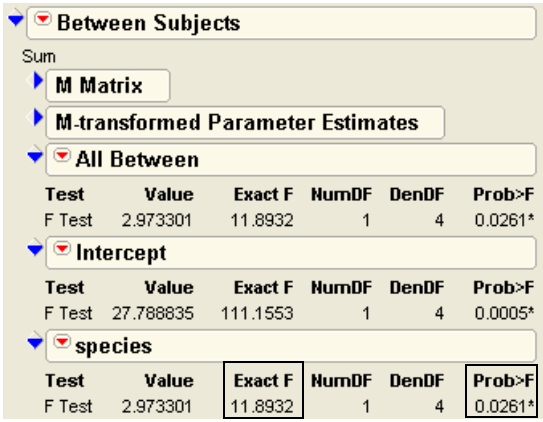


- Click **Run Model** to see the analysis results.

- When the Multivariate control panel appears, choose **Repeated Measures** from the Response popup menu.
- When the repeated measures appears, accept the default name Time and click **OK**.

The following report is the result.

Figure 14.25 Multivariate Analysis of Repeated Measures Data



The resulting fit includes the Between Subjects report for species shown in **Figure 14.25**. The report for the species effect shows the same *F*-test of 11.89 with a *p*-value of 0.0261, just as with the other methods.

Varieties of Analysis

In the previous cases, all the tests resulted in the same *F*-test for species. However, it is not generally true that different methods produce the same answer. For example, if species had more than two levels, the four multivariate tests (Wilk's lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's maximum root) each produce a different test result, and none of them would agree with the mixed model approach discussed previously.

Two more tests involving adjustments to the univariate method can be obtained from the multivariate fitting platform. With an unequal number of measurements per subject, the multivariate approach cannot be used. With unbalanced data, the mixed model approach also plunges into a diversity of methods that offer different answers from the Method of Moments that JMP uses.

Summary

When using the residual error to form *F*-statistics, ask if the row in the table corresponds to the unit of experimentation. Are you measuring variation in the way appropriate for the effect you are examining?

- When the situation does not live up to the framework of the statistical method, the analysis will be incorrect, as was method 1 (treating data table observations as experimental units) in the example above for the species test.
- Statistics offers a diversity of methods (and a diversity of results) for the same question. In this example, the different results are not wrong, they are just different.
- Statistics is not always simple. There are many ways to go astray. Educated common sense goes a long way, but there is no substitute for expert advice when the situation warrants it.

Exercises

1. Denim manufacturers are concerned with maximizing the comfort of the jeans they make. In the manufacturing process, starch is often built up in the fabric, creating a stiff jean that must be "broken in" by the wearer before they are comfortable. To minimize the break-in time for each pair of pants, the manufacturers often wash the jeans to remove as much starch as possible. This example concerns three methods of this washing. The data table Denim.jmp (Creighton, 2000) contains four columns: Method (describing the method used in washing), Size of Load (in pounds), Sand Blasted (recording whether or not the fabric was sand blasted prior to washing), Thread Wear (a measure of the destruction of threads during the process) and Starch Content (the starch content of the fabric after washing).

   (a) The three methods of washing consist of using an enzyme (alpha amalyze) to destroy the starch, chemically destroying the starch (caustic soda), and washing with an abrasive compound (pumice stone, hence the term "stone-washed jeans"). Determine if there is a significant difference in starch content due to washing method in two ways: using the Fit Y by X platform and the Fit Model platform. Compare the Summary of Fit and Analysis of Variance displays for both analyses.

   (b) Produce an LSMeans Plot for the Method factor in the above analysis. What information does this tell you? Compare the results with the means diamonds produced in the Fit Y By X plot.

(c) Fit a model of Starch Content using all three factors Size of Load, Sand Blasted, and Method. Which factors are significant?

(d) Produce LSMeans Plots for Method and Sand Blasted and interpret them.

(e) Use LSMeans Contrasts to determine if alpha amalyze is significantly different from caustic soda in its effect on Starch Content.

(f) Examine all first-level interactions of the three factors by running Fit Model and requesting **Factorial to Degree** from the Macros menu with the three factor columns selected. Which interactions are significant?

2. The file Titanic.jmp contains information on the passengers of the RMS Titanic. The four variables represent the class (first, second, third, and crew), age, sex, and survival status (yes or no) for each passenger. Use JMP to answer the following questions:

(a) Using the Fit Model platform, find a nominal logistic model to predict survival based on class, age, and sex.

(b) To evaluate the effectiveness of your model, save the probability formula from the model to the data table. Then, using Fit Y By X, prepare a contingency table to discover the percentage of times the model made the correct prediction.

(c) How do you know that the predictions in (b) are the most accurate for the given data?

3. The file Decathlon.jmp (Perkiömäki, 1995) contains decathlon scores from several winners in various competitions. Although there are ten events, they fall into groups of running, jumping, and throwing.

(a) Suppose you were going to predict scores on the 100m running event. Which of the other nine events do you think would be the best indicators of the 100m results?

(b) Run a Stepwise regression (from the Fit Model Platform) with 100m as the Y and the other nine events as X. Use the default value of 0.25 probability to enter and the Forward method. Which events are included in the resulting model?

(c) Complete a similar analysis on Pole Vault. Examine the events included in the model. Are you surprised?

15 Bivariate and Multivariate Relationships

Overview

This chapter explores the relationship between two variables, their correlation, and the relationships between more than two variables. You look for patterns and you look for points that don't fit the patterns. You see where the data points are located, where the distribution is dense, and which way it is oriented.

Detective skills are built with the experience of looking at a variety of data, and learning to look at them in a variety of different ways. As you become a better detective, you also develop better intuition for understanding more advanced techniques.

It is not easy to look at lots of variables, but the increased range of the exploration helps you make more interesting and valuable discoveries.

Bivariate Distributions

Previous chapters covered how the distribution of a response can vary depending on factors and groupings. This chapter returns to distributions as a simple unstructured batch of data. However, instead of a single variable, the focus is on the joint distribution of two or more responses.

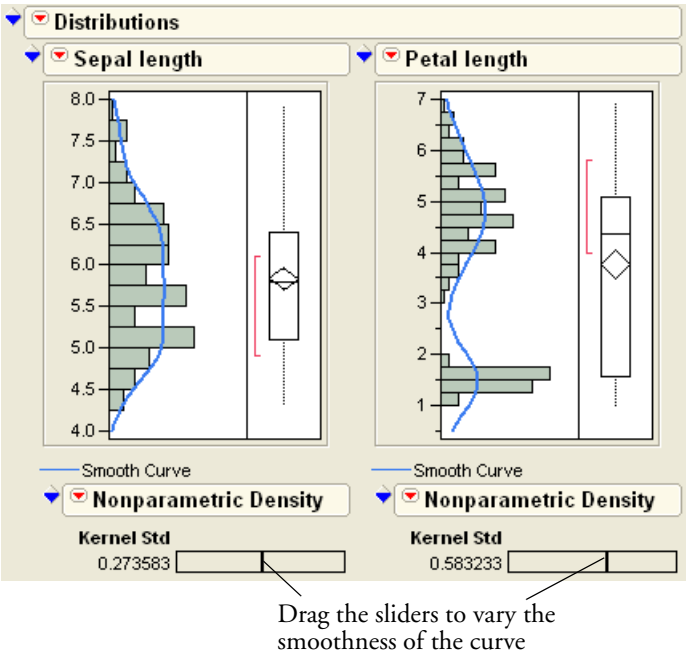
Density Estimation

As with univariate distributions, a central question is where are the data? What regions of the space are dense with data, and what areas are relatively vacant? The histogram forms a simple estimate of the density of a univariate distribution. If you want a smoother estimate of the density, JMP has an option that takes a weighted count of a sliding neighborhood of points to produce a smooth curve. This idea can be extended to several variables.

One of the most classic multivariate data sets in statistics contains the measurements of iris flowers that R. A. Fisher analyzed. Fisher's iris data are in the data table called Iris.jmp, with variables Sepal length, Sepal width, Petal length, and Petal width. First, look at the variables one at a time.

- Open Iris.jmp and choose **Analyze > Distribution** on Sepal Length and Petal Length.
- When the report appears, select **Fit Distribution > Smooth Curve** from the red triangle menu on the report title bar.
- When the smooth curve appears, drag the density slider beneath the histogram to see the effect of using a wider or narrower smoothing distribution (**Figure 15.1**).

Figure 15.1 Univariate Distribution with Smoothing Curve



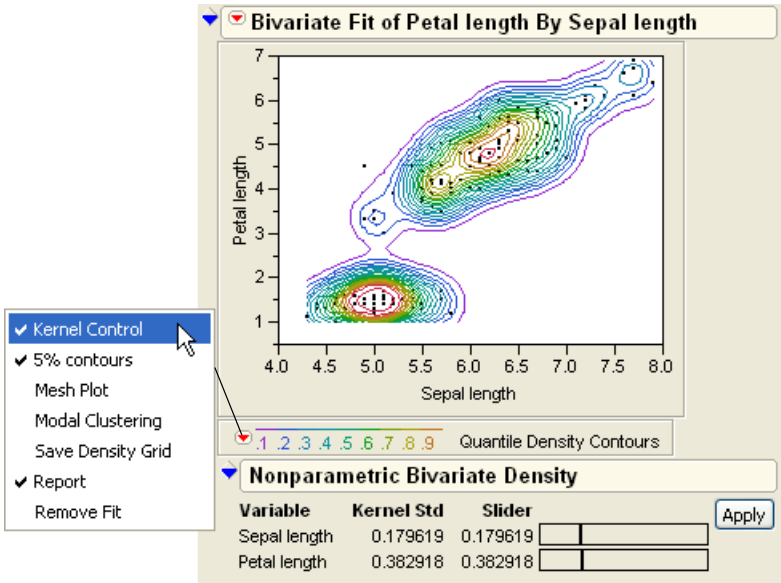
Notice in **Figure 15.1** that Petal length has an unusual distribution with two modes and a vacant area in the middle of the range. There are no petals with a length in the range from 2 to 3.

Bivariate Density Estimation

JMP has an implementation of a smoother that works with two variables to show their bivariate densities. The goal is to draw lines around areas that are dense with points. Continue with the iris data and look at Petal length and Sepal length together:

- Choose **Analyze > Fit Y by X** with Petal length as Y and Sepal Length as X.
- When the scatterplot appears, select **Nonpar Density** from the menu on the title of the plot.

Figure 15.2 Bivariate Density Estimation Curves



The result (**Figure 15.2**) is a contour graph, where the various contour lines show paths of equal density. The density is estimated for each point on a grid by taking a weighted average of the points in the neighborhood, where the weights decline with distance. Estimates done in this way are called *kernel smoothers*.

The Nonparametric Bivariate Density table beneath the plot has slider controls available for control of the vertical and horizontal width of the smoothing distribution.

- Select **Kernel Control** from the red triangle popup menu beside the legend of the contour graph.

Because it can take a while to calculate densities, they are not re-estimated until the **Apply** button is clicked.

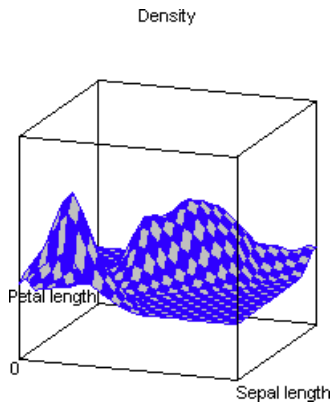
The density contours form a map showing where the data are most dense. The contours are calculated according to the quantiles, where a certain percent of the data lie outside each contour curve. These quantile density contours show where each 5% and 10% of the data are. The innermost narrow contour line encloses the densest 5% of the data. The heavy line just outside surrounds the densest 10% of the data. It is colored as the 0.9 contour because 90% of the data lie outside it. Half the data distribution is inside the solid green lines, the 50% contours. Only about 5% of the data is outside the outermost 5% contour.

One of the features of the iris length data is that there seem to be several local peaks in the density. There are two "islands" of data, one in the lower-left and one in the upper-right of the scatterplot.

These groups of locally dense data are called *clusters*, and the peaks of the density are called *modes*.

- Select **Mesh Plot** from the popup menu on the legend of the contour plot.

This produces a 3-D surface of the density, as shown to the right.



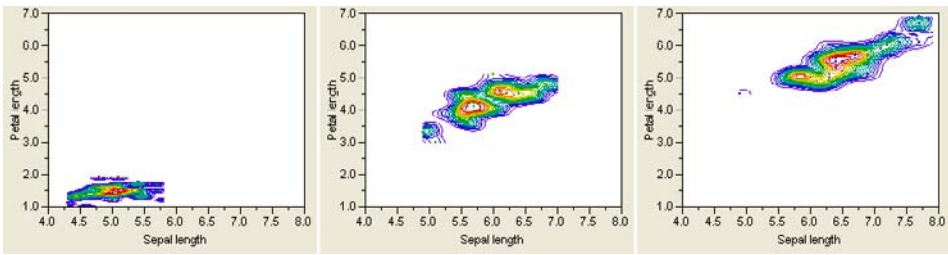
- Click and drag on the mesh plot to rotate it.

Mixtures, Modes, and Clusters

Multimodal data often comes from a mixture of several groups. Examining the data closely reveals that it is actually a collection of three species of iris: Virginica, Versicolor, and Setosa.

Conducting a bivariate density for each group results in the bivariate density plots in **Figure 15.3**. These plots have their axes adjusted to show the same scales.

Figure 15.3 Bivariate Density Curves



To classify an observation (an iris) into one of these three groups, a natural procedure would be to compare the density estimates corresponding to the petal and sepal length of a specimen over the three groups, and assign it to the group where its point is enclosed by the highest density curve. This statistical method is called *discriminant analysis* and is shown in detail in Chapter 18, "Discriminant and Cluster Analysis" on page 477.

The Elliptical Contours of the Normal Distribution

Notice that the contours of the distributions on each species are elliptical in shape. It turns out that ellipses are the characteristic shape for a bivariate Normal distribution. The Fit Y by X platform can show you a graph of these Normal contours.

- Choose **Analyze > Fit Y by X** with Petal length as Y and Sepal length as X.

When the scatterplot appears,

- Select **Group By** from the Fitting popup menu beneath the scatterplot and use Species as a grouping variable.
- Select **Density Ellipse** from the Fitting popup menu with 0.50 as the level for the ellipse.

The result of these steps is shown in **Figure 15.4**. When there is a grouping variable in effect, there is a separate estimate of the bivariate Normal density (or any fit you select) for each group. The Normal density ellipse for each group encloses the densest 50% of the estimated distribution.

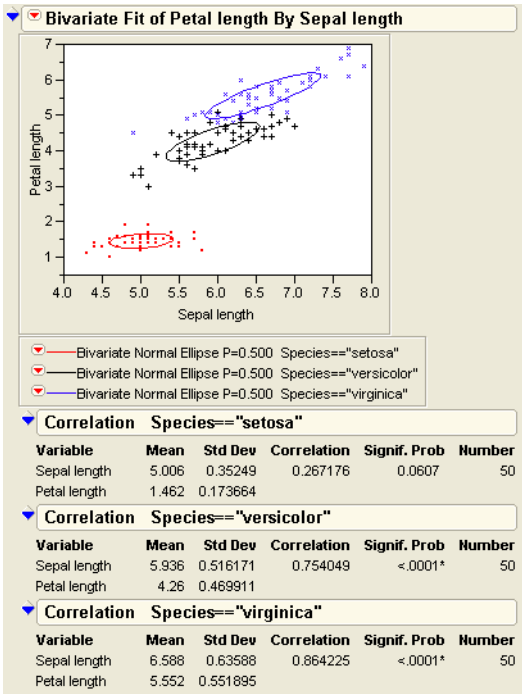
Notice that the two ellipses toward the top of the plot are fairly diagonally oriented, while the one at the bottom is not. The reports beneath the plot show the means, standard deviations, and correlation of Sepal length and Petal length for the distribution of each species. Note that the correlation is low for Setosa, and high for Versicolor and Virginica. The diagonal flattening of the elliptical contours is a sign of strong correlation. If variables are uncorrelated, then their Normal density contours appear to have a more non-diagonal shape.

One of the main uses of a correlation is to see if variables are related. You want to know if the distribution of one variable is a function of the other. When the variables are Normally distributed and uncorrelated, then the univariate distribution of one variable is the same no matter what the value of the other variable is. When the density contours have no diagonal aspect, then the density across any slice is the same no matter where you take that slice (after you Normalize the slice to have an area of one so it becomes a univariate density).

The **Density Ellipse** command in the Fit Y by X platform also gives a significance test on the correlation, which shows the *p*-value for the hypothesis that the correlation is 0.

The bivariate Normal is quite common and is very basic for analyzing data, so let's cover it in more detail with simulations.

Figure 15.4 Density Ellipses for Species Grouping Variable



Correlations and the Bivariate Normal

Describing Normally distributed bivariate data is easy because you need only the means, standard deviations, and the correlation of the two variables to completely characterize the distribution. If the distribution is not Normal, you might need a good deal more to summarize it.

Correlation is a measure, on a scale of –1 to 1, of how close two variables are to being linearly related. If you can draw a straight line through all the points of a scatterplot, then the correlation is one. The sign of the correlation reflects the slope of the regression line—a perfect negative correlation has value –1.

Simulation Exercise

As in earlier chapters, it is useful to examine simulated data created with formulas. This simulated data provides a reference point when you move on to analyze real data.

- Open Corrsim.jmp, found in the sample data subfolder.

This table has no rows, but contains formulas to generate correlated data.

- Choose **Rows > Add Rows** and enter 1000 when prompted for the number of rows wanted.

The formulas evaluate to give simulated correlations (**Figure 15.5**). There are two independent standard Normal random columns, labeled X and Y. The remaining columns (y.50, y.90, y.99, and y.1.00) have formulas constructed to produce the level of correlation indicated in the column names (0.5, 0.9, 0.99, and 1) The formula for generating a correlation r with variable X is to make the linear mix with coefficient r for X and Y computed as $\sqrt{1-r^2}$

Figure 15.5 Partial Listing of Simulated Values

Corrsim
Notes Choose Rows-->Add R
Columns (6/0)
X
y
y.50
y.90
y.99
y1.00
Rows
All rows 1000

	X	y	y.50	y.90	y.99	y1.00
1	-0.19713	-0.5179	-0.54708	-0.40316	-0.26822	-0.19713
2	0.203865	1.811595	1.670819	0.973134	0.457383	0.203865
3	0.644009	-0.67647	-0.26384	0.284741	0.542141	0.644009
4	-0.78978	-1.03112	-1.28786	-1.16026	-0.92734	-0.78978
5	2.486609	0.936286	2.054152	2.646066	2.593822	2.486609
6	-2.39251	1.187547	-0.16781	-1.63562	-2.20106	-2.39251
7	-0.1077	0.059655	-0.00219	-0.07092	-0.0982	-0.1077
8	0.61535	0.444676	0.692775	0.747645	0.671926	0.61535
9	1.06067	-1.54895	-0.81109	0.279433	0.831558	1.06067
10	1.35534	-1.69451	-0.78982	0.481187	1.102746	1.35534
11	1.499731	0.582277	1.254132	1.603566	1.566874	1.499731

You can use the Fit Y by X platform to examine the correlations:

- Choose **Analyze > Fit Y by X** with X as the X variable, and all the Y columns as the Y's.
- Hold down the Control (or ⌘) key and select **Density Ellipse** from the Fitting popup menu on the title at the top of the report. Choose **0.9** as the density level.

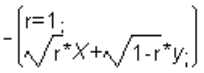
Holding down the Control (or ⌘) key causes the command to apply to all the open plots in the Fit Y by X window simultaneously.

- Do the previous step twice more with **0.95** and **0.99** as density parameters.

These steps make Normal density ellipses (**Figure 15.6**) containing 90%, 95%, and 99% of the bivariate Normal density, using the means, standard deviations, and correlation from the data.

As an exercise, create the same plot for generated data with a correlation of -1, which is the last plot shown in **Figure 15.6**. To do this:

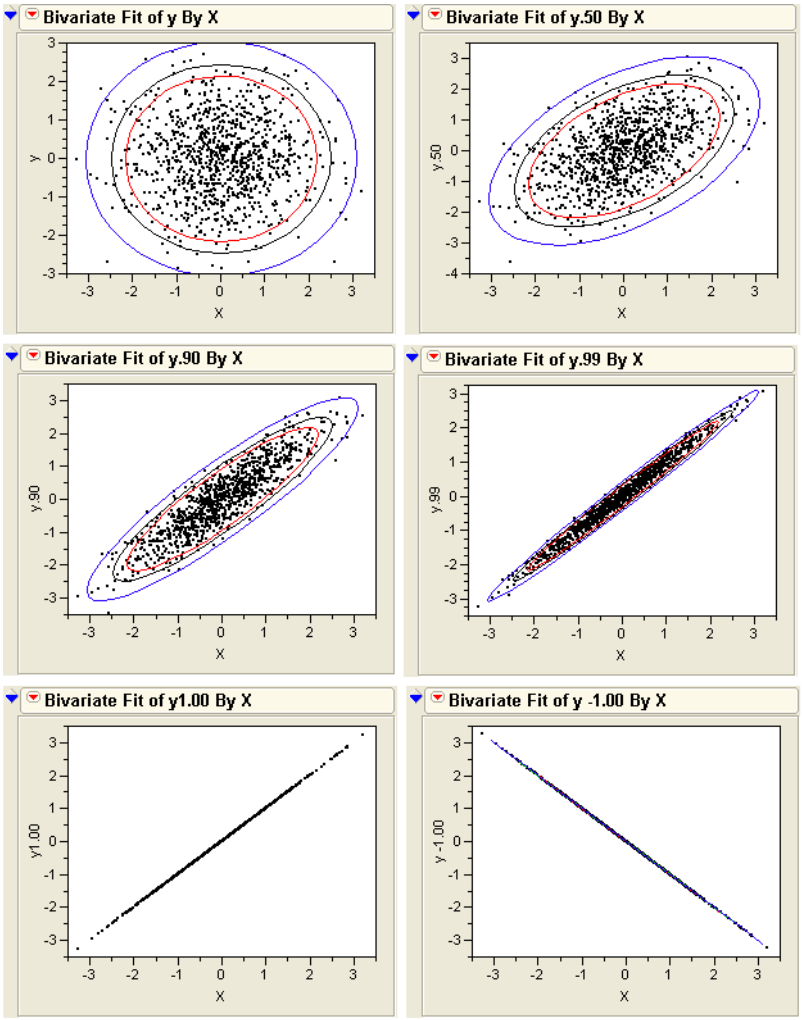
- Create a new column.
- Select **Formula** from the **Column Properties** menu on the New Column dialog or Column Info dialog.
- Enter the following formula.



Hint: Open the Formula Editor window for the variable called Y1.00. Select its formula and drag it to the Formula Editor window for the new column you are creating. With the whole formula selected, click the unary sign change button on the Formula Editor keypad.

Note in **Figure 15.6** that as the correlation grows from 0 to 1, the relationship between the variables gets stronger and stronger. The Normal density contours are circular at correlation 0 (if the axes are scaled by the standard deviations) and collapse to the line at correlation 1.

Figure 15.6 Density Ellipses for Various Correlation Coefficients



Correlations Across Many Variables

This example uses six variables. To characterize the distribution of a six-variate Normal distribution, the means, the standard deviations, and the bivariate correlations of all the pairs of variables are needed.

In a chemistry study, the solubility of 72 chemical compounds was measured with respect to six solvents (Koehler and Dunn, 1988). One purpose of the study was to see if any of the solvents were correlated—that is, to identify any pairs of solvents that acted on the chemical compounds in a similar way.

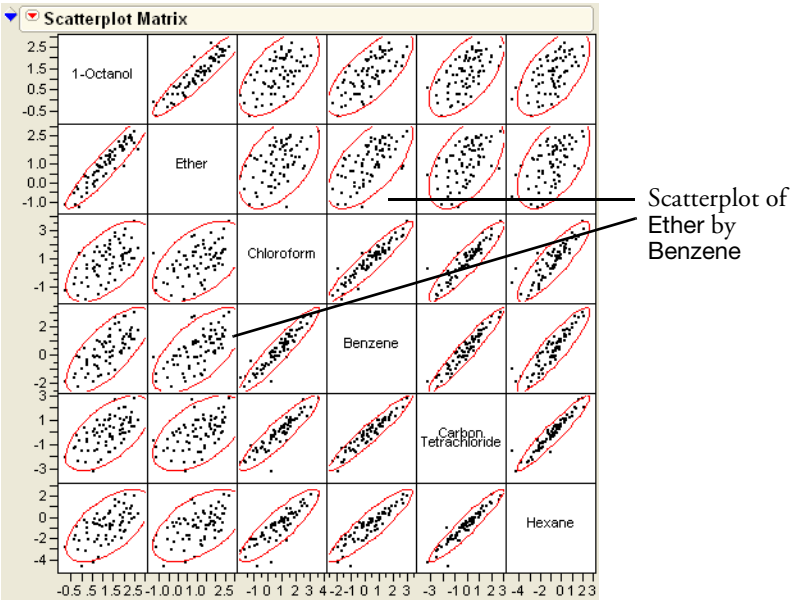
- Open Solubility.jmp to see variables (solvents) Label, 1-Octanol, Ether, Chloroform, Benzene, Carbon Tetrachloride, and Hexane.

A tag icon appears beside the Label column, which signifies that JMP will use the values in this column to label points in plots.

- Choose **Analyze > Multivariate Methods > Multivariate** with all the continuous solvent variables as Y's.

Initially, the scatterplot matrix looks like the one shown in **Figure 15.7**. Each small scatterplot can be identified by the name cells of its row and column.

Figure 15.7 Scatterplot Matrix for Six Variables



You can resize the whole matrix by resizing any one of its small scatterplots.

- Move your mouse over the corner of any scatterplot until the cursor changes into a resize arrow. Click and drag to resize the plots.

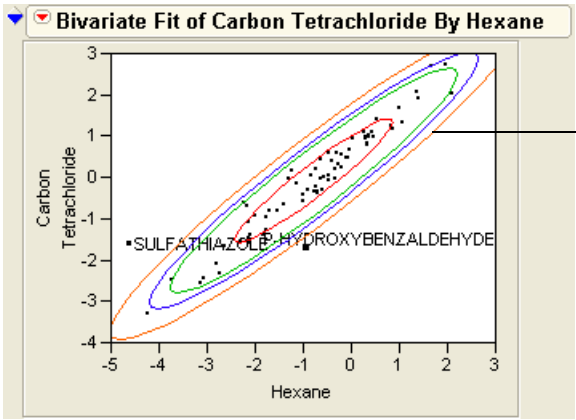
Also, you can change the row and column location of a variable in the matrix by dragging its name on the diagonal with the hand tool.

Keep the correlation report for these six variables open to use again later in this chapter.

Bivariate Outliers

Let's switch platforms to get a closer look at the relationship between carbon tetrachloride and hexane using a set of density contours.

- Choose **Analyze > Fit Y by X** with Carbon tetrachloride as Y and Hexane as X.
- Select **Density Ellipse** from the Fitting popup menu four times for arguments 0.50, 0.90, 0.95, and 0.99 to add four density contours to the plot, as shown here.



99% density ellipse. We expect 99% of the data to be inside this ellipse.

Under the assumption that the data are distributed bivariate Normal, the inside ellipse contains half the points, the next ellipse 90%, then 95%, and the outside ellipse contains 99% of the points.

Note that there are two points that are outside even the 99% ellipse.

- To make your plot look like the one above, click and then Shift-click to highlight the two outside points. With the points highlighted, choose **Rows > Label** to label them.

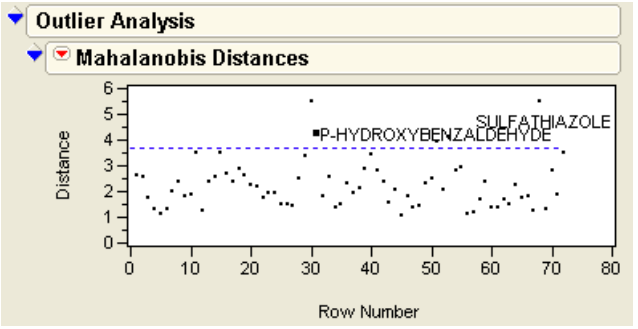
The labeled points are outliers. A point can be considered an *outlier* if its bivariate Normal density contour is associated with a very low probability. Note that "P-hydroxybenzaldehyde" is not an outlier for either variable individually (*i.e.* in a univariate sense). In the scatterplot it is near the middle of the Hexane distribution, and is barely outside the 50% limit for the Carbon tetrachloride distribution. However, it is a bivariate outlier because it falls outside the correlation pattern, which shows most of the points in a narrow diagonal elliptical area.

A common technique for computing outlier distance is the *Mahalanobis distance*. The Mahalanobis distance is computed with respect to the correlations as well as the means and standard deviations of both variables.

- Click the Multivariate platform to make it the active window and select **Outlier Analysis > Mahalanobis Distances** from its popup menu.

This command gives the Mahalanobis Distance outlier plot shown in **Figure 15.8**.

Figure 15.8 Outlier Analysis with Mahalanobis Outlier Distance Plot



The reference line is drawn using an *F*-quantile and shows the estimated distance that contains 95% of the points. In agreement with the ellipses, “Sulfathiazole” and “P-hydroxybenzaldehyde” show as prominent outliers.

- Click the scatterplot to activate it and select the brush tool from the **Tools** menu or toolbar.
- Try dragging the brush tool over these two plots to confirm that the points near the central ellipse have low outlier distances and the points outside are greater distances.
- Close this Multivariate platform window.

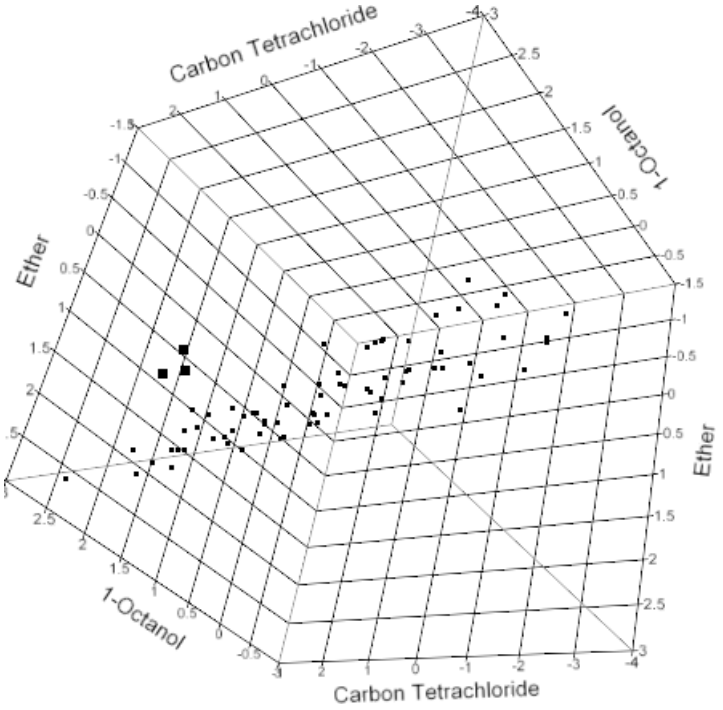
Three and More Dimensions

To consider three variables at a time, consider the first three variables, Ether, 1-octanol, and carbon tetrachloride. You can see the distribution of points with a 3-D scatterplot:

- Choose **Graph > Scatterplot 3D** and select Ether, 1-octanol, and Carbon tetrachloride as Y variables.
- When the 3-D plot appears, spin the plot and look for three-variate outliers.

The orientation in **Figure 15.9** shows three points that appear to be outlying from the rest of the points with respect to the ellipsoid-shaped distribution.

Figure 15.9 Outliers as Seen in Scatterplot 3D



Principal Components

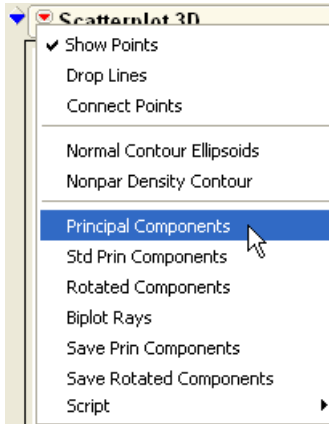
As you spin the points, notice that some directions in the space show a lot more variation in the points than other directions. This was true in two dimensions when variables were highly correlated. The long axis of the Normal ellipse showed the most variance; the short axis showed the least. Now, in three dimensions, there are three axes showing a three-dimensional ellipsoid for trivariate Normal. The solubility data seems to have the ellipsoidal contours characteristic of Normal densities, except for a few outliers.

The directions of the axes of the Normal ellipsoids are called the *principal components*. They were mentioned in the Galton example in Chapter 10, "Fitting Curves through Points: Regression" on page 235.

The *first principal component* is defined as the direction of the linear combination of the variables that has maximum variance, subject to being scaled so the sum of squares of the coefficients is one. In a 3D scatterplot, it is easy to rotate the plot and see which direction this is.

The *second principal component* is defined as the direction of the linear combination of the variables that has maximum variance, subject to it being at a right angle (orthogonal) to the first principal component. Higher principal components are defined in the same way. There are as many principal components as there are variables. The last principal component has little or no variance if there is substantial correlation among the variables. This means that there is a direction for which the Normal density hyper-ellipsoid is very thin.

The Scatterplot 3D platform can show principal components.



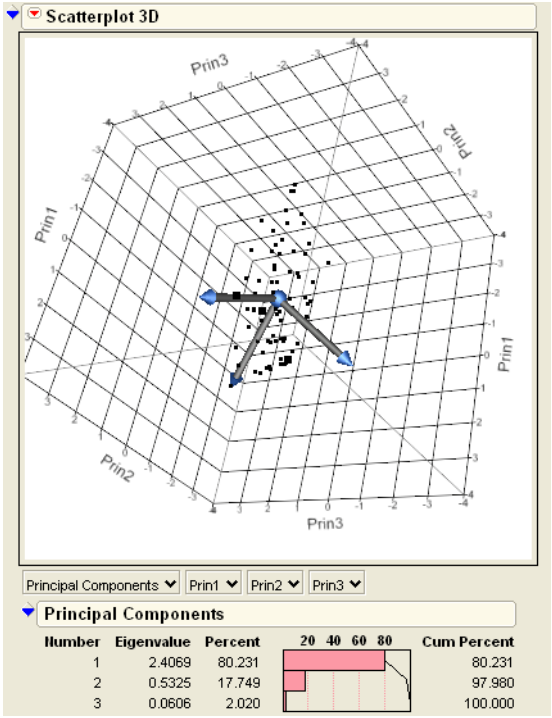
- Click the Scatterplot 3D platform shown in **Figure 15.9** to make it active.
- Select the **Principal Components** option found in the popup menu on the title bar.

This adds three principal components to the variables list and creates three new rays in the plot, as shown in **Figure 15.10**.

The directions of the principal components are shown as rays from the origin, labeled P1 for the first principal component, P2 for the second, and P3 for the third. As you rotate the plot, you see that the principal component rays correspond to the directions in the data in decreasing order of variance. You can also see that the principal components form right angles in three-dimensional space.

The Principal Components report in **Figure 15.10** shows what portion of the variance among the variables is carried by each principal component. In this example, 80% of the variance is carried by the first principal component, 17% by the second, and 2% by the third. It is the correlations in the data that make the principal components interesting and useful. If the variables are not correlated, then the principal components all carry the same variance.

Figure 15.10 Biplot Showing Principal Components and Principal Components Report

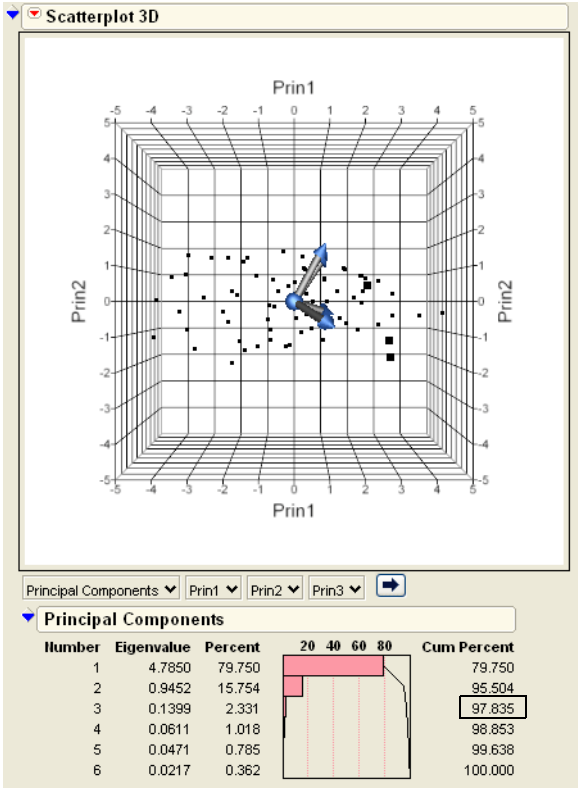


Principal Components for Six Variables

Now let's move up to more dimensions than humans can visualize. Click the Solubility.jmp data table to make it the active window and look at principal components for all six variables in the data table.

- Proceed as before. Choose **Graph > Scatterplot 3D** and select the six solvent variables as spin components.
- Select **Principal Components** in the platform popup menu, which adds six principal component axes to the plot and produces a principal component analysis.

Figure 15.11 Principal Components for Six Variables



Examine the Principal Components table (**Figure 15.11**). There is a column in the table for each of the six principal components. Note in the Cum Percent row that the first three principal components work together to account for 97.8% of the variation in six dimensions.

The first three principal components are the plot's *x*-, *y*-, and *z*-axes (**Figure 15.11**). As you spin the points, remember that you are seeing 97.8% of the variation in six dimensions as summarized by the first three principal components. This is the best three-dimensional view of six dimensions.

The rays are labeled with the names of the variables.

- Hover the cursor over one of the Principal Component arrows to see the component it represents.

You can actually measure the coordinates of each point along these axes. The measurement is accurate to the degree that the first three principal components capture the

variation in the variables. This technique of showing the points on the same plot that shows directions of the variables was pioneered by Ruben Gabriel; therefore, the plot is called a *Gabriel biplot*. The rays are actually formed from the eigenvectors in the report, which are the coefficients of the principal components on the standardized variables.

Correlation Patterns in Biplots

Note the very narrow angle between rays for Ether and 1-octanol. However, these two rays are at near right angles (in 3-D space) to the other four rays. Narrow angles between principal component rays are a sign of correlation. Principal components try to squish the data from six dimensions down to three dimensions. To represent the points most accurately in this squish, the directions for correlated variables are close together because they represent most of the same information. Thus, the Gabriel biplot shows the correlation structure of high-dimensional data.

Refer back to **Figure 15.7** to see the scatterplot matrix for all six variables. The scatterplot matrix confirms the fact that Ether and 1-octanol are highly correlated. The other four variables are also highly correlated, but there is much less correlation between these two sets of variables.

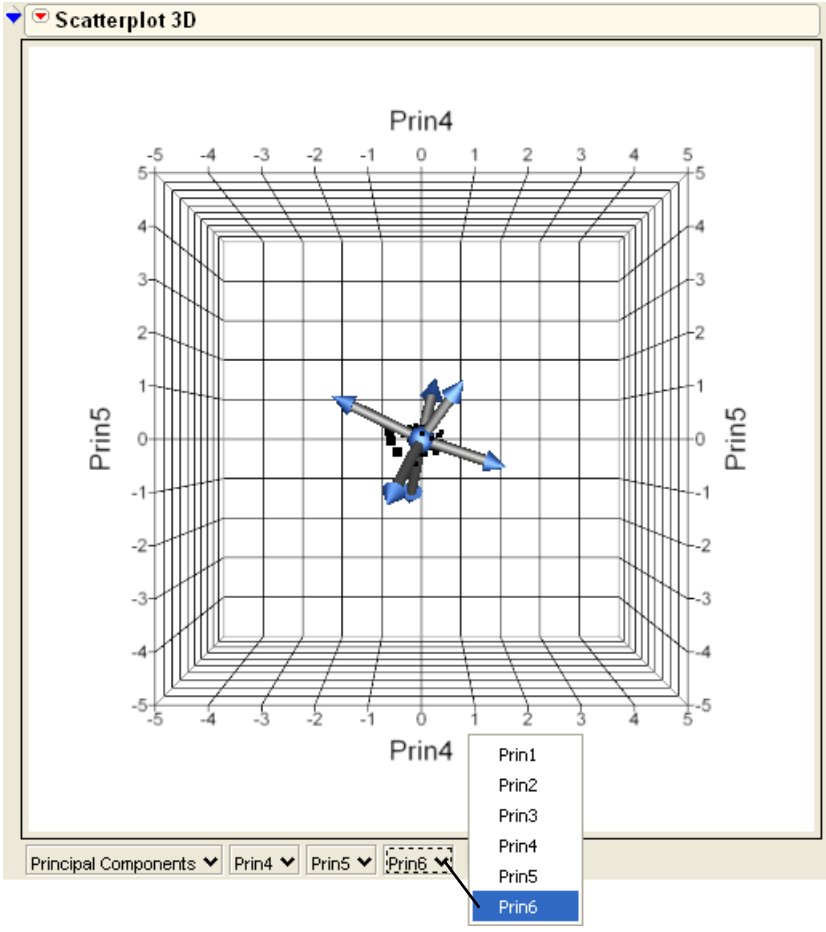
You might also consider a simple form of factor analysis, in which the components are rotated to positions so that they point in directions that correspond to clusters of variables. In JMP, this can be done in the Scatterplot 3D platform with the **Rotated Components** command.

Outliers in Six Dimensions

The Fit Y by X and the Scatterplot 3D platforms revealed extreme values in one, two, and three dimensions. These outliers also show in six dimensions. However, there could be additional outliers that violate the higher dimensional correlation pattern. In six dimensions, some of the directions of the data are quite flat, and there could be an outlier in that direction that wouldn't be revealed in an ordinary scatterplot.

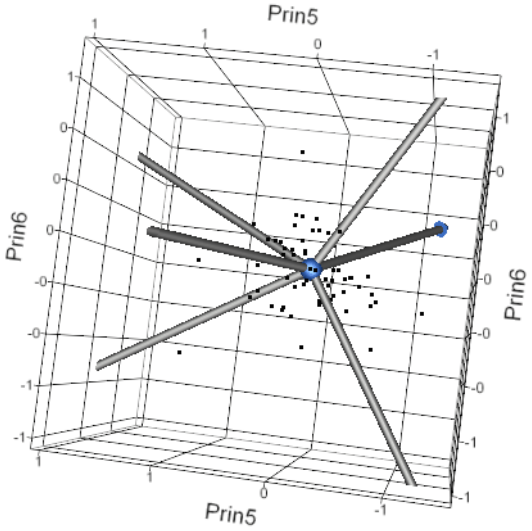
- In the Scatterplot 3D platform with six variables, change the displayed principal components to the last three principal components, as shown in **Figure 15.12**.

Figure 15.12 Biplot Showing Six Principal Components



Now you are seeing the directions in six-dimensional space that are least prominent for the data. The data points are crowded near the center because you are looking at a small percentage of the variability.

- Adjust the axes by dragging on their extremes to zoom in to the center of the plot.
- Highlight and label the points that seem to stand out.

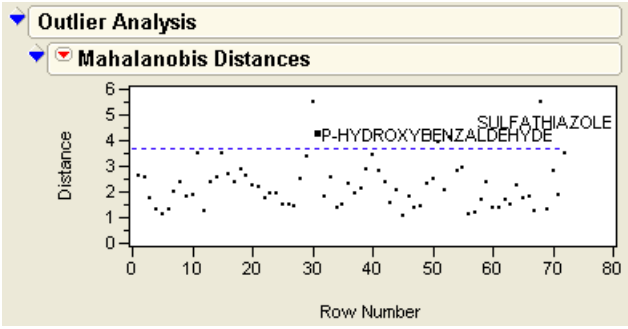


These are the points that are in the directions that are most unpopular. Sometimes they are called *class B outliers*.

All the outliers from one dimension to six dimensions should show up on an outlier distance plot that measures distance with respect to all six variables.

- Click (to activate) the Multivariate platform previously generated for all six variables.
- Select **Outlier Analysis > Mahalanobis Distances** from the platform menu to see the six-dimensional outlier distance plot in **Figure 15.13**. (If you closed the correlation window, choose **Analyze > Multivariate Methods > Multivariate** with all six responses as Y variables, and then select **Outlier Analysis** from the platform menu.)

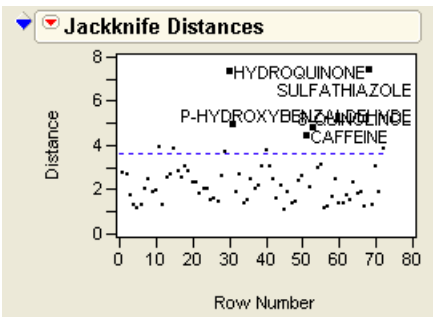
Figure 15.13 Outlier Distance Plot for Six Variables



There is a refinement to the outlier distance that can help to further isolate outliers. When you estimate the means, standard deviations, and correlations, all points—including outliers—are included in the calculations and affect these estimates, causing an outlier to disguise itself.

Suppose that as you measure the outlier distance for a point, you exclude that point from all mean, standard deviation, and correlation estimates.

This technique is called *jackknifing*. The jackknifed distances often make outliers stand out better.



To see the Jackknifed Distance plot shown to the right, select the **Outlier Analysis > Jackknife Distances** option from the popup menu at the top of the Outliers analysis.

Summary

When you have more than three variables, the relationships among them can get very complicated. Many things can be easily found in one, two, or three dimensions, but it is hard to visualize a space of more than three dimensions.

The histogram provides a good one-dimensional look at the distribution, with a smooth curve option for estimating the density. The scatterplot provides a good two-dimensional look at the distribution, with Normal ellipses or bivariate smoothers to study it. In three dimensions,

Scatterplot 3D provides the third dimension. To look at more than three dimensions, you must be creative and imaginative.

One good basic strategy for high-dimensional exploration is to take advantage of correlations and reduce the number of dimensions. The technique for this is principal components, and the graph is the Gabriel biplot, which shows all the original variables as well as the points in principal component space.

You can also use highlighting tools to brush across one distribution and see how the points highlight in another view.

The hope is that you either find patterns that help you understand the data, or points that don't fit patterns. In both cases you can make valuable discoveries.

Exercises

1. In the data table Crime.jmp, data are given for each of the 50 U.S. states concerning crime rates per 100,000 people for seven classes of crimes.

   (a) Use the Multivariate platform to produce a scatterplot matrix of all seven variables. What pattern do you see in the correlations?

   (b) Conduct an outlier analysis (using Mahalanobis distance) of the seven variables, and note the states that seem to be outliers. Do the outliers seem to be states with similar crime characteristics?

   (c) Conduct a principal components analysis on the correlations of these variables. Then, spin the principal components, assigning the *x*-, *y*-, and *z*-axes to the first three principal components. Which crimes seem to group together?

   (d) It is impossible to graph all seven variables at once. Using the eigenvalues from the principal components report, how many dimensions would you retain to accurately summarize these crime statistics?

2. The data in Socioeconomic.jmp (SAS Institute 1988) consists of five socioeconomic variables for twelve census tracts in the Los Angeles Standard Metropolitan Statistical Area.

   (a) Use the Multivariate platform to produce a scatterplot matrix of all five variables.

   (b) Conduct a principal components analysis (on the correlations) of all five variables. Considering the eigenvalues produced in the report, how many factors would you use for a subsequent rotation?

(c) Rotate the number of factors that you determined in part (b). Which variables load on each factor?

16 Design of Experiments

Overview

Designed experiments are an important technology in science and engineering. We learn by trial and error, *i.e.* experimentation, and experimental design makes that learning process efficient and reliable.

Design of Experiments (DOE) is probably the single most powerful technique you can use for product development, refinement, problem solving, scientific inquiry, and optimization of products and processes.

This chapter covers how JMP constructs and analyzes experimental designs. See the *JMP Design of Experiments* guide in the online help for details about designs not covered in this chapter.

Introduction

Experimentation is the fundamental tool of the scientific method. In an experiment, a *response* of interest is measured as some *factors* are changed systematically through a series of *runs*, also known as *trials*. Those factors that are not involved in the experiment are held (as much as is practical) at a constant level, so that any variation or effect produced by the experiment can be attributed to the changes in the factor and natural variability. The goal is to determine if and how the factors affect the response.

Experimental design addresses this goal, allowing us to learn the most about the relationship from the fewest number of runs. This is an important reason to use experimental design—it saves money by gleaning maximum knowledge from the fewest number of experimental trials.

Experimentation Is Learning

The word *experiment* has the same root as the words *expert* and *experience*. They all derive from the Latin verb *experior*, which means to try, to test, to experience, or to prove.

Experimentation is the familiar process of trial and error. Try something and see what happens. Learn from experience. Learn from changing factors and observing the results. This is the *inductive method* that encompasses most learning. Designed experiments add to this method by giving us a framework to direct our experiences in a meaningful way.

Controlling Experimental Conditions Is Essential

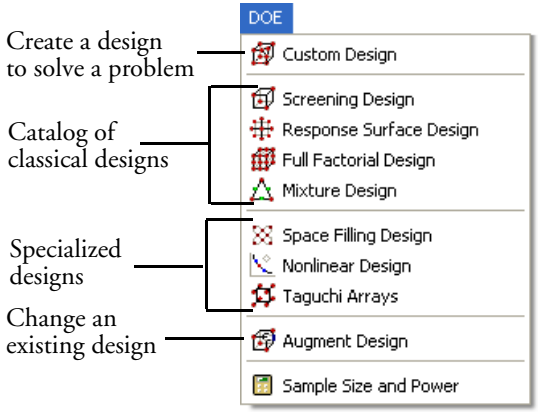
We are easily fooled unless we take care to control the experimental conditions. Their control is the critical first step in doing scientific experiments. This is the step that distinguishes experimental results from observational, happenstance data. You may obtain clues of how the world works from non-experimental data, but you cannot put full trust into learning from observational phenomena because you are never completely sure why the response changed. Were the response differences were due to changes in the factor of interest, or some change in uncontrolled variables?

Experiments Manage Random Variation within A Statistical Framework

Experiments are never exactly repeatable because of the inevitable uncontrollable, random component of the response. Understanding how to model this random variation was one of the first triumphs of the new “statistical science” of the 1920’s. Important techniques like randomization and blocking, standard statistical distributions, and tests that they developed are still in use today.

JMP DOE

The DOE platform in JMP is an environment for describing the factors, responses, and other specifications, creating a designed experiment, and saving it in a JMP table. When you select the DOE menu, you see the list of designs shown to the right.



The JMP Custom Designer builds a design for your specific problem that is consistent with your resource budget. You can use the Custom Designer for routine factor screening, response optimization, and mixture problems. Also, the Custom Designer can find designs for special conditions not covered in the lists of predefined designs.

A Simple Design

The following situation helps to acclimate you to the design capabilities of JMP.

The Experiment

Acme Piñata Corporation discovered that its piñatas were too easily broken. The company wants to perform experiments to discover what factors might be important for the peeling strength of flour paste.

In this design, we search through nine factors to discover two things.

- Which factors actually affect the peel strength?
- What settings should those factors take in order to optimize the peel strength?

The Response

Strength refers to how well two pieces of paper that are glued together resist being peeled apart.

The Factors

Batches of flour paste were prepared to determine the effect of the following nine factors on peeling strength:

Flour:1/8 cup of white unbleached flour or 1/8 cup of whole wheat flour

Sifted: flour was sifted or not sifted

Type: water-based paste or milk-based paste

Temp: mixed when liquid was cool or when liquid was warm

Salt: formula had a dash of salt or had no salt

Liquid: 4 teaspoons of liquid or 5 teaspoons of liquid

Clamp: pasted pieces were tightly clamped together or not clamped during drying

Sugar: formula contained 1/4 teaspoon or no sugar

Coat: whether the amount of paste applied was thin or thick

The Budget

There are many constraints that can be put on a design, arising from many different causes. In this case, we are only allotted enough time to complete 16 runs.

Enter and Name the Factors

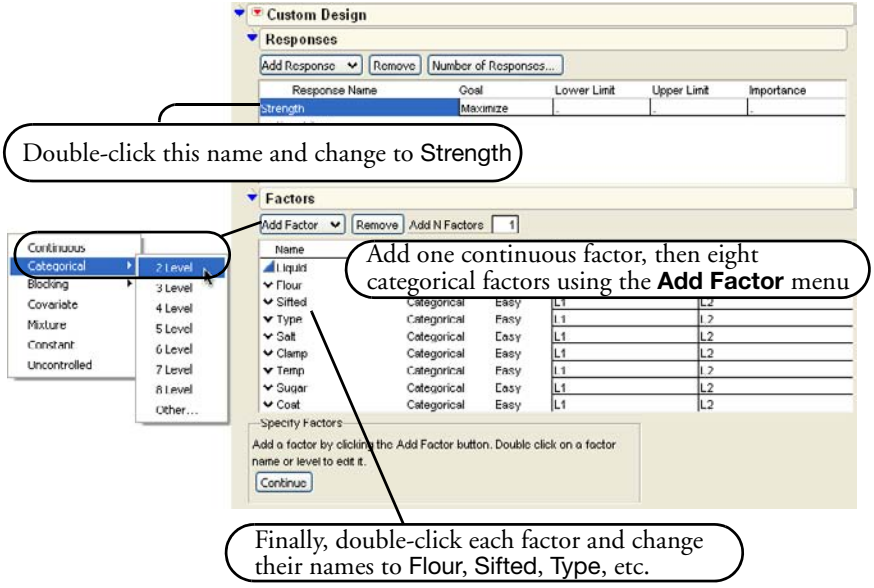
To begin, select a DOE platform and define the factors of the experiment.

- Choose **DOE > Custom Design**.

You see one default response called Y in the **Responses** panel.

- Double click on the name and change it to Strength.

Figure 16.1 Dialog for a Designating Responses and Factors



Now, complete the following steps to add the nine factors:

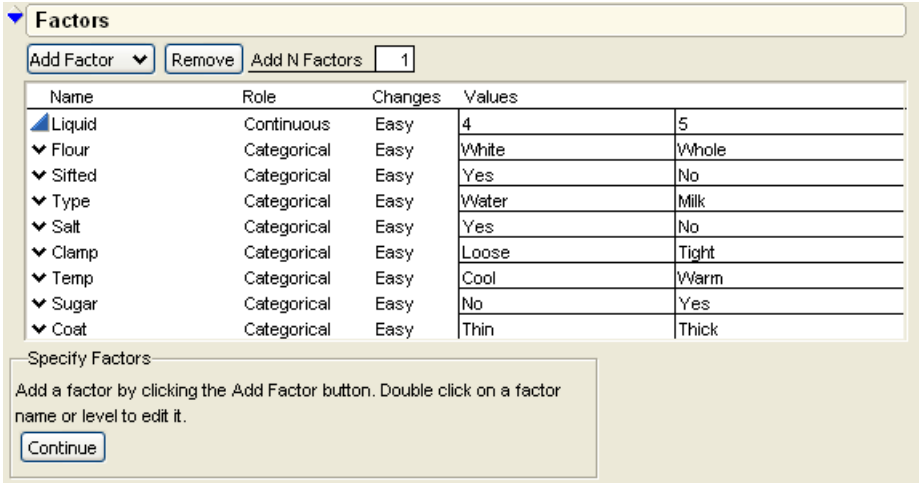
- Click the **Add Factor** button and select **Continuous** to add 1 continuous factor;
- Type 8 into the Add N Factors field and select **Add Factors > Categorical > 2 Level**.
- Double-click each factor and type its name.

You should now have the dialog shown in **Figure 16.1**.

- Now, add the appropriate values to specify the levels of each factor. For example, Liquid can take levels 4 and 5. Flour takes levels White and Whole. Continue this for each of the nine factors.

Your **Factors** panel should look like the one in **Figure 16.2**.

Figure 16.2 Factors Panel



Click **Continue** to proceed to the next step.

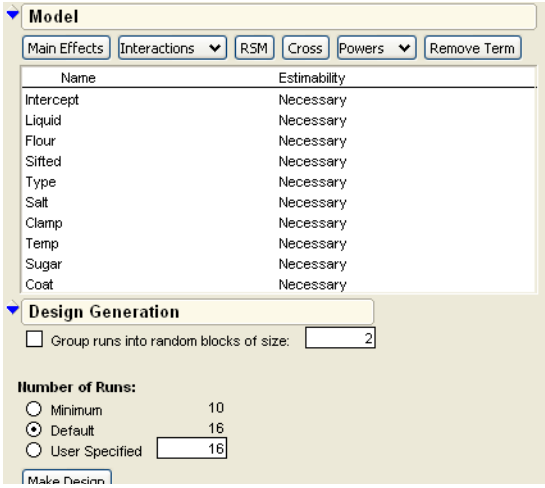
Define the Model

Three new panels appear: **Define Factor Constraints**, **Model**, and **Design Generation** (**Figure 16.3**).

We have no mathematical constraints (aside from the number of runs, addressed later), so we do not need to modify the **Define Factor Constraints** panel.

By default, the intercept and all the main effects are entered into the model. Other model terms, like interactions and powers, should be entered at this point. We only examine main effects in this first example (typical of screening designs), so we do not need to modify the **Model** panel.

Figure 16.3 Model Definition Panels



The **Design Generation** panel allows us to specify the number of experimental runs (or *trials*). JMP has several suggestions to choose from.

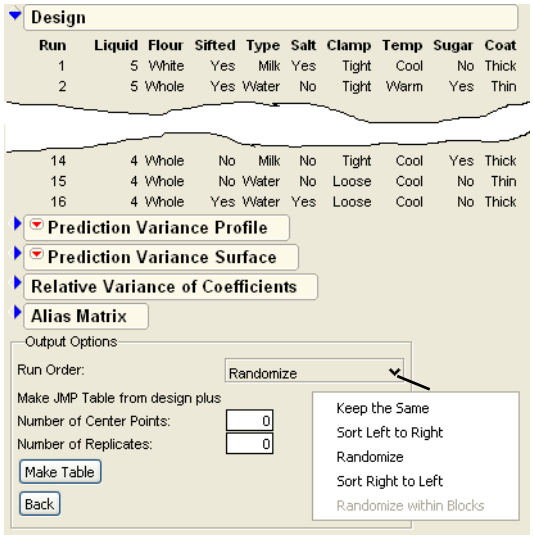
- **Minimum** is the smallest number of runs that allow estimation of the main effects.
- **Default** is a suggestion that is larger than minimum but smaller than a full factorial design, often giving balanced designs using an economical number of runs.
- Depending on your budget, you may be able to afford a different number of runs than those suggested. The **Number of Runs** box accepts any number of runs. You are in no way limited to the suggestions that JMP makes.

Make sure the **Default** selection of 16 runs is selected, then click the **Make Design** button.

JMP now searches for an optimal design. In essence, JMP searches along a mathematical surface for optimal values. Since there is always the risk of finding local maxima rather than global maxima, JMP completes several independent searches, shown in a progress bar.

Once completed, the design and several diagnostic plots appear.

Figure 16.4 Design and Diagnostics



Your design may not look like this design. JMP optimizes the design based on a mathematical criterion, but in many cases, there are several equivalent designs.

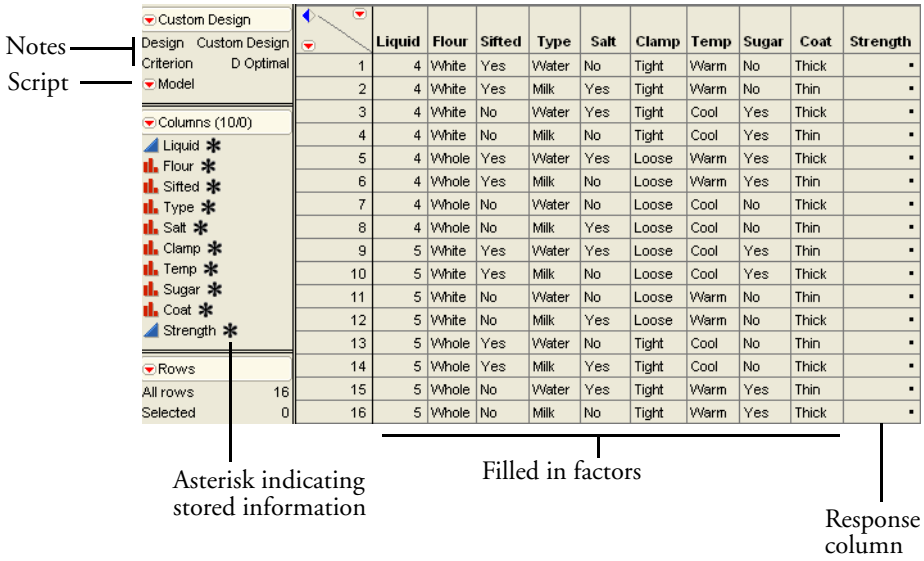
At this point, JMP is ready to prepare the data table. In most designs, it is desirable to randomize the order of the trials, but there are occasions when the runs should be sorted. For illustration, we sort this data table. The **Run Order** drop-down (**Figure 16.4**) shows these options.

Choose **Sort Right to Left** for the Run Order and click **Make Table.**

JMP generates a data table with several convenient features.

- All the factors have their settings filled in.
- The response column is present, but empty, waiting for the results of your experiment.
- A note is present (upper left corner) that shows the type of design.
- Limits and coding information is stored as column properties for each variable.
- A **Model** script holds all the information about the requested model. The **Fit Model** dialog looks for this script and completes itself automatically based on the script's contents.

Figure 16.5 Flour Data Table



	Liquid	Flour	Sifted	Type	Salt	Clamp	Temp	Sugar	Coat	Strength
1	4	White	Yes	Water	No	Tight	Warm	No	Thick	.
2	4	White	Yes	Milk	Yes	Tight	Warm	No	Thin	.
3	4	White	No	Water	Yes	Tight	Cool	Yes	Thick	.
4	4	White	No	Milk	No	Tight	Cool	Yes	Thin	.
5	4	Whole	Yes	Water	Yes	Loose	Warm	Yes	Thick	.
6	4	Whole	Yes	Milk	No	Loose	Warm	Yes	Thin	.
7	4	Whole	No	Water	No	Loose	Cool	No	Thick	.
8	4	Whole	No	Milk	Yes	Loose	Cool	No	Thin	.
9	5	White	Yes	Water	Yes	Loose	Cool	Yes	Thin	.
10	5	White	Yes	Milk	No	Loose	Cool	Yes	Thick	.
11	5	White	No	Water	No	Loose	Warm	No	Thin	.
12	5	White	No	Milk	Yes	Loose	Warm	No	Thick	.
13	5	Whole	Yes	Water	No	Tight	Cool	No	Thin	.
14	5	Whole	Yes	Milk	Yes	Tight	Cool	No	Thick	.
15	5	Whole	No	Water	Yes	Tight	Warm	Yes	Thin	.
16	5	Whole	No	Milk	No	Tight	Warm	Yes	Thick	.

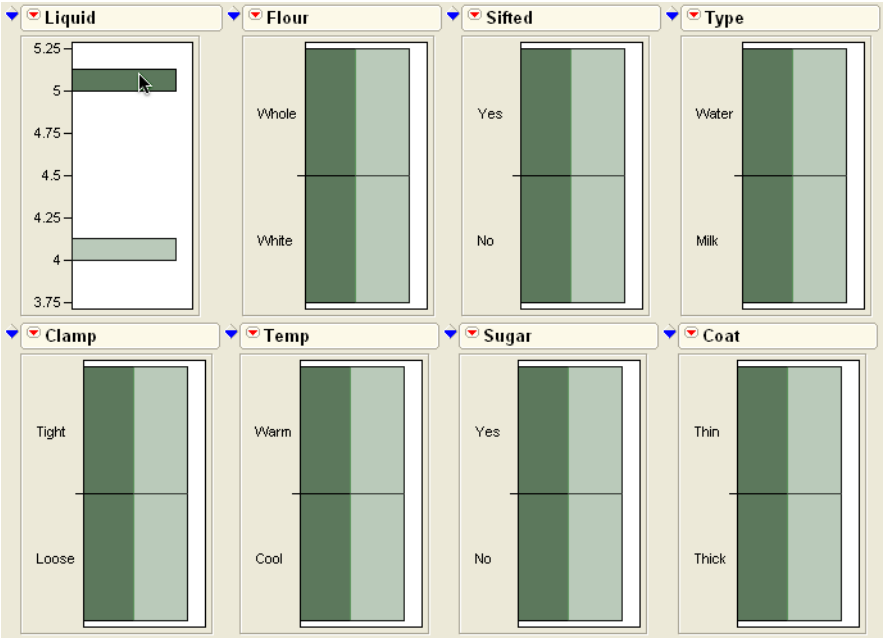
Is the Design Balanced?

It is easy to show that the Custom designer produces balanced designs when possible. To check that this design is balanced,

- Choose **Analyze > Distribution** for all nine factor variables.
- Click in each histogram bar to see that the distribution is flat for all the other variables.

The highlighted area representing the distribution for a factor level is equal in each of the other histograms, as shown in **Figure 16.6**.

Figure 16.6 Histograms Verify that the Design Is Balanced



Perform Experiment and Enter Data

At this point, you would run the 16 trials of the experiment, noting the strength readings for each trial, then entering these results into the Strength column.

- Rather than re-enter the data into the generated design data table, open the sample data table Flrpaste.jmp.

Strength
7.8
17.4
8.2
3.7
6.5
7.1
4.3
6.1
4.8
3.5
4.3
2.9
3.7
9
1.8
4.1

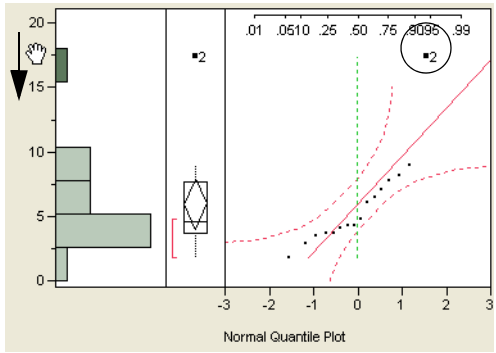
Note: The values for Strength, shown to the right, correspond to the order of the data in the Flrpaste.jmp sample data table, which may be different than the table you generated with the Custom Designer.

Examine the Response Data

As usual, a good place to start is by examining the distribution of the response, which in our example is the peel strength.

- Choose **Analyze > Distribution** and assign Strength as the **Y** variable.
- Click **OK**.
- When the histogram appears, select **Normal Quantile Plot** from the popup menu on the title bar of the histogram.

You should now see the plots shown to the right.



- Drag downward on the upper part of the histogram axis to increase the maximum so that the extreme points show clearly.
- Click on the highest point to identify it as row number 2.

The box plot and the Normal quantile plot are useful for identifying runs that have extreme values. In this case, run 2 has an unusually high peeling strength.

Analyze the Model

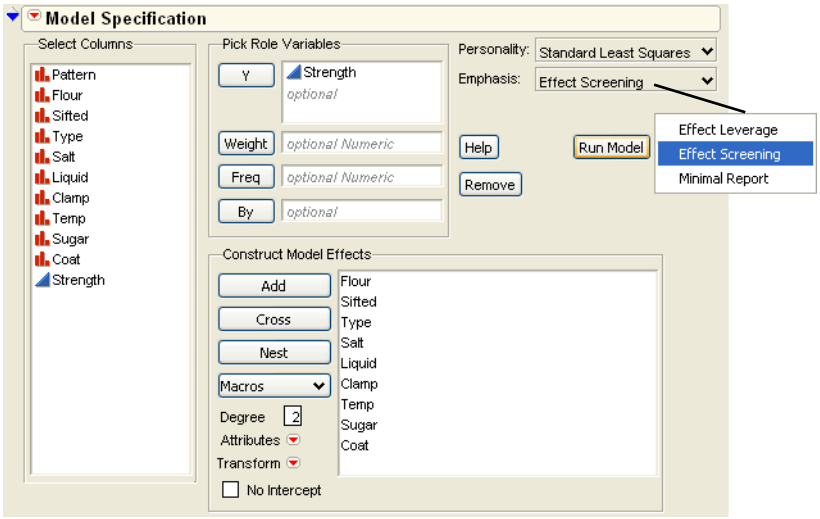
Of the nine factors in the flour paste experiment, there may be only a few that stand out in comparison with the others. The goal of the experiment is to find the factor combinations that optimize the predicted response (peeling strength); it is not to show statistical significance of model effects. This kind of experimental situation lends itself to an effect screening analysis.

- Choose **Analyze > Fit Model**.
- When the Fit Model dialog appears, select Strength as the response (Y) variable.
- Shift-click all the columns from Flour to Coat in the variable selection list and click **Add** to make nine effects in the model.

Figure 16.7 shows the completed dialog.

- Make sure that **Effect Screening** is selected on the Emphasis popup menu, and then click **Run Model**.

Figure 16.7 Fit Model Dialog for Screening Analysis of Flour Paste Main Effects



The Analysis of Variance table to the right shows that the model as a whole is not significant (p = 0.1314). The most significant factors are Flour and Type (of liquid). Note in the **Summary of Fit** table that the standard deviation of the error (Root Mean Square Error) is estimated as 2.6, a high value relative to the scale of the response. The R^2 of 0.79 is not particularly high.

Response Strength

Summary of Fit

RSquare	0.794129
RSquare Adj	0.485323
Root Mean Square Error	2.641654
Mean of Response	5.95
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	161.51000	17.9456	2.5716
Error	6	41.87000	6.9783	Prob > F
C. Total	15	203.38000		0.1314

Parameter Estimates

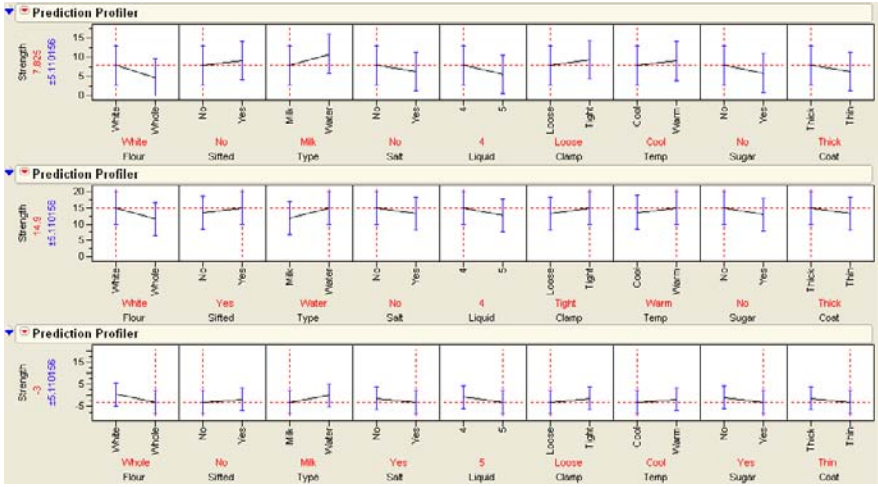
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.95	0.660413	9.01	0.0001*
Flour[White]	1.6875	0.660413	2.56	0.0432*
Sifted[No]	-0.625	0.660413	-0.95	0.3805
Type[Milk]	-1.525	0.660413	-2.31	0.0603
Salt[No]	0.775	0.660413	1.17	0.2851
Liquid[4]	1.1375	0.660413	1.72	0.1358
Clamp[Loose]	-0.775	0.660413	-1.17	0.2851
Temp[Cool]	-0.6125	0.660413	-0.93	0.3895
Sugar[No]	1	0.660413	1.51	0.1807
Coat[Thick]	0.8125	0.660413	1.23	0.2646

Note: You can double-click columns in any report to specify the number of decimals to display.

Following the statistical report, you see a Prediction Profiler, which shows the predicted response for each combination of factor settings. **Figure 16.8** shows three manipulations of the Prediction Profiler for the flour experiment. The settings of each factor are connected by a line, called the *prediction trace* or *effect trace*. You can grab and move each vertical dotted line in the Prediction Profile plots to change the factor

settings. The predicted response automatically recomputes and shows on the vertical axis, and the prediction traces are redrawn.

Figure 16.8 Screening Model Prediction Profiler for Flour Paste Experiment



The Prediction Profiler lets you look at the effect on the predicted response of changing one factor setting while holding the other factor settings constant. It can be useful for judging the importance of the factors.

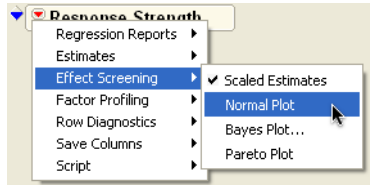
The effect traces in the plot at the top of **Figure 16.8** show a larger response for white flour than for whole wheat flour, and for water than for milk, which indicates that changing them to their higher positions increases peel strength.

The second plot in **Figure 16.8** shows what happens if you click and move the effect trace to the high response of each factor; the predicted response changes from 7.825 to 14.9. This occurs with sifted white flour, 4 teaspoons warm water, no salt, no sugar, pasted applied thickly, and clamped tightly while drying.

If you had the opposite settings for these factors (bottom plot in **Figure 16.8**), then the linear model would predict a surface strength of –3 (a clearly impossible response). The prediction is off the original scale, though you can double-click in the *y*-axis area to change the *y* scale of the plot.

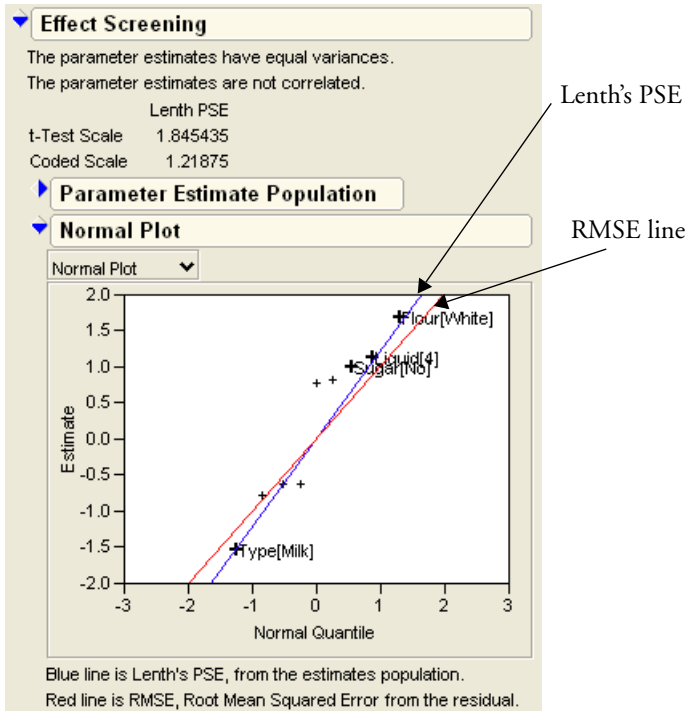
Commands in the popup menu on the Response Strength title bar give you other options.

Select **Normal Plot** from the **Effect Screening** submenu on the analysis title bar to display the Normal Plot of the parameter estimates shown in **Figure 16.9**.



The *Normal Plot* is a Normal quantile plot (Daniel 1959), which shows the parameter estimates on the vertical axis and the Normal quantiles on the horizontal axis. In a screening experiment, you expect most of the effects to be inactive, to have little or no effect on the response. If that is true, then the estimates for those effects are a realization of random noise centered at zero. What you want is a sense of the magnitude of an effect you should expect when it is truly active instead of just noise. On a Normal Plot, the active effects appear as outliers that lie away from the line that represents Normal noise.

Figure 16.9 Normal Plot Shows Most Influential Effect



Looking at responses in this way is a valid thing to do for two-level balanced designs, because the estimates are uncorrelated and all have the same variance. The Normal Plot is a useful way (and about the only way) to evaluate the results of a saturated design with no degrees of freedom for estimating the error.

The Normal Plot also shows the straight line with slope equal to the *Lenth's PSE* (pseudo standard error) estimate (Lenth 1989). This estimate is formed by taking 1.5 times the median absolute value of the estimates after removing all the estimates greater than 2.75 times the median absolute estimate in the complete set of estimates. Lenth's PSE is computed using the Normalized estimates and disregards the intercept. Effects that deviate substantially from this Normal line are automatically labeled on the plot.

Usually, most effects in a screening analysis have small values; a few have larger values. In the flour paste example, there appears to be only noise, because none of the effects separate from the Normal lines more than would be expected from a Normal distribution of the estimates. This experiment actually has the opposite of the usual distribution. There is a vacant space near the middle of the distribution—none of the estimates are very near zero. (Note in your analysis tables that the Whole-Model *F*-test shows that the model as a whole is not significant.)

So far, the experiment has not been conclusive for most of the factors, although white flour does appear to give better results than whole wheat flour. The experiment might bear repeating with better control of variability and a better experimental procedure.

Details of the Design

There are a few details of this design that we did not discuss during the analysis. For example, how the nine factors plus an intercept can be investigated with only 16 runs. This is possible because we assume that all interactions have no effect and they are confounded with the main effects. The interactions are not estimable if added to the model.

Confounding Structure

Since this experiment needs to look at many factors in a very few runs, many main effects are confounded with two-way interactions. In other words, we assume that there is no interaction among the main effects, so any interaction terms are zero. Since we assume they are zero, we allow the design to use a single coefficient to estimate the main effects and the interaction terms that the effect is confounded with. It is informative to look at the confounding structure of the design. To do this:

- Return to the Custom Design window where you specified the factors, runs, and model.
- Open the **Aliasing of Effects** node by clicking on the triangle next to the outline header.

Figure 16.10 Aliasing Report

Alias Matrix

Effect	1 2	1 3	1 4	1 5	1 6	1 7	1 8	6 7	6 8	6 9	7 8	7 9	8 9
Intercept	0	0	0	0	0	0	0	0	0	0	0	0	0
Flour	0	0	0	0	0	0	0	0	-1	0	0	0	0
Sifted	0	0	0	0	0	0	0	0	0	0	0	1	0
Type	0	0	0	0	0	0	0	0	0	0	1	0	0
Salt	0	0	0	0	0	0	0	0	0	0	0	0	0
Liquid	0	0	0	0	0	-1	0	0	0	0	0	0	0
Clamp	0	0	0	0	0	0	-1	0	0	0	0	0	0
Temp	0	0	0	-1	0	0	0	0	0	0	0	0	0
Sugar	0	0	0	0	-1	0	0	0	0	0	0	0	0
Coat	0	0	0	0	0	0	0	0	0	0	0	0	0

Here you can see that the Temp effect is confounded with at least one second-order interaction (and possibly more in the section of the report that has been cut out). The confounding involves columns **1** and **5**, indicated by the **1 5** at the top of the column containing the -1. **1 5** is a compact way of designating the Flour*Liquid term, since Flour is the first variable in the effects list and Liquid is the fifth. Luckily, you do not have to understand the details of aliasing in order to use the Custom Designer.

Using the Custom Designer

Modify a Design Interactively

There is a **Back** button at several stages in the design dialog that allows you to go back to a previous step and modify the design. For example, you can modify a design by adding quadratic terms to the model, removing the center points, or removing the replicate.

How the Custom Designer Works

The Custom Designer starts with a random design where each point is inside the range of each factor. The computational method is an iterative algorithm called *coordinate exchange*. Each iteration of the algorithm involves testing every value of each factor in the design to determine if replacing that value increases the optimality criterion. If so, the new value replaces the old. This process continues until no replacement occurs in an entire iteration.

To avoid converging to a local optimum, the whole process is repeated several times using a different random start. The designer displays the best of these designs.

Sometimes a design problem can have several equivalent solutions. Equivalent solutions are designs having equal precision for estimating the model coefficients as a group. When this is true, the design algorithm will generate different (but equivalent) designs if you press the **Back** and **Make Design** buttons repeatedly.

Custom designs give the most flexibility of all design choices. The Custom Designer gives you the following options:

- continuous factors
- categorical factors with arbitrary numbers of levels
- mixture ingredients
- covariates (factors that already have unchangeable values and design around them)
- blocking with arbitrary numbers of runs per block
- interaction terms and polynomial terms for continuous factors
- inequality constraints on the factors
- choice of number of experimental runs to do, which can be any number greater than or equal to the number of terms in the model
- selecting factors (or combinations of factors) whose parameters are only estimated if possible

After specifying all your requirements, the Custom Design generates an appropriate optimal design for those requirements. In cases where a classical design (such as a factorial) is optimal, the Custom Designer finds them. Therefore, the Custom Designer can serve any number or combination of factors.

Using the Screening Platform

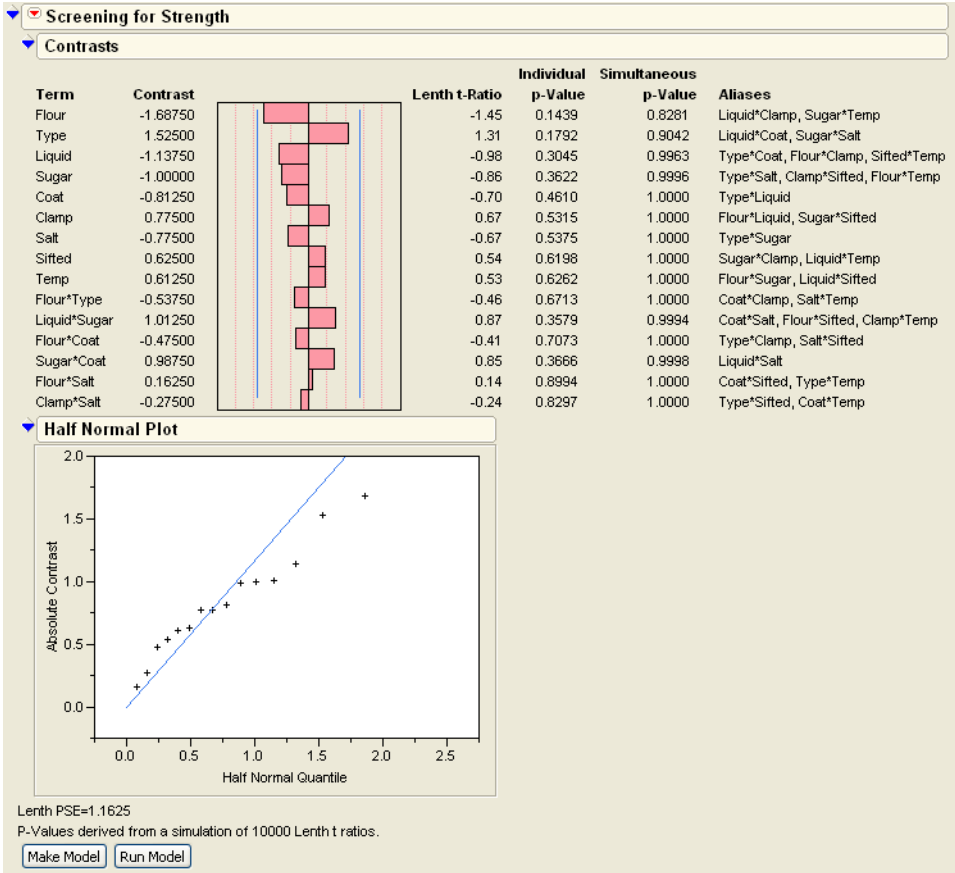
Screening designs like the flour paste experiment can be analyzed with the Screening platform rather than the Fit Model platform. Although the Fit Model platform performs excellent general-purpose analyses, the Screening platform uses calculations that are especially suited for screening designs, and in some cases is better at finding active factors.

To analyze this data using the Screening platform,

- Select **Analyze > Modeling > Screening**.
- Add the effects (all the categorical factors except Pattern) as **X** and Strength as **Y**.

This produces the report shown in **Figure 16.11**. The report is divided into two sections. The upper half shows *p*-values and other numerical quantities that describe the proposed model. The bottom half shows a half-normal plot that is useful for visually identifying active model effects.

Figure 16.11 Initial Screening Report



Contrasts and *p*-values

The second column of the text report contains contrasts, which are equivalent to parameter estimates in orthogonal designs like this one. The *t*-ratios are calculated using a simulation, and they give individual *p*-values that are interpreted in the usual way. Simultaneous *p*-values are used in models where multiple comparisons are needed.

In this example, there are no significant *p*-values.

Half-Normal Plot

The half-normal plot visually shows active effects. If there were any active effects in this model, they would be highlighted and labeled in this plot. Since there are no effects showing as active, we conclude that there are no significant effects in the model.

Screening for Interactions: The Reactor Data

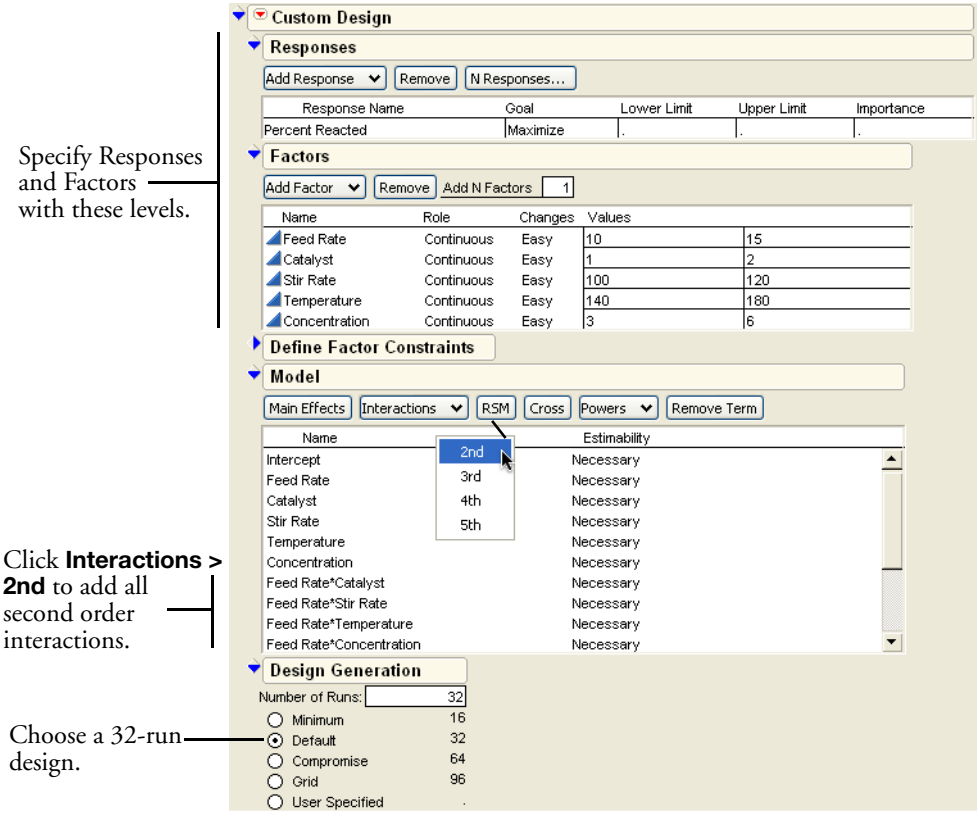
Now that you've seen the basic flow of a design in JMP, we will illustrate a more complicated one that involves both main effects and interactions.

Box, Hunter, and Hunter (2005, p. 259) discuss a study of chemical reactors that has five two-level factors, Feed Rate, Catalyst, Stir Rate, Temperature, and Concentration. The purpose of the study is to find the best combination of settings for optimal reactor output. It is also known that there may be interactions among the factors.

A full factorial for five factors requires $2^5 = 32$ runs. You can generate the design table using the **Full Factorial** selection on the **DOE** main menu. However, using the Custom Designer is the preferred method.

- Select **DOE > Custom Design** and specify the response and factor settings as shown in **Figure 16.12**.
- Click **Interactions > 2nd** to quickly add all second order interactions.
- Specify 32 runs.
- Click **Make Design**.

Figure 16.12 Designing the Reactor Experiment



The design generated by JMP DOE and the results of this experiment are in the sample JMP data table called Reactor 32 Runs.jmp.

To analyze the reactor data, open the table (found in the Design Experiment subfolder of the sample data) called Reactor 32 Runs.JMP.

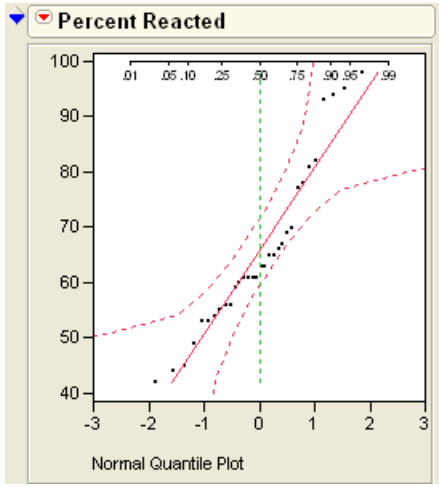
Figure 16.13 shows a partial listing of the data.

Figure 16.13 Design Table and Data for Reactor Example

	Feed Rate	Catalyst	Stir Rate	Temperature	Concentration	Percent Reacted
1	10	1	100	140	3	61
2	10	1	100	140	6	56
3	10	1	100	180	3	69
4	10	1	100	180	6	44
5	10	1	120	140	3	53
6	10	1	120	140	6	59
7	10	1	120	180	3	66
8	10	1	120	180	6	49
9	10	2	100	140	3	63
10	10	2	100	140	6	70
11	10	2	100	180	3	94
12	10	2	100	180	6	78

It is useful to begin the analysis with a quick look at the response data.

- Choose **Analyze > Distribution** to look at the distribution of the response variable Percent Reacted.
- Select the **Normal Quantile Plot** option from the popup menu on the histogram title bar to see the Normal Quantile Plot shown at right.



Since this is a screening experiment, we begin the exploration with the Screening platform.

- Select **Analyze > Modeling > Screening**.
- Assign Feed Rate, Catalyst, Stir Rate, Temperature, and Concentration to **X**, and Percent Reacted to **Y**.
- Click **OK**.

The Contrasts section of the output, **Figure 16.14**, shows that Catalyst, Temperature, Concentration, and some interactions are significant.

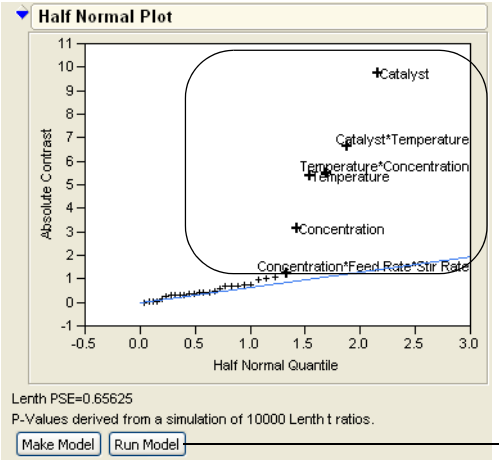
Figure 16.14 Reactor Contrasts Output

Contrasts

Term	Contrast	Lenth t-Ratio	Individual p-Value	Simultaneous p-Value
Catalyst	9.75000	14.86	<.0001*	<.0001*
Temperature	5.37500	8.19	0.0001*	0.0003*
Concentration	-3.12500	-4.76	0.0009*	0.0151*
Feed Rate	-0.68750	-1.05	0.2804	0.9998
Stir Rate	-0.31250	-0.48	0.6384	1.0000
Catalyst*Temperature	6.62500	10.10	<.0001*	0.0002*
Catalyst*Concentration	1.00000	1.52	0.1282	0.9538
Temperature*Concentration	-5.50000	-8.38	<.0001*	0.0003*
Catalyst*Feed Rate	0.68750	1.05	0.2804	0.9998
Temperature*Feed Rate	-0.43750	-0.67	0.4930	1.0000
Concentration*Feed Rate	0.06250	0.10	0.9250	1.0000
Catalyst*Stir Rate	0.43750	0.67	0.5197	1.0000
Temperature*Stir Rate	1.06250	1.62	0.1102	0.9216
Concentration*Stir Rate	0.43750	0.67	0.4930	1.0000
Feed Rate*Stir Rate	0.37500	0.57	0.5789	1.0000
Catalyst*Temperature*Concentration	-0.12500	-0.19	0.8521	1.0000
Catalyst*Temperature*Feed Rate	0.68750	1.05	0.2804	0.9998
Catalyst*Concentration*Feed Rate	-0.93750	-1.43	0.1482	0.9751
Temperature*Concentration*Feed Rate	0.31250	0.48	0.6384	1.0000
Catalyst*Temperature*Stir Rate	0.56250	0.86	0.3749	1.0000
Catalyst*Concentration*Stir Rate	0.06250	0.10	0.9250	1.0000
Temperature*Concentration*Stir Rate	0.06250	0.10	0.9250	1.0000
Catalyst*Feed Rate*Stir Rate	0.75000	1.14	0.2429	0.9990
Temperature*Feed Rate*Stir Rate	-0.37500	-0.57	0.5789	1.0000
Concentration*Feed Rate*Stir Rate	-1.25000	-1.90	0.0688	0.7557
Catalyst*Temperature*Concentration*Feed Rate	0.31250	0.48	0.6384	1.0000
Catalyst*Temperature*Concentration*Stir Rate	-0.31250	-0.48	0.6384	1.0000
Catalyst*Temperature*Feed Rate*Stir Rate	-1.57e-15	-0.00	1.0000	1.0000
Catalyst*Concentration*Feed Rate*Stir Rate	0.75000	1.14	0.2429	0.9990
Temperature*Concentration*Feed Rate*Stir Rate	0.50000	0.76	0.4293	1.0000
Catalyst*Temperature*Concentration*Feed Rate*Stir Rate	-0.25000	-0.38	0.7061	1.0000

The half-normal plot confirms this activity. Active effects separate from the diagonal line, and JMP highlights and labels them.

Figure 16.15 Reactor Half-Normal Plot



Significant Effects are Labeled.

Click to make or run model.

The two buttons below the plot let us analyze this model using the **Fit Model** platform. **Make Model** launches the Fit Model dialog with effects pre-assigned. **Run Model** does the same thing, but also runs the model and displays the report. Using the **Run Model** button is the same as using the **Make Model** button, then pressing Run Model in the resulting dialog.

- Click **Run Model**.

This creates a model containing all active effects. If there are main effects that are involved in significant interactions, they too are added.

Note: If an error message appears about a missing effect, click **Continue**.

Let's look closer at the interaction of Concentration and Temperature.

- Select **Factor Profiling > Profiler**.

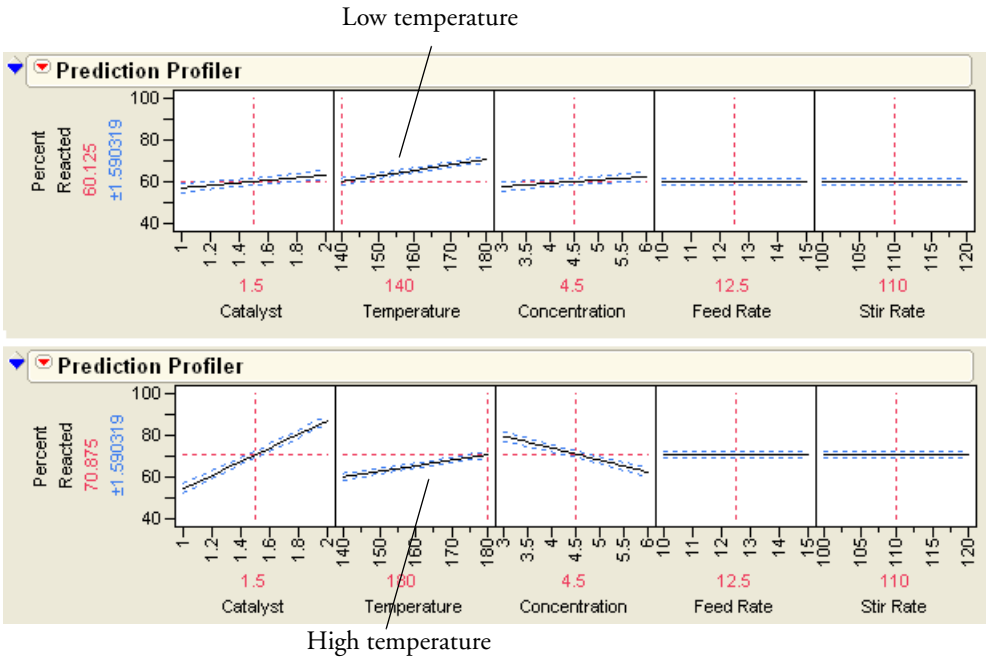
The profiler appears at the bottom of the report.

- In the profiler, click the levels of Temperature repeatedly to alternate its setting from 140 to 180.

Now watch the slope on the profile for Concentration or Catalyst. The slopes change dramatically as temperature is changed, which indicates an interaction. When there is no interaction, only the heights of the profile should change, not the slopes. Watch the slope for Catalyst, too, because it also interacts with Temperature.

The Prediction Profiler at the top of **Figure 16.16** show responses when temperature is at its low setting. The lower set of plots show what happens when temperature is at its higher setting.

Figure 16.16 Effect of Changing Temperature Levels

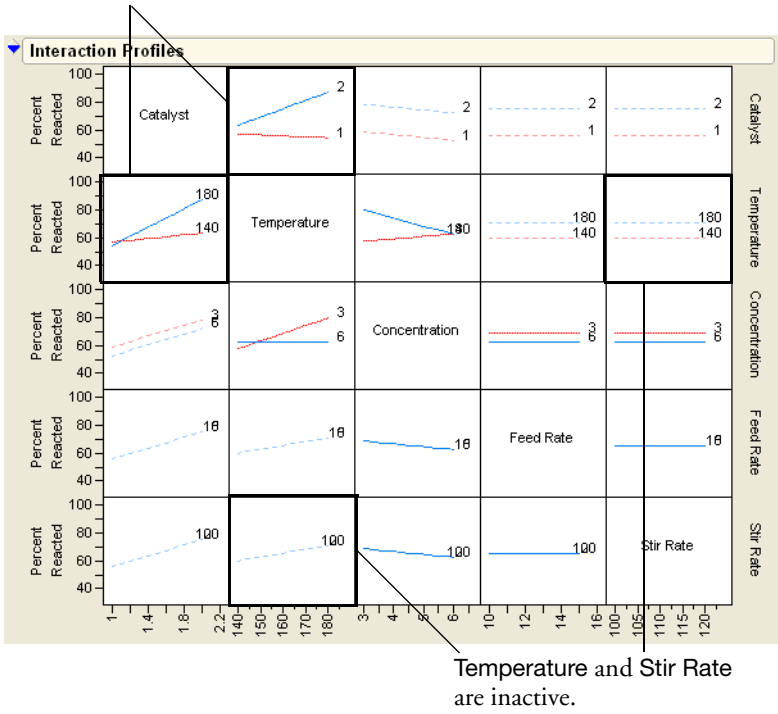


This slope change that is caused by interaction can be seen for all interactions in one picture as follows:

- Use **Factor Profiling > Interaction Plot** from the popup menu on the analysis title bar to **Figure 16.17**.

These are profile plots that show the interactions between all pairs of variables in the reactor experiment.

Figure 16.17 Interaction Plots for Five-Factor Reactor Experiment



In an interaction plot, the *y*-axes are the response. Each small plot shows the effect of two factors on the response. One factor (associated with the column of the matrix of plots) is on the *x*-axis. This factor's effect shows as the slope of the lines in the plot. The other factor becomes multiple prediction profiles (lines) as it varies from low to high. This factor shows its effect on the response as the vertical separation of the profile lines. If there is an interaction, then the slopes are different for different profile lines, like those in the Temperature by Catalyst plot.

Note: The lines of a cell in the interaction plot are dotted when there is no corresponding interaction term in the model.

The Prediction Profiler plots in **Figure 16.16** indicated that Temperature is active and Stir Rate is inactive. In the matrix of Interaction Profile plots, look at the Stir Rate column for row Temperature; Temperature causes the lines to separate, but doesn't determine the slope of the lines. Look at the plot when the factors are reversed (row Stir Rate and column Temperature); the lines are sloped, but they don't separate.

Recall that Temperature interacted with Catalyst and Concentration. This is evident by the differing slopes showing in the Temperature by Catalyst and the Temperature by Concentration Interaction Profile plots.

Response Surface Designs

Response surface designs are useful for modeling a curved (quadratic) surface to continuous factors. If a minimum or maximum response exists inside the factor region, a response surface model can pinpoint it. Three distinct values for each factor are necessary to fit a quadratic function, so the standard two-level designs cannot fit curved surfaces.

The Experiment

Suppose the objective of an industrial experiment is to minimize the unpleasant odor of a chemical. It is known that the odor varies with temperature (temp), gas-liquid ratio (gl ratio), and packing height (ht). The experimenter wants to collect data over a wide range of values for these variables to see if a response surface can identify values that give a minimum odor (adapted from John, 1971).

Response Surface Designs in JMP

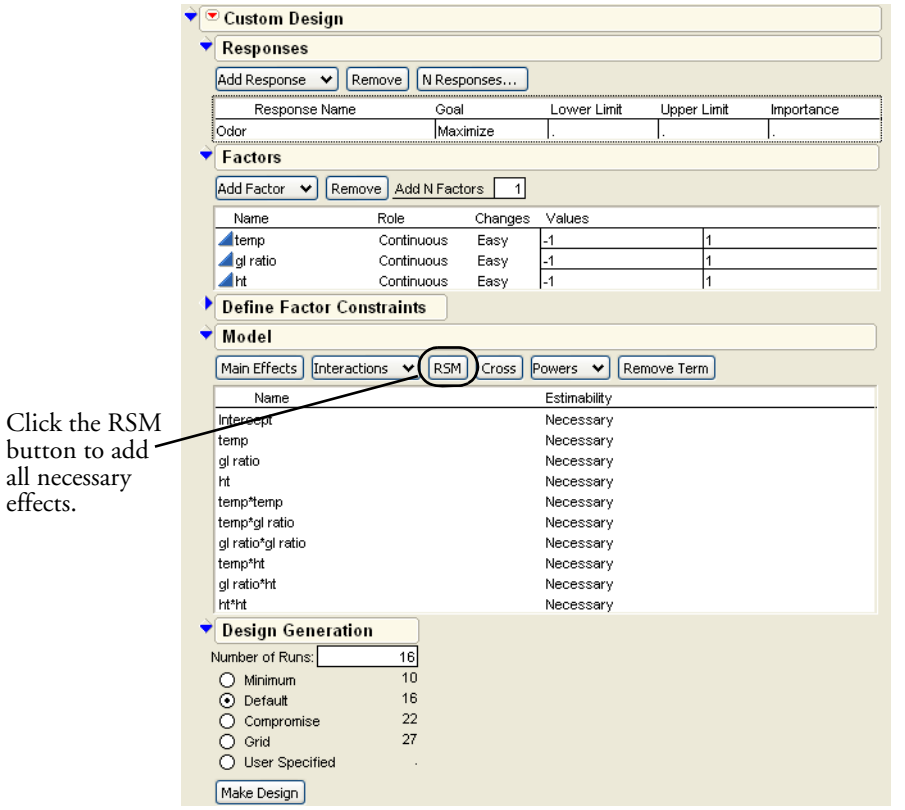
To generate a response surface design,

- Choose **DOE > Custom Design**.
- In the resulting dialog, enter the response and three factors (shown in **Figure 16.18**).
- Click the RSM button to add the interaction and power terms that are necessary to model the quadratic surface.
- Select 16 runs.
- Choose **Make Design.**

When the design appears,

- Choose **Sort Right to Left** from the **Run Order** drop-down list.
- Click **Make Table**.

Figure 16.18 Design Dialog to Specify Factors



The table appears with missing values for the Odor response value.

- Enter the response values shown in as in the Odor column.

Figure 16.19 Design Table for a Central Composite Response Surface Design

Custom Design

Design	Custom Design
Criterion	I Optimal
Model	

Columns (4/0)

- temp *
- gl ratio *
- ht *
- Odor *

Rows

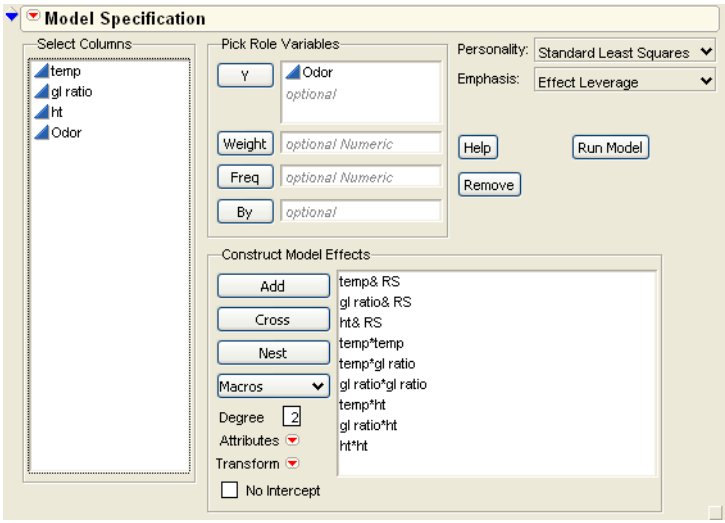
All rows	18
Selected	0
Excluded	0
Hidden	0
Labelled	0

	temp	gl ratio	ht	Odor
1	-1	-1	-1	150
2	1	-1	-1	99
3	0	0	-1	35
4	-1	1	-1	84
5	1	1	-1	87
6	0	-1	0	77
7	-1	0	0	43
8	0	0	0	5
9	0	0	0	5
10	0	0	0	15
11	0	0	0	3
12	1	0	0	42
13	0	1	0	28
14	-1	-1	1	109
15	1	-1	1	83
16	0	0	1	12
17	-1	1	1	63
18	1	1	1	61

Like all JMP tables generated by the DOE facility, the Table Property called **Model** contains the JSL script that generates the completed Fit Model dialog for analyzing the design after data are collected.

Select **Analyze > Fit Model** to see the completed dialog in **Figure 16.20**.

Figure 16.20 Fit Model Dialog



The effects appear in the Model Effects list as shown in **Figure 16.20**, with the &RS notation on the main effects (temp, gl ratio, and ht). This notation indicates that these terms are to be subjected to a curvature analysis.

Click **Run Model** on the Fit Model dialog to see the analysis.

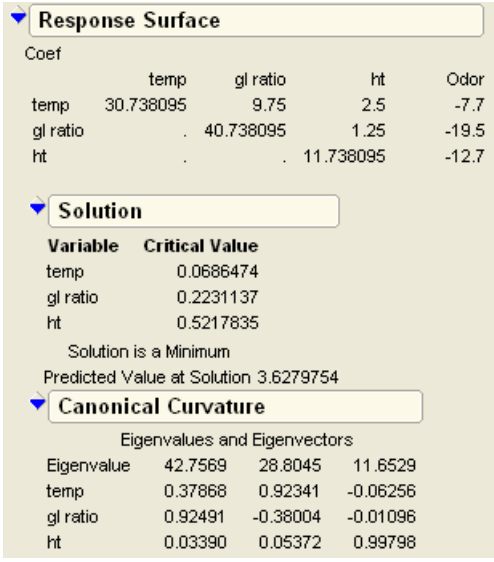
The standard least squares analysis tables appear with an additional report outline level called Response Surface.

Open the Response Surface outline level to see the tables shown in **Figure 16.21**.

- The first table is a summary of the parameter estimates.
- The Solution table lists the critical values of the surface and tells the kind of solution (maximum, minimum, or saddle point). The critical values are where the surface has a slope of zero, which could be an optimum depending on the curvature.
- The Canonical Curvature table shows eigenvalues and eigenvectors of the effects. The eigenvectors are the directions of the principal curvatures. The eigenvalue associated with each direction tells whether it is decreasing slope, like a maximum (negative eigenvalue), or increasing slope, like a minimum (positive eigenvalue).

The Solution table in this example shows the solution to be a minimum.

Figure 16.21 Response Surface Model and Analysis Results



Response Surface

Coef

	temp	gl ratio	ht	Odor
temp	30.738095	9.75	2.5	-7.7
gl ratio	.	40.738095	1.25	-19.5
ht	.	.	11.738095	-12.7

Solution

Variable	Critical Value
temp	0.0686474
gl ratio	0.2231137
ht	0.5217835

Solution is a Minimum

Predicted Value at Solution 3.6279754

Canonical Curvature

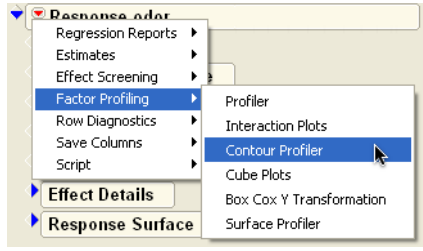
Eigenvalues and Eigenvectors

Eigenvalue	42.7569	28.8045	11.6529
temp	0.37868	0.92341	-0.06256
gl ratio	0.92491	-0.38004	-0.01096
ht	0.03390	0.05372	0.99798

Plotting Surface Effects

If there are more than two factors, you can see a contour plot of any two factors at intervals of a third factor by using the Contour Profiler.

- Choose **Factor Profiling > Contour Profiler** found in the popup menu on the report title, as shown here.



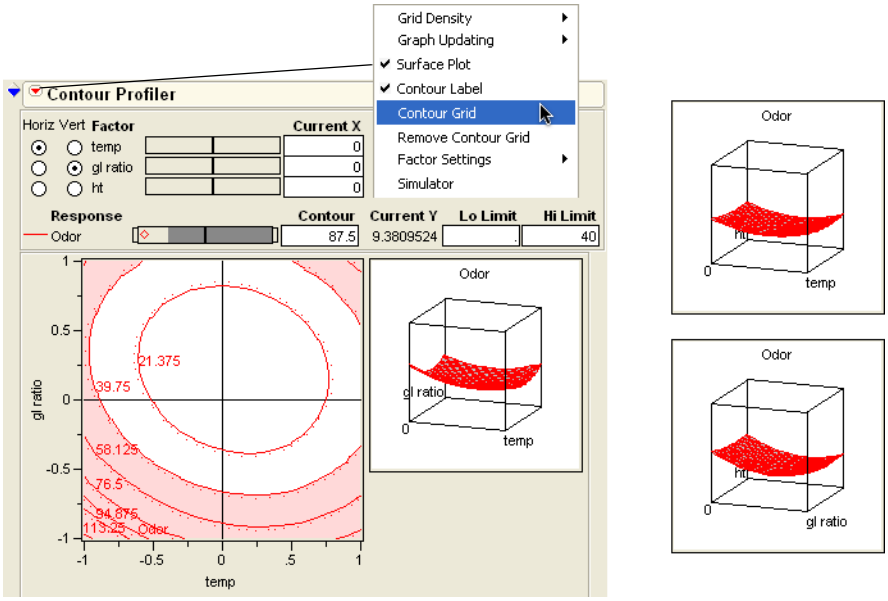
The **Contour Profiler** displays a panel that lets you use interactive sliders to vary one factor's values and observe the effect on the other two factors. You can also vary one factor and see the effect on a mesh plot of the other two factors. **Figure 16.22** shows contours of ht as a function of temp and gl ratio. The mesh plots on the right in **Figure 16.22** show the response surface in the combinations of the three factors taken two at a time.

- Enter 40 as the Hi Limit specification.

Entering a value defines and shades the region of acceptable values for the three variables.

- Optionally, use the **Contour Grid** option in the Contour Profiler popup to add grid lines with specified values to the contour plot.

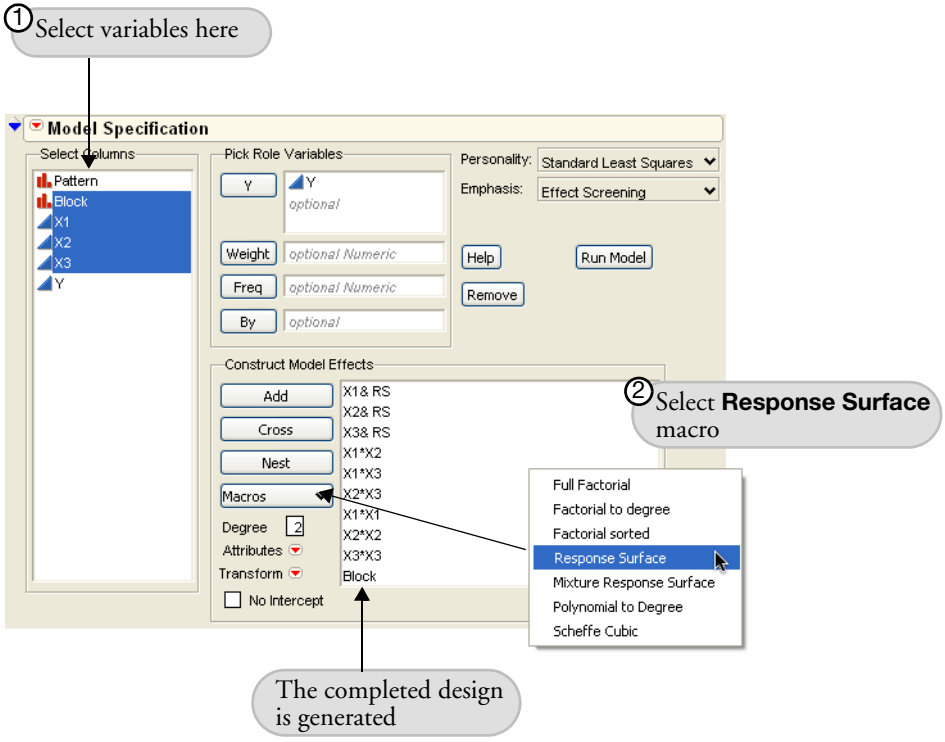
Figure 16.22 Contour Profiler with Contours and Mesh Plot



Designating RSM Designs Manually

JMP completes the Fit Model dialog automatically when you build the design using JMP tools, but you can generate a response surface analysis for any effects by selecting them in the Select Columns list and choosing **Response Surface** from the effect **Macros**, as illustrated in **Figure 16.23**.

Figure 16.23 Fit Model Dialog for Response Surface Design



The Prediction Variance Profiler

Although most design types have at least two factors, the following examples have a single continuous factor and compare designs for quadratic and cubic models. The purpose of these examples is to introduce the prediction variance profile plot.

A Quadratic Model

Follow these steps to create a simple quadratic model with a single continuous factor.

- Select **DOE > Custom Design.** Add one continuous factor and click **Continue**.
- Select **2nd** from the Powers popup menu in the Model panel to create a quadratic term.
- Use the default number of runs, 6, and click **Make Design**.

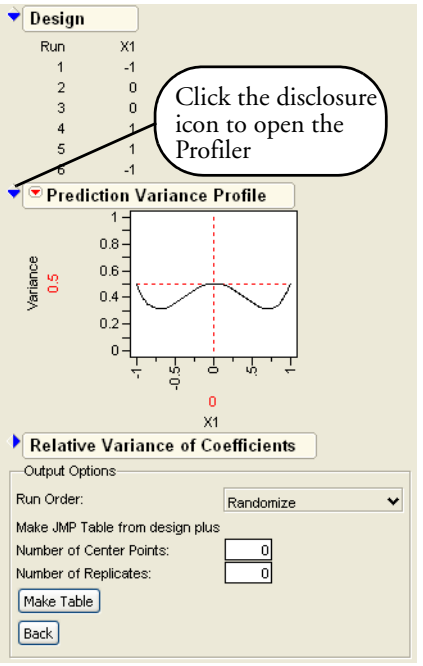
When the design appears,

- Open the **Prediction Variance Profile**.

For continuous factors, the initial setting is at the mid-range of the factor values. For categorical factors, the initial setting is the first level. If the design model is quadratic, then the prediction variance function is quartic. The three design points are –1, 0, and 1. The prediction variance profile shows that the variance is a maximum at each of these points, on the interval –1 to 1. The *y*-axis is the relative variance of prediction of the expected value of the response.

This prediction variance is a relative variance. The actual variance would be this value times the error variance.

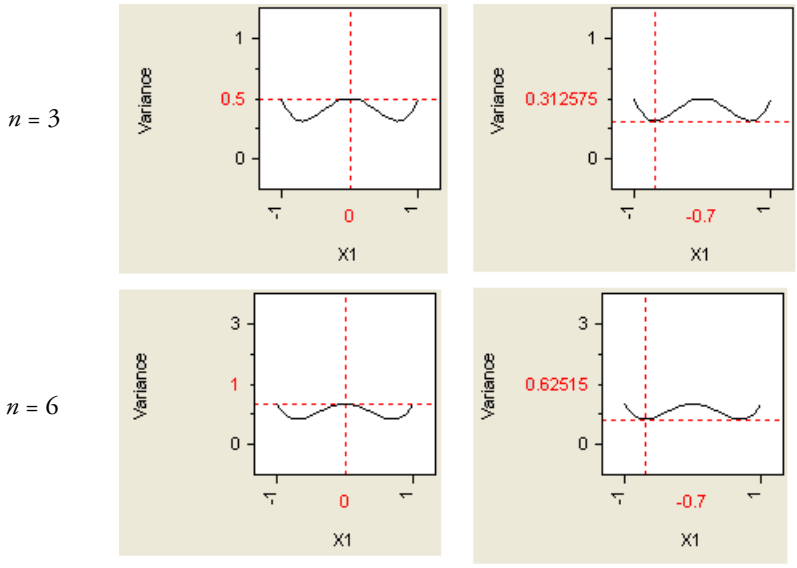
When you choose a sample size, you are deciding how much variance in the expected response you are willing to tolerate. As the number of runs increases, the prediction variance curve decreases.



To compare profile plots, click **Back** and choose **Minimum** in the Design Generation panel, which gives a sample size of 3.

This produces a curve that has the same shape as the previous plot, but the maxima are at 1 instead of 0.5. **Figure 16.24** compares plots for sample size 6 and sample size 3 for this quadratic model example. You can see the prediction variance increase as the sample size decreases. These profiles are for middle variance and lowest variance, for sample sizes 6 (bottom charts) and 3 (top charts).

Figure 16.24 Comparison of Prediction Variance Profiles

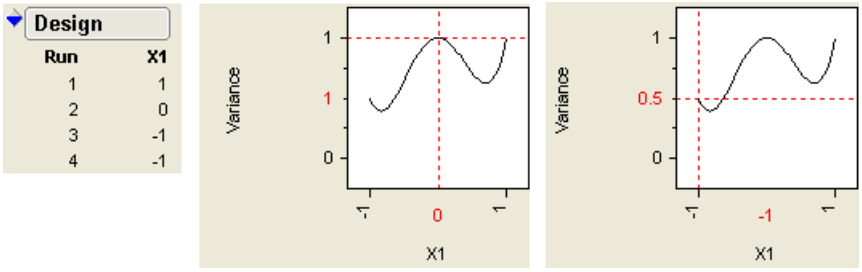


Note: You can double-click on the axis to set a factor level precisely.

For a final look at the Prediction Variance Profile for the quadratic model, click **Back** and enter a sample size of 4 in the Design Generation panel and click **Make Design**.

The sample size of 4 adds a point at –1 (**Figure 16.25**). Therefore, the variance of prediction at –1 is lower (half the value) than the other sample points. The symmetry of the plot is related to the balance of the factor settings. When the design points are balanced, the plot is symmetric, like those in **Figure 16.24**. When the design is unbalanced, the prediction plot is not symmetric, as shown in **Figure 16.25**.

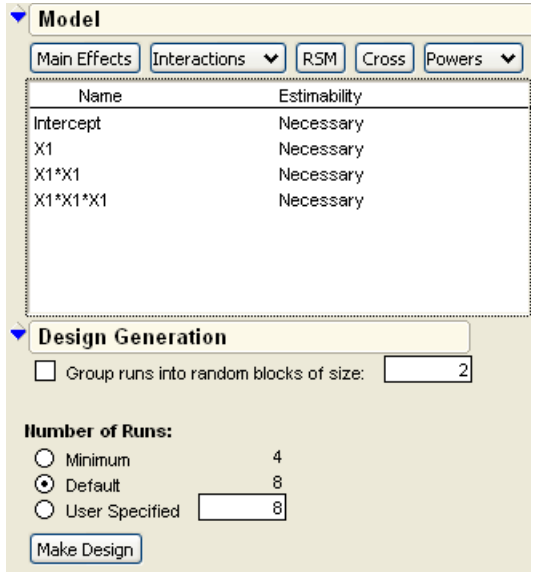
Figure 16.25 Sample Size of Four for the One-Factor Quadratic Model



A Cubic Model

The runs in the quadratic model are equally spaced. This is not true for the single-factor cubic model shown in this section. To create a one-factor cubic model, follow the same steps as in “A Quadratic Model” on page 442.

- In addition, add a cubic term to the model with the **Powers** popup menu.



Use the Default number of runs in the Design Generation panel.

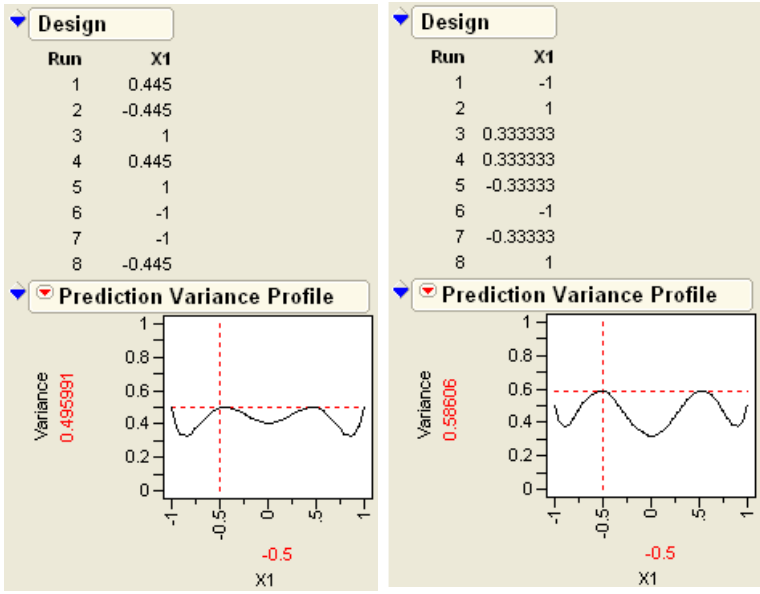
- Click **Make Design** to continue.
- Open the Prediction Variance Profile Plot to see the Prediction Variance Profile and its associated design shown in **Figure 16.26**.

The cubic model has a variance profile that is a 6th degree polynomial.

Note that the points are not equally-spaced in *X*. This design has a better prediction variance profile than the equally-spaced design with the same number of runs, though this result might seem counter-intuitive.

To produce the equally spaced design used here, make a new data table with a column named X1 containing the values shown in **Figure 16.26**. Then, select **DOE > Custom Design** and add a covariate factor. Select the newly made column. Then, add the X1*X1 and X1*X1*X1 terms using the **Powers** button. Using 8 runs, make the design to reveal the Prediction Variance Profiler shown in **Figure 16.26**.

Figure 16.26 Comparison of Prediction Variance Profiles



The design on the left is an optimal design from the Custom designer. The design on the right is an equally-spaced design. At values of X1 = -0.5, the optimal design is better due to its lower prediction variance.

Design Issues

So far, the discussion has been on the particulars of how to get certain designs. But why are these designs good? Here is some general advice, though all these points have limitations or drawbacks as well.

Multifactor designs yield information faster and better. Experiments that examine only one factor and therefore only ask one question at a time are inefficient.

For example, suppose that you have one factor with two levels, and you need 16 runs to see a difference that meets your requirements. Increasing to seven factors (and running each as a separate experiment) would take $7 \times 16 = 104$ runs. However, with a multifactor experimental design, all seven factors can be estimated in those 16 runs, and have the same standard error of the difference, saving you 7/8's of your runs.

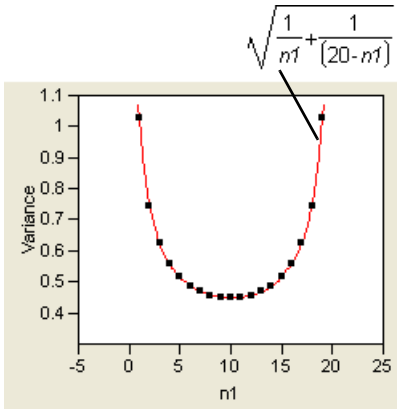
But that is not all. With only 16 runs you can still check for two-way interactions among the seven factors. The effect of one factor may depend on the settings of another factor. This cannot be discovered with experiments that vary only one factor at a time.

Experimentation allows you to build a comprehensive model and understand more of the dynamics of the response's behavior.

Balancing and orthogonality. Let's illustrate the balancing issue: In the simplest case of two groups, the standard error of the difference between the two means is proportional to the square root of the sum of the reciprocals of the sample size, as shown in the formula to the right.

$$\sqrt{\frac{1}{n1}+\frac{1}{n2}}$$

The plot shown to the right is a graph of this formula as a function of *n1*, where *n1* + *n2* is kept constant at 20. You can see that the smallest variance occurs when the design is balanced; where *n1* = *n2* = 10.



You can clearly see that if the design is balanced, you minimize the standard errors of differences, and thus maximize the power of the tests. In practice, a slight imbalance is still reasonable.

Do as many runs as you can afford. The more information you have, the more you know. You should always try to perform an experiment with as many runs as possible. Since experimental runs are costly, you have to balance the limiting factor of the cost of trial runs with the amount of information you expect to get from the experiment.

The literature of experimental design emphasizes orthogonal designs. They are great designs, but come only in certain sizes (*e.g.* powers of 2). Suppose you have a perfectly designed (orthogonal) experiment with 16 runs. Given the opportunity to do another virtually free run, therefore using 17 runs, but making the design unbalanced and non-orthogonal, should you accept the run?

Yes. *Definitely.* Adding runs never subtracts information.

Take values at the extremes of the factor range.

This has two advantages and one disadvantage:

- It maximizes the power. By separating the values, you are increasing the parameter for the difference, making it easy to distinguish it from zero.
- It keeps the applicable range of the experiment wide. When you use the prediction formula outside the range of the data, its variance is high, and it is unreliable for other reasons.
- The disadvantage is that the true response might not be approximated as well by a simple model over a larger range.

Let's illustrate why choosing a wide range of values is good.

The data table shown to the right is a simple arrangement of 12 points. There are 4 points at extreme values of X, 4 points at moderate range, and 4 points at the center. Examine subsets of the 12 points, to see which ones perform better:

	X	Y
1	-4	-1
2	-4	-3
3	-2	-2
4	-2	0
5	0	-1
6	0	-1
7	0	1
8	0	1
9	2	0
10	2	2
11	4	1
12	4	3

- Exclude the extremes of ±4.
- Exclude the intermediate points of ±2.
- Exclude the center points.

To exclude points,

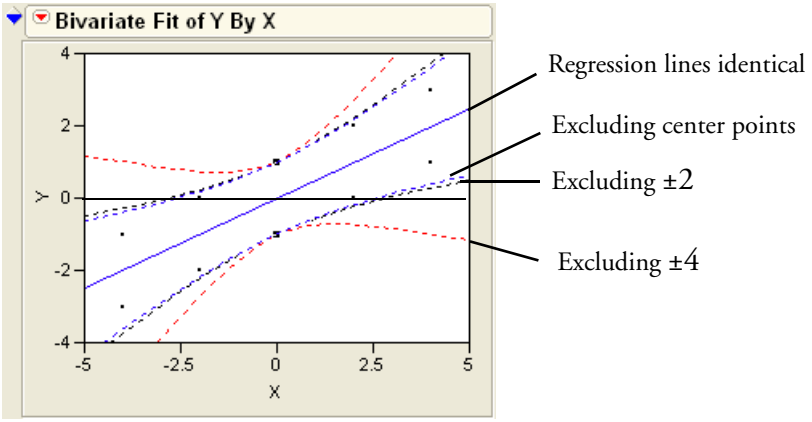
- Highlight the rows to exclude in the data table and select **Rows > Exclude**.

Figure 16.27 compares the confidence interval of the regression line in the three subsets. Each situation has the same number of observations, the same error variance, and the same residuals. The only difference is in the spacing of the X values, but the subsamples show the following differences:

- The fit excluding the points at the extremes of ±4 has the widely flared confidence curves. The confidence curves cross the horizontal line at the mean, indicating that it is significant at 0.05.
- The fit excluding the intermediate range points of ±2 has much less flared confidence curves. The confidence curves do not cross the horizontal line at the mean, indicating that it is not significant at 0.05.
- The fit excluding the center points was the best, though not much different than the previous case.
- All the fits had the same size confidence limits in the middle.

For given sample size, the better design is the one that spaces out the points farther.

Figure 16.27 Confidence Interval of the Regression Line

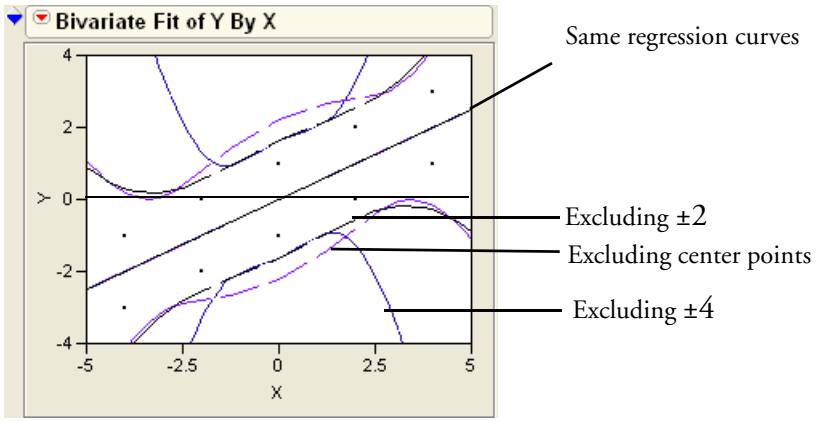


If you can't afford to put values at the absolute extremes, try to come close by designing the experiment so that many subsets of variables cover all combinations of levels.

Put a few runs in the center too, if this makes sense. There are two reasons to do this:

- This allows for estimation of curvature.
- If you have several runs at the same point, you can estimate pure error and do a lack-of-fit test.

Using the same data as before, in the same three scenarios, look at a quadratic curve instead of a line. Note the confidence limits for the curve that excluded the center points have a much wider interval at the middle. Curvature is permitted by the model but is not well supported by the design.



Center points are also usually replicated points that allow for an independent estimate of pure error, which can be used in a lack-of-fit test. For the same number of points, the design with center points can detect lack-of-fit and nonlinearities.

Randomize the assignment of runs. This is to neutralize any inadvertent effects that may be present in a given sequence.

Routine Screening Examples

This section gives short examples showing how to use the Custom Designer to generate specific types of screening designs.

Main Effects Only

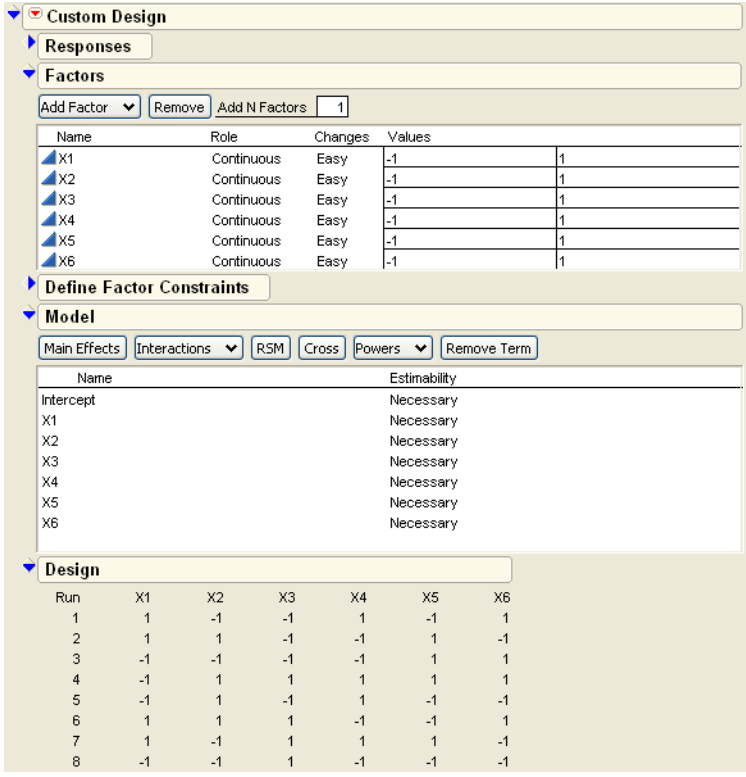
- Select **DOE > Custom Design**.
- Enter the number of factors you want (six for this example) into the Factors panel and click **Continue** as shown in **Figure 16.28**.

This example uses six factors. Because there are no complex terms in the model, no further action is needed in the Model panel. The default number of runs (8) is correct for the main-effects-only model.

The result is a resolution-3 screening design. All main effects are estimable but are confounded with two-factor interactions.

- Click **Make Design** to see the Factor Design table in **Figure 16.28.**

Figure 16.28 A Main Effects Only Screening Design



All Two-Factor Interactions Involving Only One Factor

Sometimes there is reason to believe that some two-factor interactions may be important. The following example illustrates adding all the two-factor interactions involving a single factor. The example has five continuous factors.

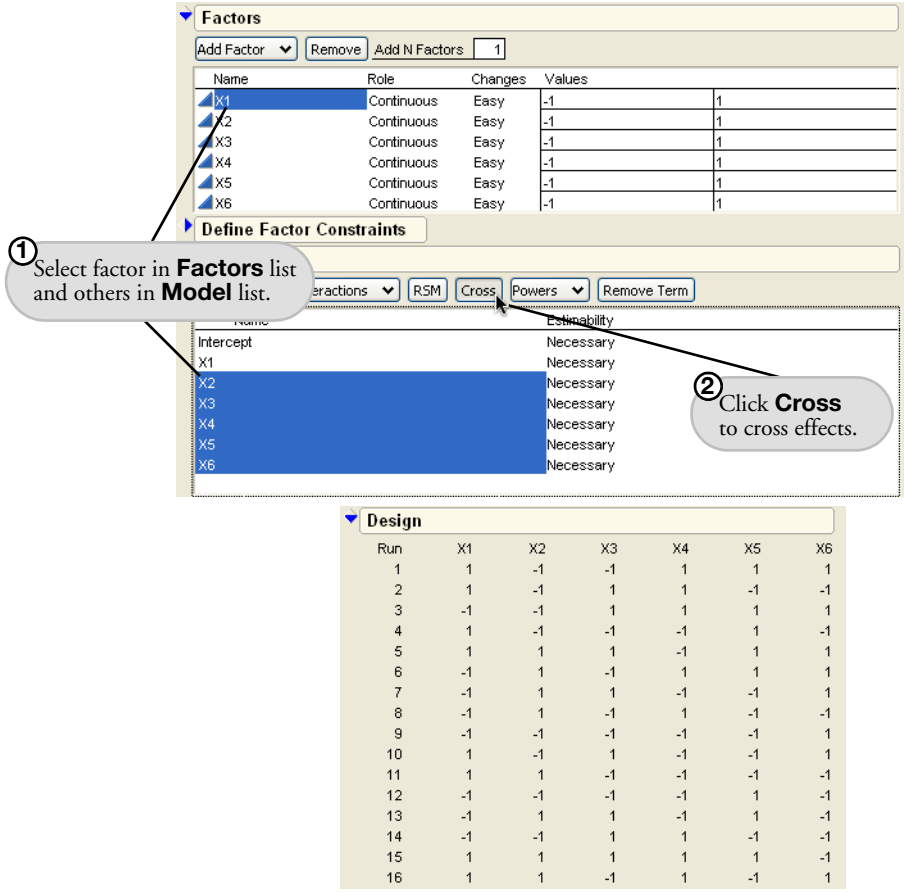
This design is a resolution-4 design equivalent to folding over on the factor for which all two-factor interactions are estimable.

To get a specific set of crossed factors (rather than all interactions or response surface terms),

- Select the factor to cross (X1, for example) in the Factors table.
- Select the other factors in the Model Table and click **Cross** to see the interactions in the model table, as shown in **Figure 16.29**.

The default sample size for designs with only two-level factors is the smallest power of two that is larger than the number of terms in the design model. For example, in **Figure 16.29**, there are 9 terms in the model, and 2^4=16 is the smallest power of two that is greater than 9.

Figure 16.29 Two-Factor Interactions That Involve Only One of the Factors



All Two-Factor Interactions

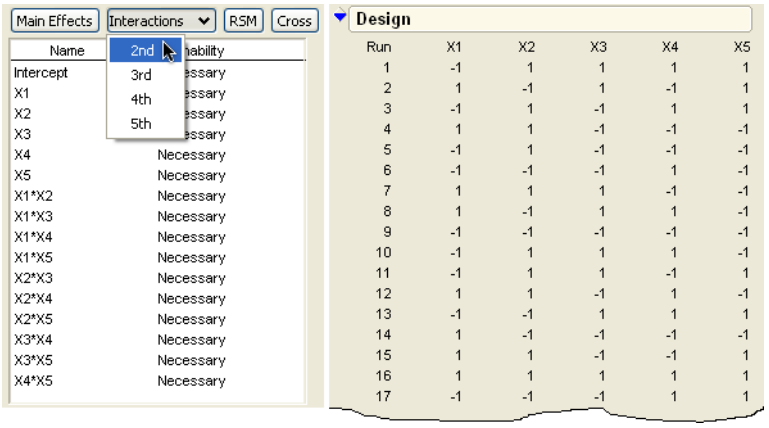
In situations where there are few factors and experimental runs are cheap, you can run screening experiments that allow for estimating all the two-factor interactions. The Custom Design interface makes this simple (see **Figure 16.30**).

- Enter the number of factors (5 in this example).
- Click **Continue** and choose **2nd** from the **Interactions** popup in the Model outline.

Click **Make Design**.

Figure 16.30 shows a partial listing of the two-factor design with all interactions. The default design has the minimum-power-of-two sample size consistent with fitting the model.

Figure 16.30 All Two-Factor Interactions



Design Strategies Glossary

Significance. You want a statistical hypothesis test to show significance at meaningful levels. For example, you want to show statistically that a new drug is safe and effective as a regulatory agency. Perhaps you want to show a significant result to publish in a scientific journal. The *p*-value shows the significance. The *p*-value is a function of both the estimated effect size and the estimated variance of the experimental error. It shows how unlikely so extreme a statistic would be due only to random error.

Effect Precision. Effect precision is the difference in the expected response attributed to a change in a factor. This precision is the factor's coefficient, or the estimated parameter in the prediction expression.

Effect Size. The standard error of the effect size measures how well the experimental design estimates the response. This involves the experimental error variance, but it can be determined relative to this variance before the experiment is performed.

Screening Designs. These designs sort through many factors to find those that are most important, *i.e.* those that the response is most sensitive to. Such factors are the vital few that account for most of the variation in the response.

Prediction Designs. Rather than looking at how factors contribute to a response, you instead want to develop the most accurate prediction.

Optimization. You want to find the factor settings that optimize a criterion. For example, you may want factor values that optimize quality, yield, or cost while minimizing bad side effects.

Robustness. You want to determine operational settings that are least affected by variation in uncontrollable factors.

Formulation. In addition to these goals, sometimes there is a constraint. For example, factors that are mixture proportions must to sum to 1.

Coding. For continuous factors, coding transforms the data from the natural scale to a -1 to 1 scale, so that effect sizes are relative to the range. For categorical factors, coding creates design columns that are numeric indicator columns for the categories.

Balanced. A design is balanced when there are the same number of runs for each level of a factor, and for various combinations of levels of factors.

Factorial. All combinations of levels are run. If there are k factors and each has two levels, the number of factorial runs is 2^k.

Fractional Factorial. A fraction of a factorial design where some factors are formed by interactions of other factors. They are frequently used for screening designs. A fractional factorial design also has a sample size that is a power of two. If k is the number of factors, the number of runs is 2^{k-p} where $p < k$. Fractional factorial designs are orthogonal.

Plackett-Burman. *Plackett-Burman* designs are an alternative to fractional factorials for screening. Since there are no two-level fractional factorial designs with sample sizes between 16 and 32 runs. a useful characteristic of these designs is that the sample size is a multiple of four rather than a power of two. However, there are 20-run, 24-run, and 28-run Plackett-Burman designs.

The main effects are orthogonal and two-factor interactions are only partially confounded with main effects. This is different from resolution-3 fractional factorial where two-factor interactions are indistinguishable from main effects.

In cases of *effect sparsity* (where most effects are assumed to have no measurable effect on the response), a stepwise regression approach can allow for removing some insignificant main effects while adding highly significant and only somewhat correlated two-factor interactions.

Resolution. The *resolution* number is a way to describe the degree of confounding, usually focusing on main effects and two-way interactions. Higher-order interactions are assumed to be zero.

In resolution-3 designs, main effects are not confounded with other main effects, but two-factor interactions are confounded with main effects. Only main effects are included in the model. For the main effects to be meaningful, two-factor interactions are assumed to be zero or negligible.

In resolution-4 designs, main effects are not confounded with each other or with two-factor interactions, but some two-factor interactions can be confounded with each other. Some two-factor interactions can be modeled without being confounded. Other two-factor interactions can be modeled with the understanding that they are confounded with two-factor interactions included in the model. Three-factor interactions are assumed to be negligible.

In resolution-5 designs, there is no confounding between main effects, between two-factor interactions, or between main effects and two-factor interactions. That is, all two-factor interactions are estimable.

Minimum-Aberration. A *minimum-aberration design* has a minimum number of confoundings for a given resolution.

Response Surface. Response Surface Methodology (RSM) is a technique invented that finds the optimal response within specified ranges of the factors. These designs can fit a second-order prediction equation for the response. The quadratic terms in these equations model the curvature in the true response function. If a maximum or minimum exists inside the factor region, RSM can find it.

In industrial applications, RSM designs usually involve a small number of factors, because the required number of runs increases dramatically with the number of factors.

Mixture Design. Any design where some of the factors are mixture factors. Mixture factors express percentages of a compound and must therefore sum to one. Classical mixture designs include the *simplex centroid*, *simplex lattice*, and *ABCD*.

Space-Filling. A design that seeks to distribute points evenly throughout a factor space.

Taguchi. The goal of the Taguchi Method is to find control factor settings that generate acceptable responses despite natural environmental and process variability. In each experiment, Taguchi's design approach employs two designs called the *inner* and *outer* array. The Taguchi experiment is the cross product of these two arrays. The control factors, used to tweak the process, form the inner array. The noise factors, associated with process or environmental

variability, form the outer array. Taguchi's signal-to-noise ratios are functions of the observed responses over an outer array. The Taguchi designer in JMP supports all these features of the Taguchi method. The inner and outer array design lists use the traditional Taguchi orthogonal arrays such as L4, L8, L16, and so forth.

Folding. If an effect is confounded with another, you can add runs to the experiment and flip certain high/low values of one or more factors to remove their confounding.

Augmentation. If you want to estimate more effects, or learn more about the effects you have already examined, adding runs to the experiment in an optimal way assures that you learn the most from the added runs.

***D*-Optimality.** Designs that are *D*-Optimal maximize a criterion so that you learn the most about the parameters (they minimize the generalized variance of the estimates). This is an excellent all-purpose criterion, especially in screening design situations.

***I*-Optimality.** Designs that are *I*-Optimal maximize a criterion so that the model predicts best over the region of interest (they minimize the average variance of prediction over the experimental region). This is a very good criterion for response surface optimization situations.

17 Exploratory Modeling

Overview

Exploratory modeling (sometimes known as *data mining*) is the process of exploring large amounts of data, usually using an automated method, to find patterns and discoveries. JMP has two platforms especially designed for exploratory modeling: the Partition platform and the Neural Net platform.

The Partition platforms recursively partitions data, automatically splitting the data at optimum points. The result is a decision tree that classifies each observation into a group. The classic example is turning a table of symptoms and diagnoses of a certain illness into a hierarchy of assessments to be evaluated on new patients.

The Neural Net platform implements a standard type of neural network. Neural nets are used to predict one or more response variables from a flexible network of functions of input variables. They can be very good predictors, and are useful when the underlying functional form of the response surface is not important.

The Partition Platform

The Partition platform is used to recursively partition a data set in ways similar to CART™, CHAID™, and C4.5. The technique is often taught as a data mining technique, because

- It is good for exploring relationships without having a good prior model.
- It handles large problems easily.
- The results are very interpretable.

The factor columns (*X*'s) can be either continuous or categorical. If an *X* is continuous, then the splits (partitions) are created by a *cutting value*. The sample is divided into values below and above this cutting value. If the *X* is categorical, then the sample is divided into two groups of levels.

The response column (*Y*) can also be either continuous or categorical. If *Y* is continuous, then the platform fits means, and creates splits which most significantly separate the means by examining the sums of squares due to the means differences. If *Y* is categorical, then the response rates (the estimated probability for each response level) become the fitted value, and the most significant split is determined by the largest likelihood-ratio chi-square statistic. In either case, the split is chosen to maximize the difference in the responses between the two branches of the split.

The Partition platform displays slightly different outputs, depending on whether the *Y*-variables in the model are continuous or categorical. In **Figure 17.1**, each point represents a response from the category and partition it is in. The *y*-position is random within the *y* partition, and the *x*-position is a random permutation so the points are in the same rectangle but at different positions in successive analyses. **Figure 17.2** shows the corresponding case for a continuous response.

Figure 17.1 Output with Categorical Response

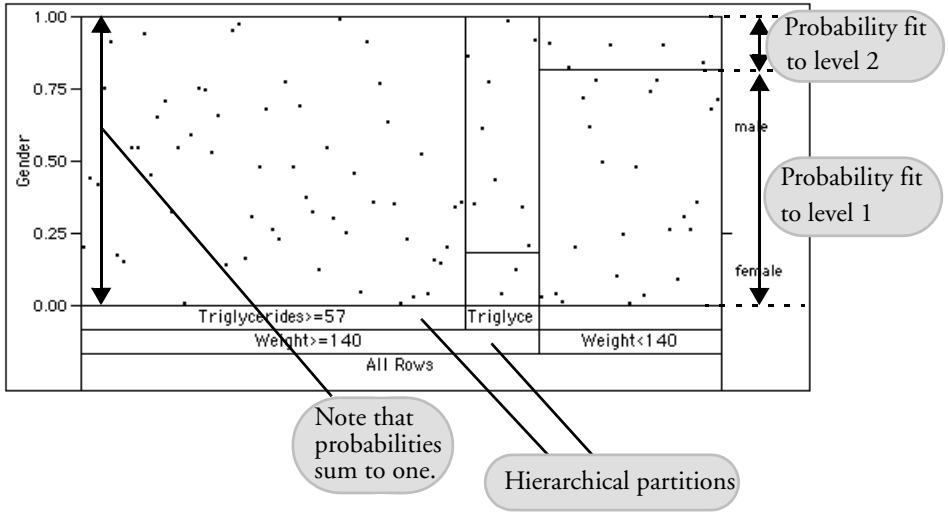
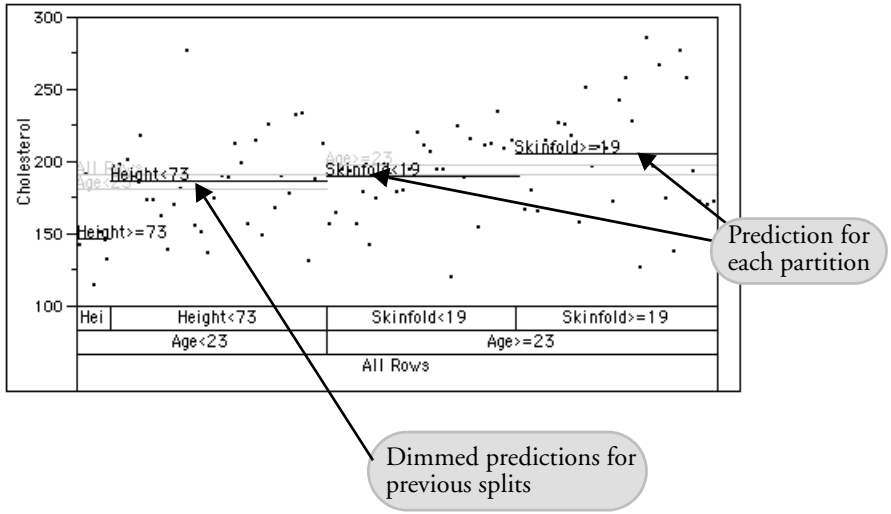


Figure 17.2 Output for Continuous Responses



Modeling with Recursive Trees

As an example of a typical analysis, open the Lipid Data.jmp data table. This data contains results from blood tests, physical measurements, and medical history for 95 subjects.

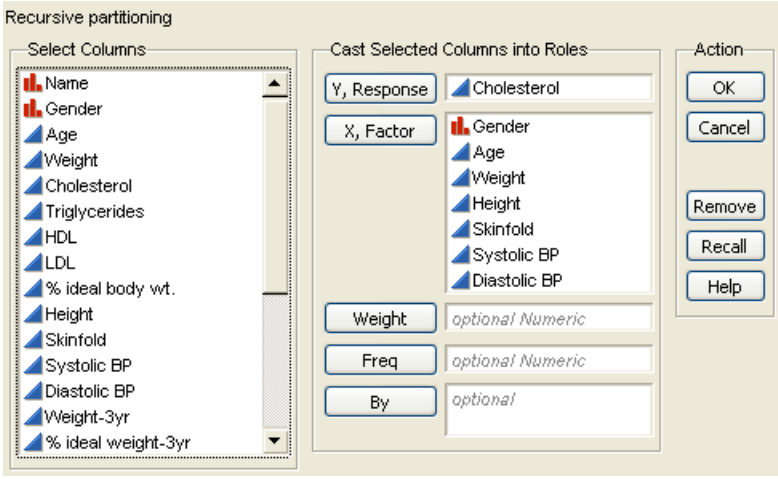
Cholesterol tests are invasive (requiring the extraction of blood) and require laboratory procedures to obtain their results. Suppose these researchers are interested in using non-

invasive, external measurements and information from questionnaires to determine which patients are likely to have high cholesterol levels. Specifically, they want to predict the values stored in the Cholesterol column with information found in the Gender, Age, Weight, Skinfold, Systolic BP, and Diastolic BP columns.

To begin the analysis,

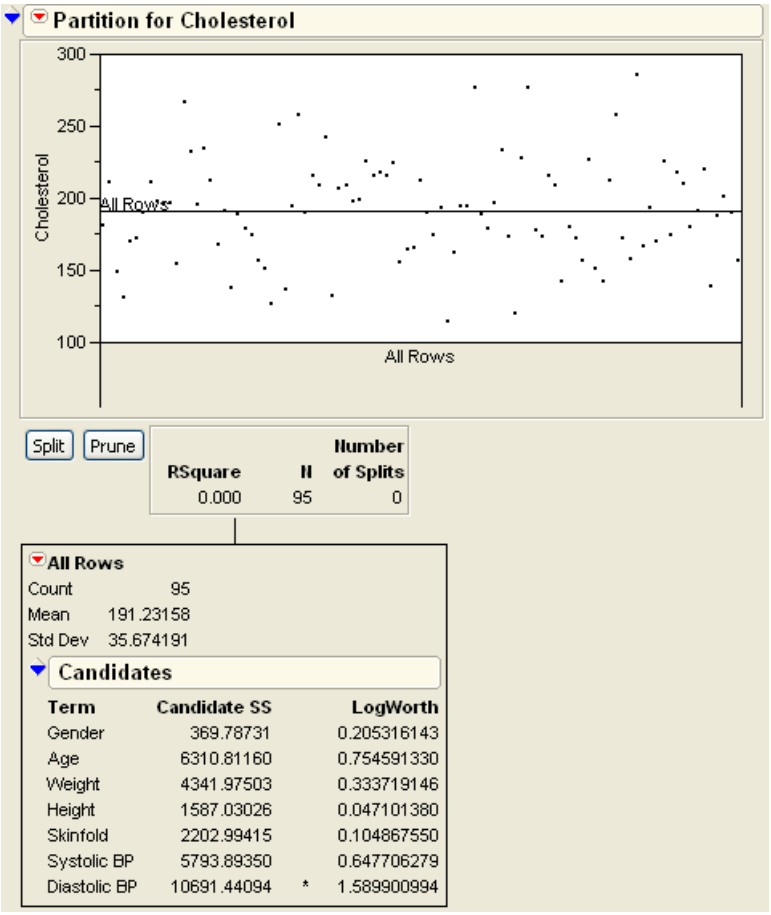
- Select **Analyze > Modeling > Partition**.
- Assign the variables as shown in **Figure 17.3**.
- Click **OK**.

Figure 17.3 Partition Dialog



The initial Partition report appears, as in **Figure 17.4**. By default, the **Candidates** node of the report is closed, but is opened here for illustration. Note that no partitioning has happened yet—all of the data are placed in a single group whose estimate is the mean cholesterol value. In order to begin the partitioning process, you must interactively request splits.

Figure 17.4 Initial Lipid Partition Report

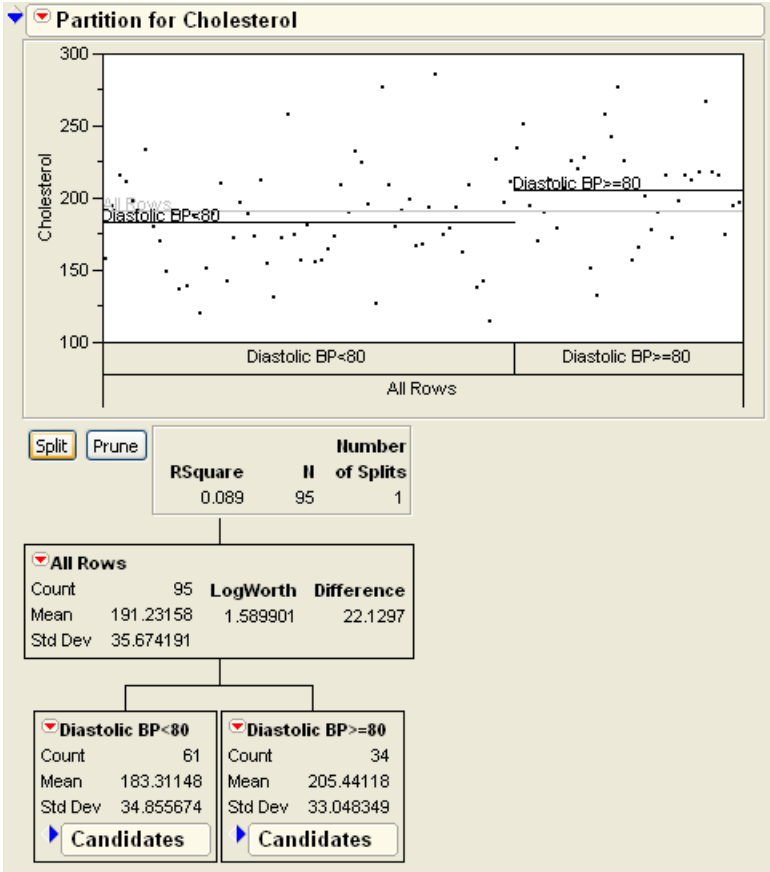


To determine the optimum split, each *x*-value is considered. The one that results in the highest reduction in total sum of squares is the optimum split, and is used to create a new branch of the tree. In this example, a split using Diastolic BP results in a reduction of 10691.44094 in the total SS, so it is used as the splitting variable.

Click the **Split** button to cause the split to take place.

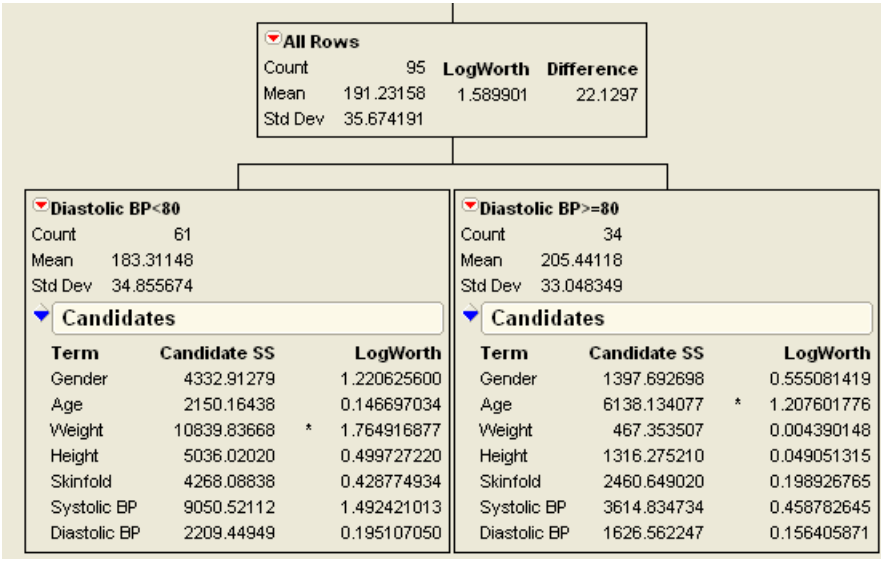
The resulting report is shown in **Figure 17.5**. As expected, the Diastolic BP variable is involved, splitting at the value 80. People with a diastolic blood pressure less than 80 tend to have lower cholesterol (in fact, a mean of 183.3) than those with blood pressure above 80 (whose mean is 205.4).

Figure 17.5 First Split of Lipid Data



An examination of the candidates report (**Figure 17.6**) shows the possibilities for the second split. Under the **Weight** leaf, a split in the Weight variable would produce a 10839.83-unit reduction in the sum of squares. Under the **Diastolic BP>80** leaf, a split in the Age variable would produce a 6138.13-unit reduction. Therefore, pressing the **Split** button splits under the **Weight** leaf.

Figure 17.6 Candidates for Second Split



Press the **Split** button to conduct the second split.

The resulting report is shown in **Figure 17.7**, with its corresponding tree shown in **Figure 17.8**.

Figure 17.7 Plot After Second Split

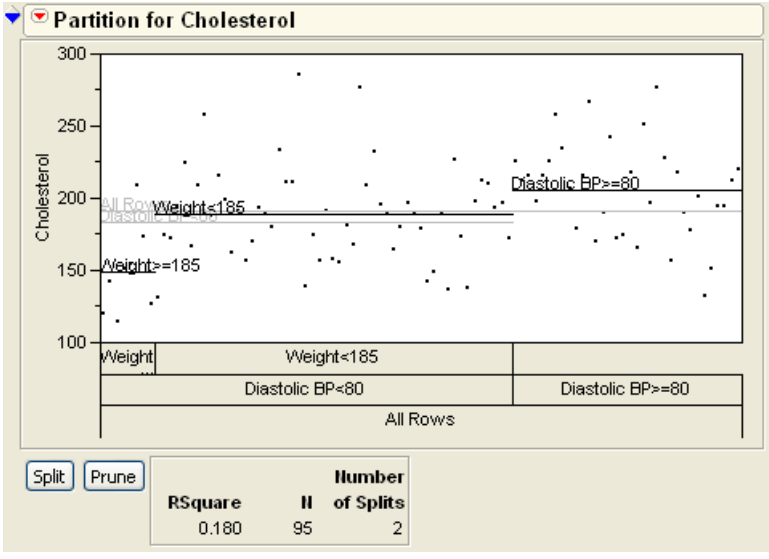
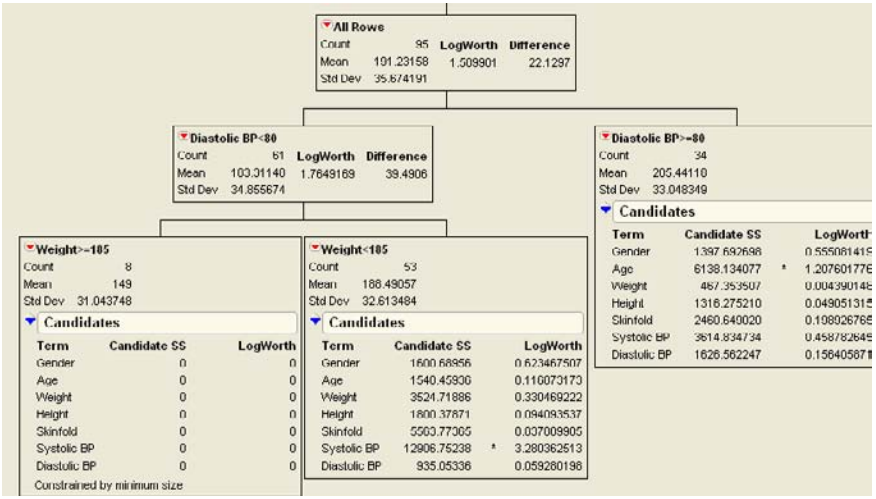


Figure 17.8 Tree After Second Split



This second split shows that of the people with diastolic blood pressure less than 80, their weight is the best predictor of cholesterol. The model predicts that those who weigh more than 185 pounds have a cholesterol of 149—less than those that weigh less than 185 pounds, whose average cholesterol is predicted as 188.5.

Splitting can continue in this manner until you are satisfied with the predictive power of the model. As opposed to software that continues splitting until a criterion is met, JMP allows you to be the judge of the effectiveness of the model.

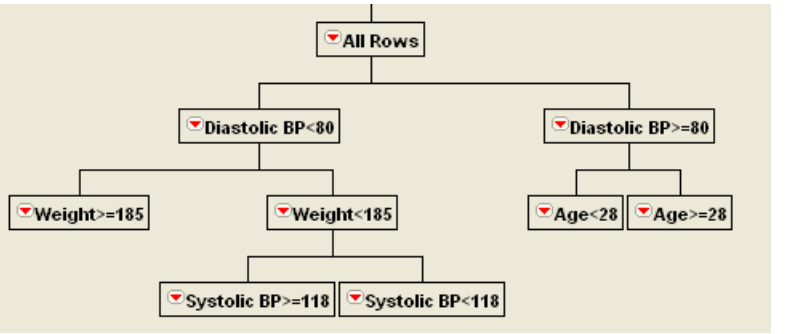
- Press the **Split** button two more times, to produce a total of four splits.

Viewing Large Trees

With several levels of partitioning, tree reports can become quite large. JMP has several ways to ease the viewing of these large trees.

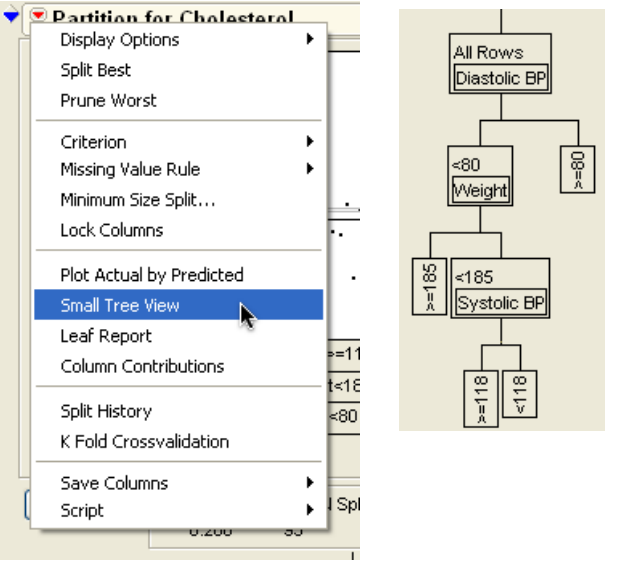
- Use the scroll tool on the tool bar ( ) or **Tools** menu to easily scroll around the report.
- Use the **Display Options**, found in the platform popup menu, to turn of parts of the report that are not relevant to your investigation. As an example, **Figure 17.9** shows the current lipid data set after four splits, with **Split Stats** and **Split Candidates** turned off.

Figure 17.9 Lipid Data After Four Splits



- Select **Small Tree View** from the menu on the title bar of the partition report. This option toggles a compact view of the tree, appended to the right of the main partition graph. **Figure 17.10** shows the Small Tree View corresponding to **Figure 17.9**.

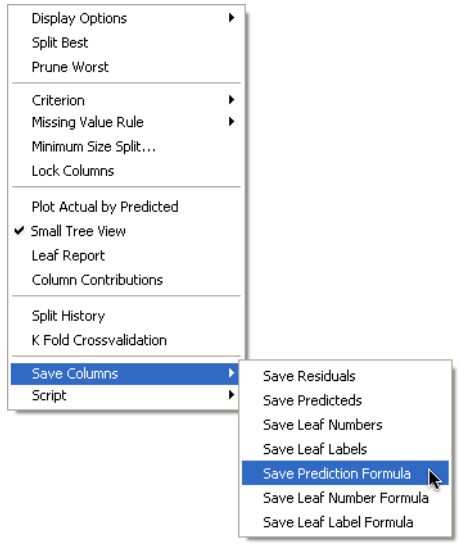
Figure 17.10 Small Tree View



Note: The font used for the small tree view is controlled by the **Small** font, found on the Fonts tab of the JMP preferences. **Figure 17.10** uses an 8-point Arial font.

Saving Results

The **Save Columns** submenu shows the options for saving results. All commands create a new column in the data table for storing their values.The commands that contain the word Formula (**Save Prediction Formula**, for example) store a formula in the column, where the other commands save values only. As an example,



- Select **Save Prediction Formula** from the **Save Columns** submenu.

This adds a column to the report named Cholesterol Predictor. It contains a formula that duplicates the partitions of the tree. To see the formula,

- Right-click in the title area of the Cholesterol Predictor column and select **Formula** from the menu that appears.

The formula for this example is shown in **Figure 17.11**. A series of `If` statements duplicates the model.

Figure 17.11 Prediction Formula

```
If( Diastolic BP<80 ⇒ If( Weight>=185 ⇒ 149
                          else ⇒ If( Systolic BP>=118 ⇒ 182.914893617021
                                     else ⇒ 232.166666666667 ) )
    else ⇒ 205.441176470588 )
```

Other Save commands are as follows.

- **Save Leaf Numbers** saves the leaf numbers of the tree to a column in the data table.
- **Save Leaf Labels** saves leaf labels of the tree to the data table. The labels document each branch that the row would trace along the tree, with each branch separated by &.
- **Save Prediction Formula** saves the prediction formula to a column in the data table. The formula is made up of nested conditional clauses.
- **Save Leaf Number Formula** saves a formula that computes the leaf number to a column in the data table.

- **Save Leaf Label Formula** saves a formula that computes the leaf label to a column in the data table.

Neural Networks

The Neural Net platform implements a standard type of neural network. Neural nets are used to predict one or more response variables from a flexible network of functions of S-shaped input variables. Neural networks can be very good predictors when it is not necessary to know the functional form of the response surface. Technical details of the particular functions used in the implementation are found in the *JMP Statistics and Graphics Guide*.

This section uses the Peanuts.jmp sample data file, from an experiment concerning a device for automatically shelling peanuts. A reciprocating grid is used to automatically shell the peanuts. The length and frequency of this stroke, as well as the spacing of the peanuts, are factors in the experiment. Kernel damage, shelling time, and the number of unshelled peanuts need to be predicted. We illustrate the procedure with unshelled peanuts, leaving the other two responses as exercises.

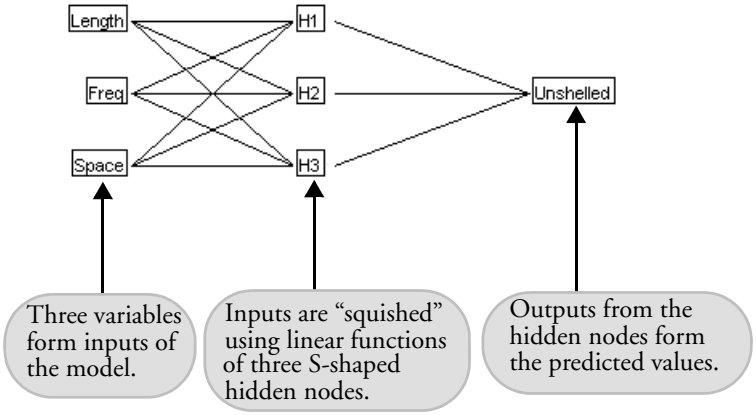
- Open the Peanuts.jmp sample data file.
- Select **Analyze > Modeling > Neural Net**.
- Assign Unshelled to Y and Length, Freq, and Space to X.
- Click **OK**.

When the Neural Net control panel appears,

- Select **Diagram** from the platform popup menu, located on the title bar of the panel.

The Neural Net diagram appears, as shown in **Figure 17.12**. The diagram illustrates that the factor columns are “squished” through three hidden nodes, whose outputs are used to form the predicted values.

Figure 17.12 Neural Net Diagram



The Neural Net control panel has the following options that allow you to quickly explore several models.

- **Hidden Nodes** is the most important number to specify. A value too low underfits the model, while a number too high overfits. There are no hard-and-fast rules on how many hidden nodes to specify; experience and exploration usually reveal the number to use.
- **Overfit Penalty** helps prevent the model from overfitting. When a neural net is overfit, the parameters are too big, so this criterion helps keep the parameters (weights) small. The penalty is often called *lambda* or *weight decay*. A zero value causes overfitting and causes difficulty in convergence. With lots of data, the overfit penalty has less of an effect. With very little data, the overfit penalty will damp down the estimates a lot. We recommended that you try several values between 0.0001 and 0.01.
- **Number of Tours** sets the number of tours. Neural nets tend to have a lot of local minima, so that in order to more likely find global minima, the model is fit many times (tours) at different random starting values. Twenty tours is recommended. If this takes up too much time, then specify fewer. If you don't trust that you have a global optimum, then specify more.
- **Max Iterations** is the number of iterations JMP takes on each tour before reporting nonconvergence.
- **Converge Criterion** is the relative change in the objective function that an iteration must meet to decide it has converged.
- **Log the tours** shows the best objective function value at each tour.
- **Log the iteration** shows the objective function for each iteration.

- **Log the estimates** shows the estimates at each iteration.
- **Save iterations in table** saves the estimates at each iteration in a data table.

Modeling with Neural Networks

Click **Go** to begin the fitting process.

When the iterations finish, results of the fit are shown as in **Figure 17.13**.

Note: These results are based on random starting points, so your results may vary from those shown here.

Figure 17.13 Fitting Results

Current Fit Results

	Objective	1	Converged At Best
SSE	0.8212767064	3	Converged Worse Than Best
Penalty	0.0855461321	0	Stuck on Flat
Total	0.9068228385	0	Failed to Improve
N	20	16	Reached Max Iter
Nparm	16		

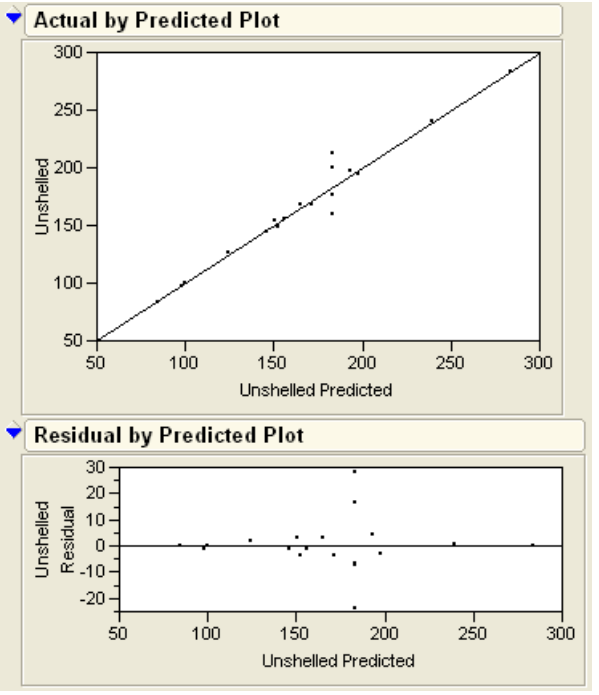
Y	SSE	RMSE	SSE Scaled	RSME Scaled	RSquare
Unshelled	1866.569602	10.8009537	0.8212767064	0.2265608	0.9568

In this example, one of the twenty tours converged at a maximum. Three others converged, but at a non-maximal point. Sixteen iterations did not converge after reaching the maximum number of iterations.

Following the results summary, parameter estimates and diagnostic plots are included (closed by default). The parameter estimates are rarely interesting (JMP doesn't even bother to calculate standard errors for them).

The Actual by Predicted plot and the Residual plot (**Figure 17.14**) are used similarly to their counterparts in linear regression. Use them to judge the predictive power of the model. In this example, the model fits fairly well, with no glaring problems in the residual plot.

Figure 17.14 Neural Net Plots



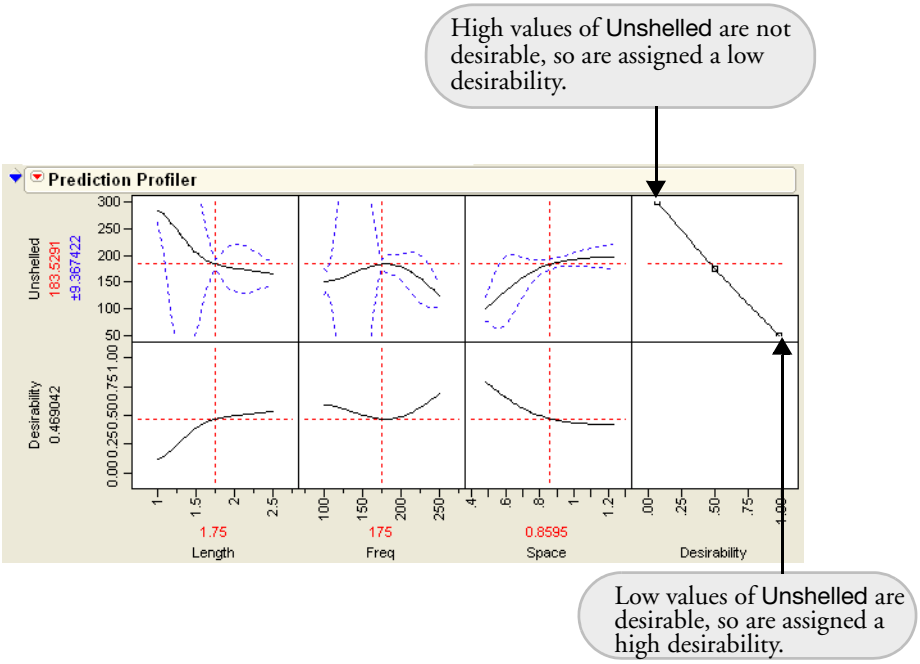
Profiles in Neural Nets

The slices through the response surface are informative.

- From the platform drop-down menu, select **Profiler**.

The Prediction Profiler (**Figure 17.15**) clearly shows the nonlinear nature of the model. Running the model with more hidden nodes increases the flexibility of these curves; running with fewer stiffens them.

Figure 17.15 Prediction Profiler



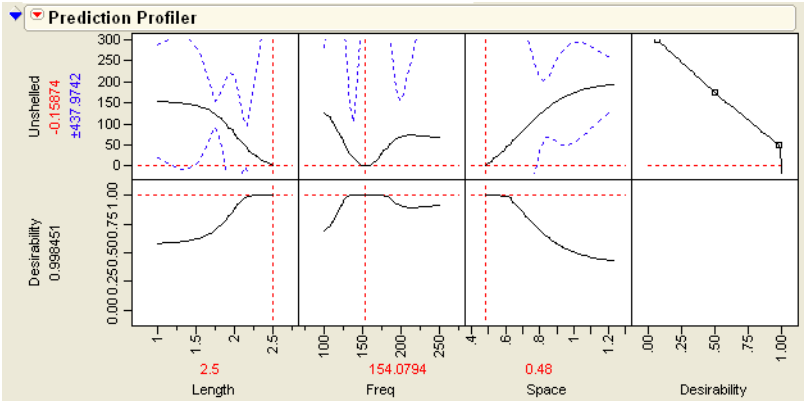
The Profiler retains all of the features used in analyzing linear models and response surfaces (discussed in "Analyze the Model" on page 421 of this book and in the *JMP Statistics and Graphics Guide*). Since we are interested in minimizing the number of unshelled peanuts, we utilize the Profiler's **Desirability Functions**, which are automatically shown (**Figure 17.15**).

To have JMP automatically compute the optimum value,

- Select **Maximize Desirability** from the platform drop-down menu.

JMP computes the maximum desirability, which in this example is a low value of Unshelled. Results are shown in **Figure 17.16**.

Figure 17.16 Maximized Desirability



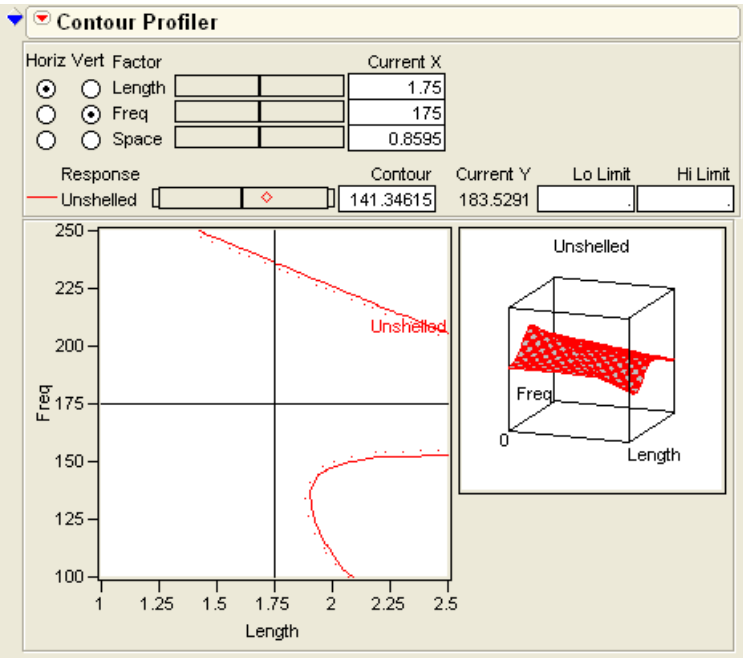
Optimal settings for the factors are shown (on screen in red) below the plots. In this case, optimal Unshelled values came from setting Length =2.5, Freq = 154, and Space = 0.48.

In addition to seeing two-dimensional slices through the response surface, the Contour Profiler can be used to visualize contours (sometimes called level curves) and mesh plots of the response surface.

- Select **Contour Profiler** from the platform drop-down menu.

The contours and mesh plot for this example are shown in **Figure 17.17**.

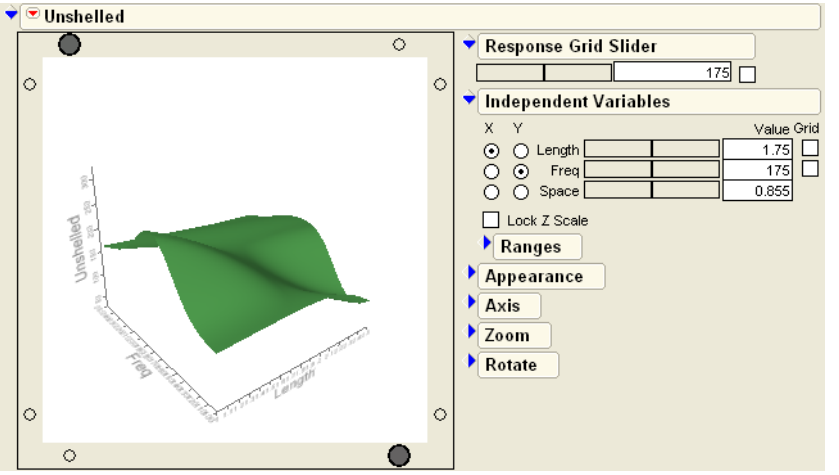
Figure 17.17 Contour Profiler



You can also get a good view of the response surface using the Surface Profiler.

Select **Surface Profiler** from the platform drop-down menu.

Figure 17.18 Surface Profiler



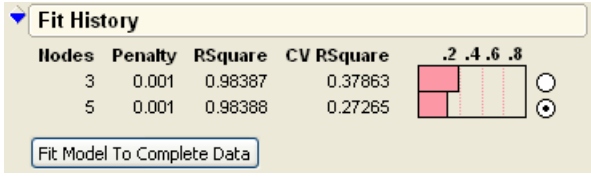
Using Cross-Validation

JMP offers two kinds of cross-validation, which allows for part of the data set to be used to fit the model, while the other part is used to test the model.

- **Random Holdback** partitions the data into a training set and a testing (holdback) set. You specify the percentage held back in the Neural Net Control Panel. The default holdback is 66.667%. Note that this technique is best used with large data sets.
- **K-Fold Cross-Validation** is a technique more suited to small data sets. The data is randomly divided into *k* groups. The model is re-estimated *k* times, holding back a different group each time. The number of groups (and, therefore, the fraction used for holdback) is specified in the Neural Net Control Panel. Complete details are in the *JMP Statistics and Graphics Guide* or the help system.

In either case, a Fit History window appears that describes the model fit for the training data. Here, we have used random hold-back cross-validation with three and five hidden nodes. Radio buttons on the left let you select the model that is the best-fitting, and a button lets you execute that fit for the complete data set.

Figure 17.19 Fit History Window



Saving Columns

Results from Neural Net analyses can be saved to columns in the data table. The following save options are available.

Save Hidden and Scaled Cols makes new columns containing the scaled data used to make the estimates, as well as the predicted values and the values of the hidden columns.

Save Predicted and Limits creates three new columns in the data table for each response variable. One column holds the predicted values, while the other two hold upper and lower 95% confidence intervals.

Save Formulas creates a new column in the data table for each response variable. This column holds the prediction formula for each response, so predicted values are calculated for each row. This option is useful if rows are added to the data table, since predicted values are

automatically calculated. Use **Save Predicted** if formulas are not desired. In this example, this command produces columns for each hidden layer, plus the following prediction formula.

$$\begin{bmatrix} -5.9269665257824 \\ +2.85916968262167*H1\ Formula \\ +3.72154127291047*H2\ Formula \\ +2.6589351260612*H3\ Formula \end{bmatrix} *47.6735334276125+168.15$$

Save Profile Formulas saves formulas almost equivalent to the formulas made with the **Save Formulas** command, but the hidden layer calculations are embedded into the final predictor formulas rather than made through an intermediate hidden layer. In this example, a column with the following formula is added.

$$\begin{bmatrix} -5.9269665257824 \\ +2.85916968262167*\text{Squish}\begin{bmatrix} 7.77629851927113 \\ +-1.5618760467786*Length \\ +-0.0385714849102*Freq \\ +5.11839597790781*Space \end{bmatrix} \\ +3.72154127291047*\text{Squish}\begin{bmatrix} -19.232317010146 \\ +0.53489901477582*Length \\ +0.08172738178618*Freq \\ +7.43806607367231*Space \end{bmatrix} \\ +2.6589351260612*\text{Squish}\begin{bmatrix} 29.9066419974379 \\ +-6.0538475749994*Length \\ +-0.0954096133777*Freq \\ +-6.2410071219191*Space \end{bmatrix} \end{bmatrix} \\ *47.6735334276125 \\ +168.15$$

Exercises

1. As shown in this chapter, the Lipids.jmp sample data set contains blood measurements, physical measurements, and questionnaire data from subjects in a California hospital. Repeat the Partition analysis of this chapter to explore models for

   (a) HDL cholesterol

   (b) LDL cholesterol

   (c) Triglyceride levels

2. Use the Peanuts.jmp sample data set and the Neural Net platform to complete this exercise. The factors of Freq, Space, and Length are described earlier in this chapter. Use these factors for the following:

   (a) Create a model for Time, the time to complete the shelling process. Use the Profiler and Desirability Functions to find the optimum settings for the factors that minimize the time of shelling.

   (b) Create a model for Damaged, the number of damaged peanuts after shelling is complete. Find values of the factors that minimize the number of damaged peanuts.

   (c) Compare the values found in the text of the chapter (**Figure 17.16**) and the values you found in parts (a) and (b) of this question. What settings would you recommend to the manufacturer?

18 Discriminant and Cluster Analysis

Overview

Both discriminant analysis and cluster analysis classify observations into groups. The difference is that discriminant analysis has actual groups to predict, whereas cluster analysis forms groups of points that are close together.

Discriminant Analysis

Discriminant analysis is appropriate for situations where you want to classify a categorical variable based on values of continuous variables. For example, you may be interested in the voting preferences (Democrat, Republican, or Other) of people of various ages and income levels. Or, you may want to classify animals into different species based on physical measurements of the animal.

There is a strong similarity between discriminant analysis and logistic regression. In logistic regression, the classification variable is random and predicted by the continuous variables, whereas in discriminant analysis the classifications are fixed, and the continuous factors are random variables. However, in both cases, a categorical value is predicted by continuous variables.

The discrimination is most effective when there are large differences among the mean values of the different groups. Larger separations of the means make it easier to determine the classifications.

The classification of values is completed using a *discriminant function*. This function is quite similar to a regression equation—it uses linear combinations of the continuous values to assign each observation into a categorical group.

The example in this section deals with a trace chemical analysis of cherts. Cherts are rocks formed mainly of silicon, and are useful to archaeologists in determining the history of a region. By determining the original location of cherts, inferences can be drawn about the peoples that used them in tool making. Klawiter (2000) was interested in finding a model that predicted the location of a chert sample based on a trace element analysis. A subset of his data is found in the data table Cherts.jmp.

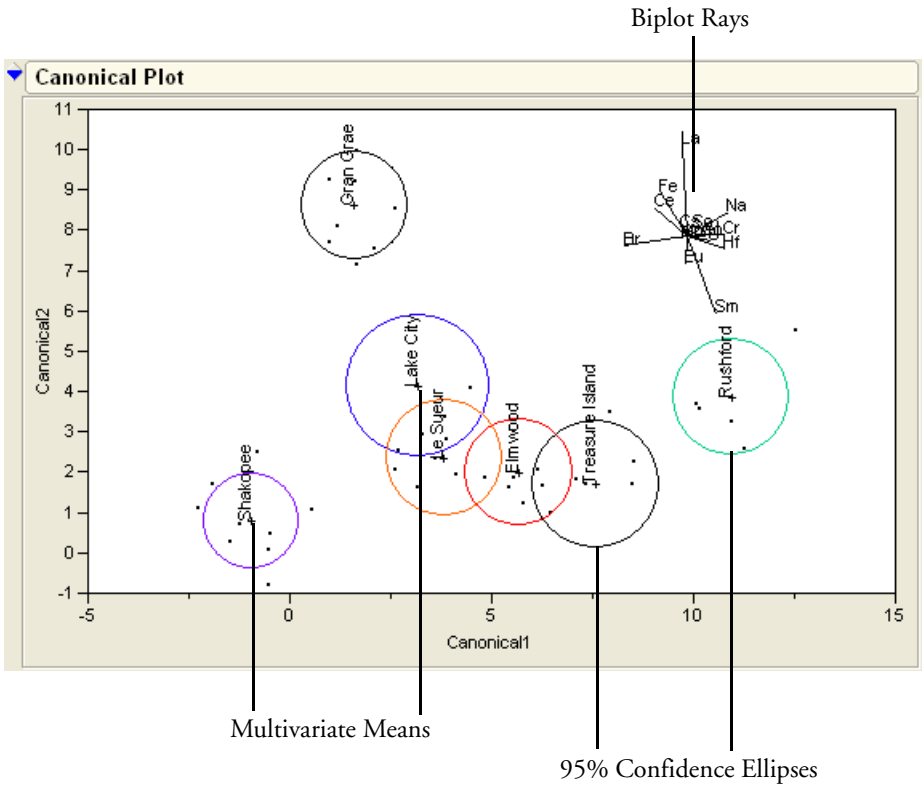
- Open the Cherts.jmp data table.
- Select **Analyze > Multivariate Methods > Discriminant**.
- Assign all the chemical names as Y and location name as X.
- Click **OK**.

The discriminant analysis report consists of two basic parts, the canonical plot and scores output.

Canonical Plot

The Canonical Plot shows the points and multivariate means in the two dimensions that best separate the groups. The canonical plot for this example is shown in **Figure 18.1**. Note that the biplot rays, which show the directions of the original variables in the canonical space, have been moved to better show the canonical graph. Click in the center of the biplot rays and drag them to move them around the report.

Figure 18.1 Canonical Plot of the Cherts Data



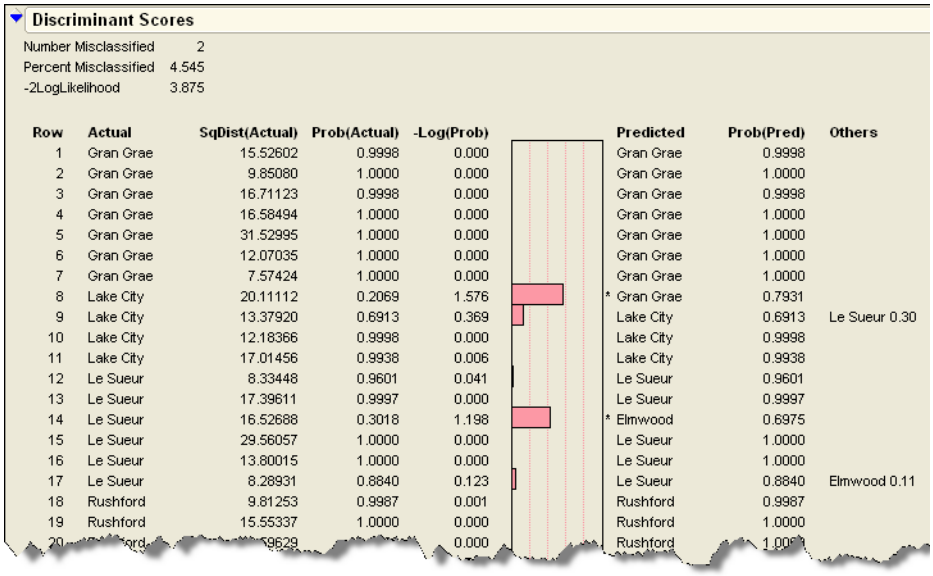
Each multivariate mean is surrounded by a 95% confidence ellipse, which appears circular in canonical space. In this example, the multivariate means for Shakopee, Gran Grae, and Rushford are more separated from the cluster of locations near the center of the graph.

Discriminant Scores

The scores report shows how well each point is classified. The first five columns of the report represent the actual (observed) data values, showing row numbers, the actual classification, the distance to the mean of that classification, and the associated probability. JMP graphs

–log(prob) in a histogram to show the loss in log-likelihood when a point is predicted poorly. When the histogram bar is large, the point is being poorly predicted. A portion of the discriminant scores for this example are shown in **Figure 18.2**.

Figure 18.2 Portion of Discriminant Scores Report



Discriminant Scores

Number Misclassified 2
Percent Misclassified 4.545
-2LogLikelihood 3.875

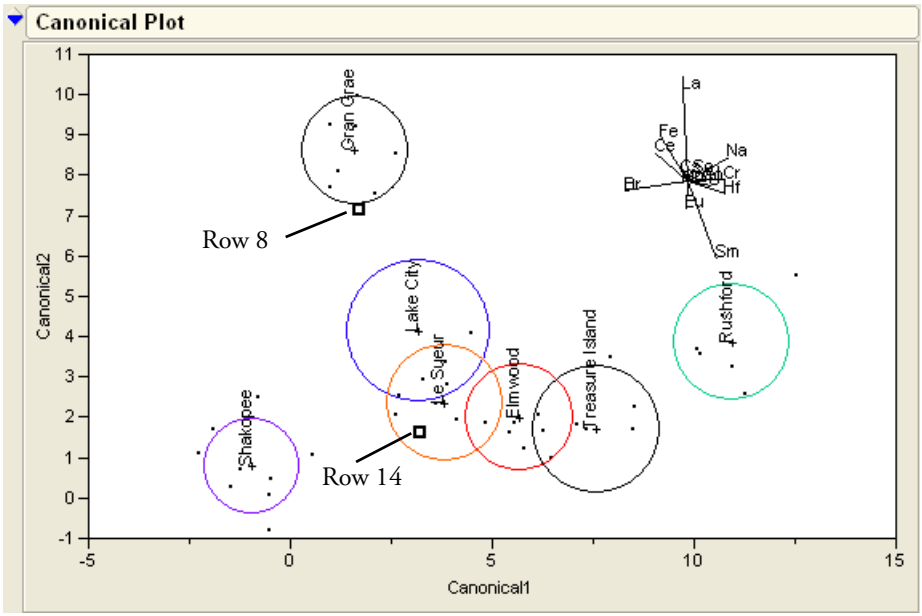
Row	Actual	SqDist(Actual)	Prob(Actual)	-Log(Prob)	Predicted	Prob(Pred)	Others
1	Gran Grae	15.52602	0.9998	0.000	Gran Grae	0.9998	
2	Gran Grae	9.85080	1.0000	0.000	Gran Grae	1.0000	
3	Gran Grae	16.71123	0.9998	0.000	Gran Grae	0.9998	
4	Gran Grae	16.58494	1.0000	0.000	Gran Grae	1.0000	
5	Gran Grae	31.52995	1.0000	0.000	Gran Grae	1.0000	
6	Gran Grae	12.07035	1.0000	0.000	Gran Grae	1.0000	
7	Gran Grae	7.57424	1.0000	0.000	Gran Grae	1.0000	
8	Lake City	20.11112	0.2069	1.576	* Gran Grae	0.7931	
9	Lake City	13.37920	0.6913	0.369	Lake City	0.6913	Le Sueur 0.30
10	Lake City	12.18366	0.9998	0.000	Lake City	0.9998	
11	Lake City	17.01456	0.9938	0.006	Lake City	0.9938	
12	Le Sueur	8.33448	0.9601	0.041	Le Sueur	0.9601	
13	Le Sueur	17.39611	0.9997	0.000	Le Sueur	0.9997	
14	Le Sueur	16.52688	0.3018	1.198	* Elmwood	0.6975	
15	Le Sueur	29.56057	1.0000	0.000	Le Sueur	1.0000	
16	Le Sueur	13.80015	1.0000	0.000	Le Sueur	1.0000	
17	Le Sueur	8.28931	0.8840	0.123	Le Sueur	0.8840	Elmwood 0.11
18	Rushford	9.81253	0.9987	0.001	Rushford	0.9987	
19	Rushford	15.55337	1.0000	0.000	Rushford	1.0000	
20	Rushford	9629		0.000	Rushford	1.000	

The predictions for rows 8 and 14 are incorrect, notated by an asterisk to the right of the plot. Why were these rows misclassified? Examining them in the canonical plot gives some insight.

- From the platform menu on the title bar of the Discriminant report, select **Score Options > Select Misclassified Rows**.
- Select **Rows > Markers** and select the square marker for these rows.

The result of this selection is shown in **Figure 18.3**.

Figure 18.3 Misclassified Rows



Row 8, although actually from Lake City, is very close to Gran Grae in canonical space. This closeness is the likely reason it was misclassified. Row 14, on the other hand, is close to Le Sueur, its actual value. It was misclassified because it was closer to another center, though this is not apparent in this projection of the 7-dimensional space.

Another quick way to examine misclassifications is to look at the Counts report (**Figure 18.4**) found below the discrimination scores. Zeros on the non-diagonal entries indicate perfect classification. The misclassified rows 8 and 14 are represented by the 1's in the non-diagonal entries.

Figure 18.4 Counts Report

Counts: Actual Rows by Predicted Columns

	Elmwood	Gran Grae	Lake City	Le Sueur	Rushford	Shakopee	Treasure Island
Elmwood	7	0	0	0	0	0	0
Gran Grae	0	7	0	0	0	0	0
Lake City	0	1	3	0	0	0	0
Le Sueur	1	0	0	5	0	0	0
Rushford	0	0	0	0	6	0	0
Shakopee	0	0	0	0	0	9	0
Treasure Island	0	0	0	0	0	0	5

Misclassified Rows

Cluster Analysis

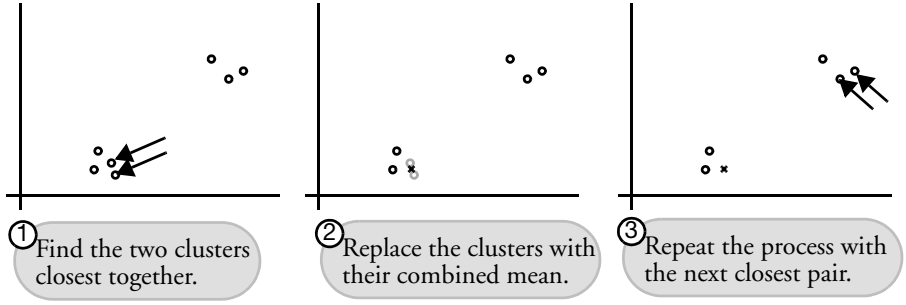
Cluster analysis is the process of dividing a set of observations into a number of groups where points inside each group are close to each other. JMP provides several methods of clustering, but here, we show only hierarchical clustering. Clustering groups observations into clusters based on some measure of distance. JMP measures distance in the simple Euclidean way. There are dozens of measures of the proximity of observations, but the essential point is that observations that are "close" to each other are joined together in groups. Each of them has the objective of minimizing within-cluster variation and maximizing between-cluster variation.

Hierarchical clustering is actually quite simple.

- Start with each point in its own cluster.
- Find the two clusters that are closest together in multivariate space.
- Combine these two clusters into a single group centered at their combined mean.
- Repeat this process until all clusters are combined into one group.

This process is illustrated in **Figure 18.5**.

Figure 18.5 Illustration of Clustering Process

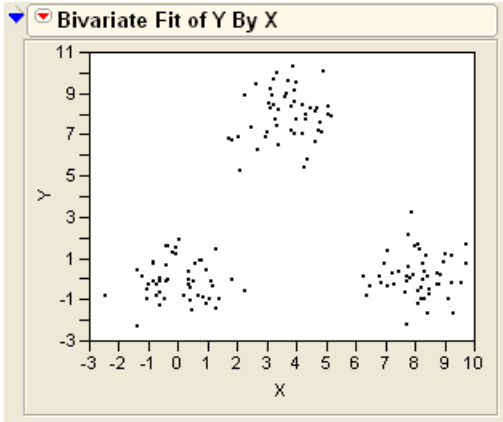


As a simple example, examine the SimulatedClusters.jmp data table.

- Open the SimulatedClusters.jmp data table.
- Select **Analyze > Fit Y By X**.
- Assign Y to **Y** and X to **X**, then click **OK**.

The results are shown in **Figure 18.6**.

Figure 18.6 Scatterplot of Simulated Data

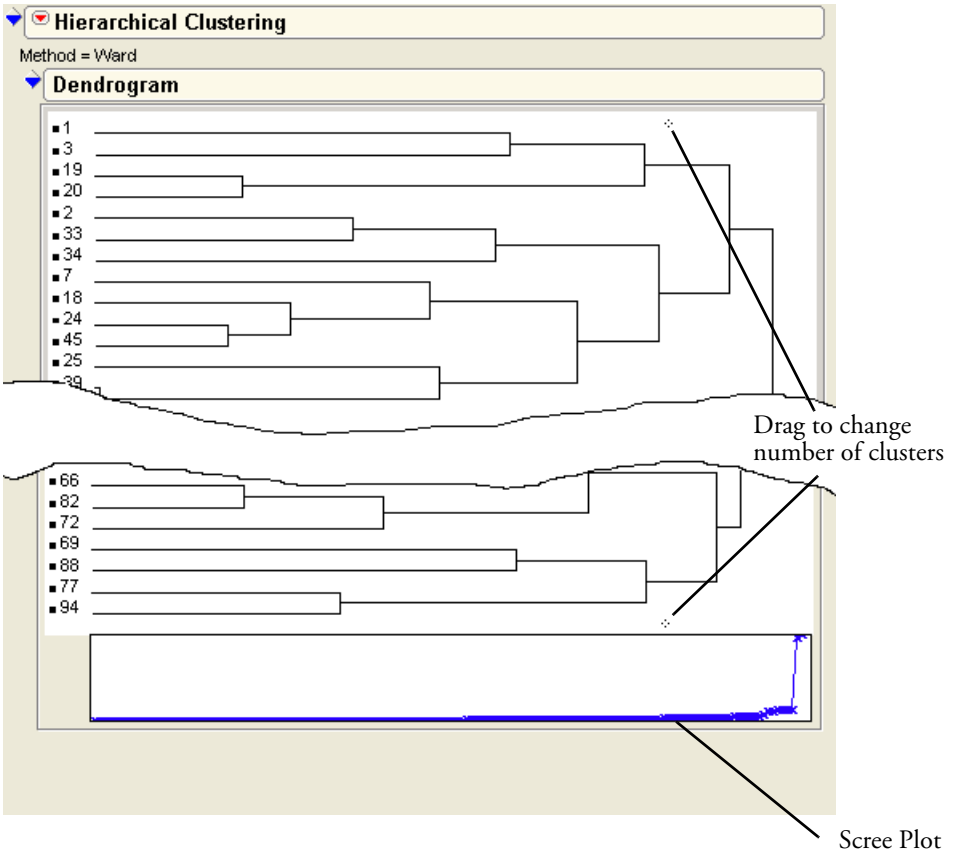


Obviously, the data clump together into three clusters. To analyze them with the clustering platform,

- Select **Analyze > Multivariate Methods > Cluster**.
- Assign X and Y to **Y, Columns**.
- Click **OK**.

The report appears as in **Figure 18.7**.

Figure 18.7 Dendrogram and Scree Plot



The top portion of the report shows a dendrogram, a visual tree-like representation of the clustering process. Branches that merge on the left were joined together earlier in the iterative algorithm.

Note the small diamonds at the bottom and the top of the dendrogram. These draggable diamonds adjust the number of clusters in the model.

Although there is no standard criterion for the best number of clusters to include, the Scree Plot (shown below the dendrogram) is shown to offer some guidance. Scree is a term for the rubble that accumulates at the bottom of steep cliffs, which this plot resembles. The place where the Scree Plot changes from a sharp downward slope to a more level slope (not always as obvious as in this example) is an indication of the number of clusters to include.

This example uses artificial data that was manufactured to have three clusters. Not surprisingly, the scree plot is very steep up to the point of three clusters, where it levels off.

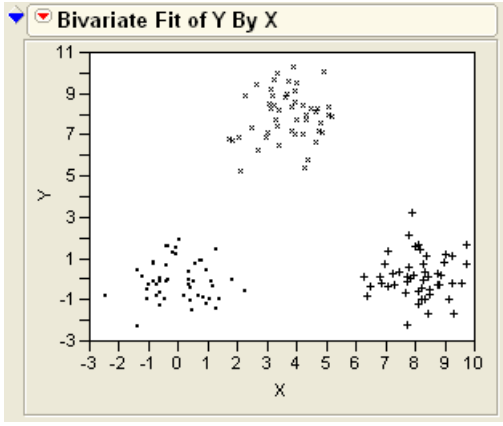
- From the platform drop-down menu, select both **Color Clusters** and **Mark Clusters**.

This assigns a special color and marker to each observation which changes dynamically as you change the number of clusters in the model. To see this,

- Drag the windows so that you can see both the Fit Y By X scatterplot and the dendrogram at the same time.
- Drag the number of cluster diamond to the right, observing the changes in colors and markers in the scatterplot.
- Move the marker to the point where there are three clusters. This should correspond to the level-off point of the Scree Plot.

The scatterplot should now look similar to the one in **Figure 18.8**.

Figure 18.8 Three Clusters



Once you decide that you have an appropriate number of clusters, you can save a column in the data table that holds the cluster value for each point.

- From the platform pop-up menu, select **Save Clusters**.

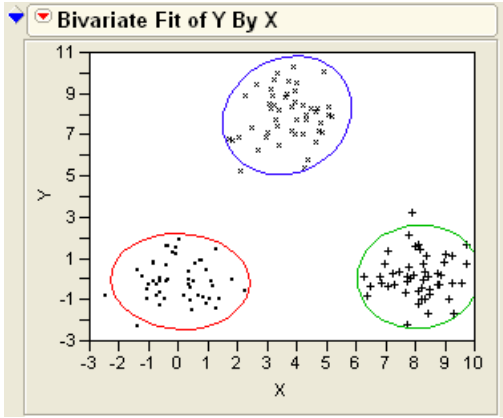
The cluster values are often useful in subsequent analyses and graphs. For example, you can draw density ellipses around each group.

- From the Fit Y By X report's drop-down menu, select **Group By**.

- In the dialog that appears, select the Cluster column and click **OK**.
- From the Fit Y By X drop-down menu select **Density Ellipse > .95**.

The display should now appear as in **Figure 18.9**.

Figure 18.9 Clusters with Density Ellipses



A Real-World Example

The data set Teeth.jmp contains measurements of the number of certain teeth for a variety of mammals. A cluster analysis can show which of these mammals have similar dental profiles.

- Open the data set Teeth.jmp.
- Select the Mammal column and assign it the **Label** role (**Cols > Label/Unlabel**).
- Select **Analyze > Multivariate Methods > Cluster**.
- Assign all the dental variables to **Y, Columns**.
- Click **OK**.

When the dendrogram appears,

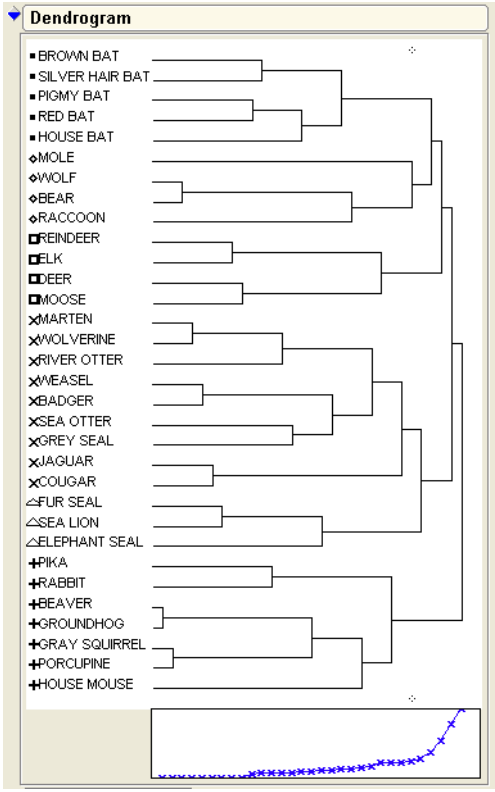
- Select **Mark Clusters** from the platform drop-down menu.

The Scree Plot (bottom of **Figure 18.10**) does not show a clear number of clusters for the model. However, there seems to be some leveling-off at a point around six or seven clusters.

- Drag the number of clusters diamond to the point corresponding to 6 clusters.

The report should appear as in **Figure 18.10**.

Figure 18.10 Dendrogram of Teeth Data



Examine the mammals classified in each cluster. Some conclusions based on these clusters are as follows.

- The five varieties of bats are in a cluster by themselves at the top of the dendrogram.
- Small mammals (like the rabbit, beaver, and squirrel) form a cluster at the bottom of the dendrogram.
- Seals and sea lions form a cluster, although not the same one as the sea otter and grey seal.
- Large mammals like the reindeer and moose form a cluster.

Exercises

1. A classic example of discriminant analysis uses Fisher's Iris data, stored in the data table Iris.jmp. Three species of irises (setosa, virginica, and versicolor) were measured on four variables (sepal length, sepal width, petal length, and petal width). Use the discriminant analysis platform to make a model that classifies the flowers into their respective species using the four measurements.
2. The data set Birth Death.jmp contains mortality (*i.e.* birth and death) rates for several countries. Use the cluster analysis platform to determine which countries share similar mortality characteristics. What do you notice that is similar among the countries that group together?

19 Statistical Quality Control

Overview

Some statistics are for proving things. Some statistics are for discovering things. And some statistics are to keep an eye on things, watching to make sure something stays within specified limits.

The watching statistics are needed mostly in industry for systems of machines in production processes that sometimes stray from proper adjustment. These statistics monitor variation, and their job is to distinguish the usual random variation (called *common causes*) from abnormal change (called *special causes*).

These statistics are usually from a time series, and the patterns they exhibit over time are clues to what is happening to the production process. If they are to be useful, the data for these statistics need to be collected and analyzed promptly so that any problems they detect can be fixed.

This whole area of statistics is called *Statistical Process Control* (SPC) or *Statistical Quality Control* (SQC). The most basic tool is a graph called a *control chart* (or *Shewhart control chart*, named for the inventor, Walter Shewhart). In some industries, SQC techniques are taught to everyone—engineers, mechanics, shop floor operators, even managers.

The use of SQC techniques became especially popular in the 1980s as industry began to better-understand the issues of quality, after the pioneering effort of Japanese industry and under the leadership of Edward Deming and Joseph Juran.

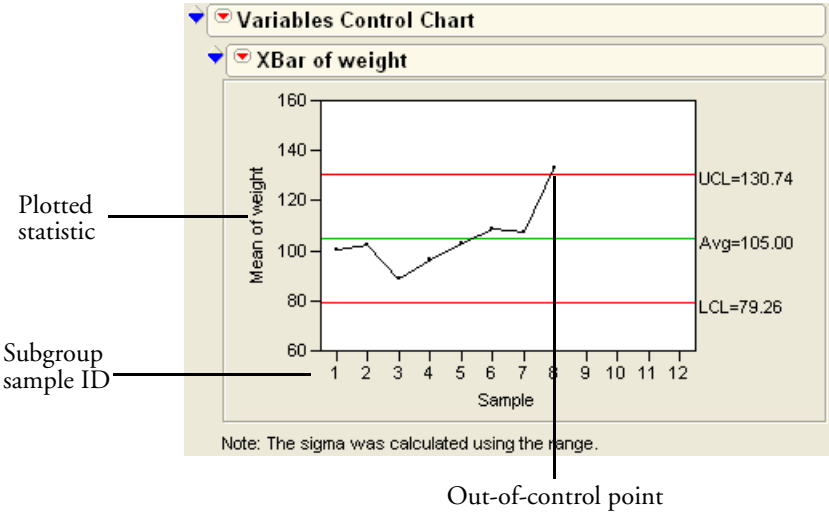
Control Charts and Shewhart Charts

Control charts are a graphical and analytical tool for deciding whether a process is in a state of statistical control. Control charts in JMP are automatically updated when rows are added to the current data table. In this way, control charts can be used to monitor an ongoing process.

Figure 19.1 shows a control chart that illustrates characteristics of most control charts:

- Each point represents a summary statistic computed from a subgroup sample of measurements of a quality characteristic.
- The vertical axis of a control chart is scaled in the same units as the summary statistics specified by the type of control chart.
- The horizontal axis of a control chart identifies the subgroup samples.
- The center line on a Shewhart control chart indicates the average (expected) value of the summary statistic when the process is in statistical control.
- The upper and lower control limits, labeled UCL and LCL, give the range of variation to be expected in the summary statistic when the process is in statistical control.
- A point outside the control limits signals the presence of a special cause of variation.

Figure 19.1 Control Chart Example



Control charts are broadly classified as *variables charts* and *attributes charts*.

Variables Charts

Control Charts for variables (variables charts) are used when the quality characteristic to be monitored is measured on a continuous scale. There are different kinds of variables control charts based on the subgroup sample summary statistic plotted on the chart, which can be the mean, the range, or the standard deviation of a measurement, an individual measurement itself, or a moving range. For quality characteristics measured on a continuous scale, it is typical to analyze both the process mean and its variability by showing a Mean chart aligned above its corresponding *r*-(range) or *s*- (standard deviation) chart. If you are charting individual response measurements, the Individual Measurement chart is aligned above its corresponding Moving Range chart. JMP automatically arranges charts in this fashion.

Attributes Charts

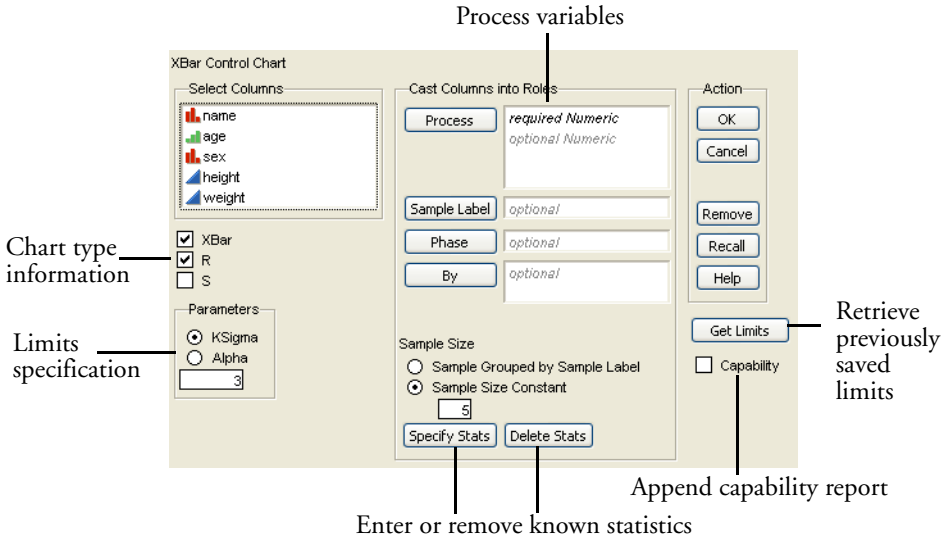
Control Charts for attributes (attributes charts) are used when the quality characteristic of a process is measured by counting the number or the proportion of nonconformities (defects) in an item, or by counting the number or proportion of nonconforming (defective) items in a subgroup sample.

The Control Chart Launch Dialog

When you select a particular control chart from the **Graph > Control Charts** menu, you see the appropriate Control Chart launch dialog. An example is shown in **Figure 19.2**. You can think of the control chart dialog as a composite of up to four panels in that request four kinds of information:

- Process information
- Chart type information
- Test requests
- Limits specification

Figure 19.2 Control Chart Dialog



Specific information shown in the dialog varies according to the kind of chart you request. Through interaction with this dialog, you specify exactly how you want the charts created. The next sections discuss the kinds of information needed to complete the Control Chart dialog and talk about types of control charts.

Process Information

The dialog displays a list of all columns in the current data table and has buttons to specify the variables to be analyzed, the subgroup sample size, and (optionally) the subgroup sample ID.

The following buttons have different functions depending on the chart type you choose.

Process

identifies variables for charting.

- For variables charts, specify the measurements.
- For attributes charts, specify the defect count or defect proportion.

Sample Label

lets you specify a variable whose values label the horizontal axis and can also identify unequal subgroup sizes. If no sample ID variable is specified, the samples are identified by their sequence number.

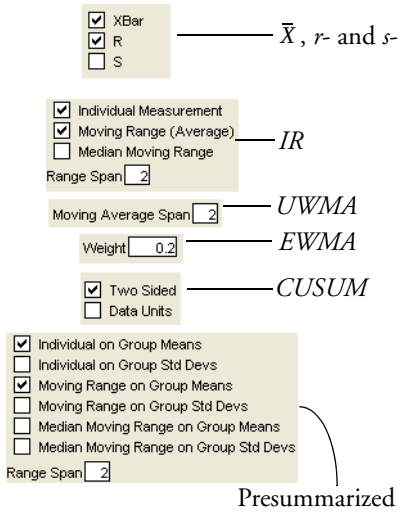
- If the subsamples are the same size, check the Sample Size Constant radio button and enter the size into the text box. If you entered a Sample ID variable, its values are used to label the horizontal axis.
- If the subsamples have an unequal number of rows or have missing values, check the **Sample Grouped by Sample Label** radio button and remove sample size information from the Constant Sample Size text box.

For attributes charts (*p*, *np*, *c*, and *u*), this variable is the subgroup sample size. In Variable charts, it identifies the sample. When the chart type is IR, a Range Span text entry box appears. The range span specifies the number of consecutive measurements from which the moving ranges are computed. These chart types are described in more detail later.

Chart Type Information

The Chart Type panel varies based on the selected chart type.

- The Shewhart Variable charts menu selection gives **XBar**, **R**, and **S** check boxes. The IR menu selection has check-box options for the Individual Measurement, Moving Range, and Median Moving range charts.
- Attributes chart selections are the **P**, **NP**, **C**, and **U** charts. There are no additional specifications for attributes charts.
- The uniformly weighted moving average (**UWMA**) and exponentially weighted moving average (**EWMA**) selections are special charts for means.



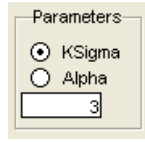
Descriptions and examples of specific kinds of charts are given later in this chapter.

Limits Specification Panel

The Limits Specification Panel allows you to specify control limits computations by entering a value for *k* (**K Sigma**) or by entering a probability α (**Alpha**), or by retrieving a limits value from a previously created Limits Table (discussed later).

There must be a specification of either **K Sigma** or **Alpha**.

K Sigma



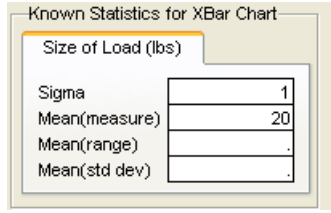
allows specification of control limits in terms of a multiple of the sample standard error. K Sigma specifies control limits at *k* sample standard errors above and below the expected value, which shows as the center line. To specify *k*, the number of sigmas, click K Sigma and enter a positive *k* value into the box. The usual choice for *k* is 3.

Alpha

specifies control limits (also called probability limits) in terms of the probability that a single subgroup statistic exceeds its control limits, assuming that the process is in control. To specify α, click the Alpha radio button and enter the probability you want. Reasonable choices for alpha are 0.01 or 0.001.

Using Known Statistics

If you click the **Specify Stats** button on the Control Charts Launch dialog, a tab with editable fields is appended to the bottom of the launch dialog. This lets you enter known statistics for the process variable. The Control Chart platform uses those entries to construct control charts. The example to the right shows 1 as the standard deviation of the process variable and 20 as the mean measurement.



Types of Control Charts for Variables

Control charts for variables are classified according to the subgroup summary statistic plotted on the chart:

- **XBar** charts display subgroup means (averages).
- **R** charts display subgroup ranges (maximum –minimum).
- **S** charts display subgroup standard deviations.

The **IR** selection gives two additional chart types:

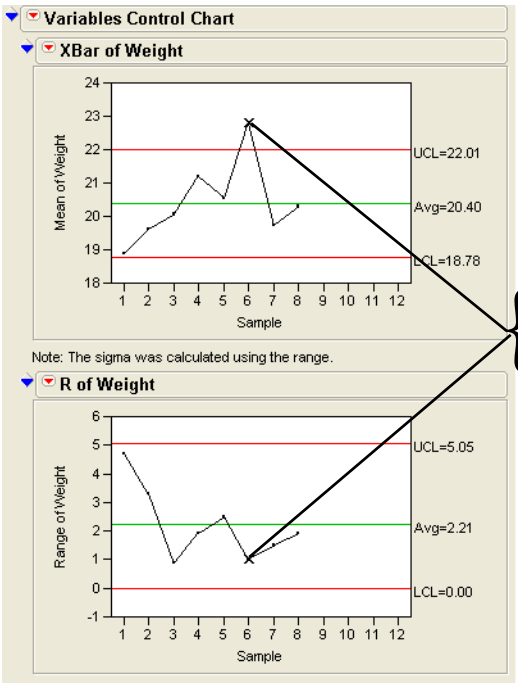
- **Individual Measurement** chart type displays individual measurements.
- **Moving Range** chart type displays moving ranges of two or more successive measurements. Moving ranges are computed for the number of consecutive measurements you enter in the Range Span box. The default range span is 2. Because moving ranges are correlated, these charts should be interpreted with care.

Mean, R, and S Charts

For quality characteristics measured on a continuous scale, a typical analysis shows both the process mean and its variability with a $\bar{X}$-chart aligned above its corresponding *r*- or *s*-chart.

The example in **Figure 19.3** uses the Coating.jmp data (taken from the *ASTM Manual on Presentation of Data and Control Charts*). The quality characteristic of interest is the Weight column. A subgroup sample of four is chosen. The $\bar{X}$-chart and an *r*- chart for the process show that sample six is above the upper control limit (UCL), which indicates that the process is not in statistical control. To check the sample values, you can click the sample summary point on either control chart and the corresponding rows highlight in the data table.

Figure 19.3 Variables Charts for Coating Data



	19	19	5	20.8	.
	20	20	5	21.6	21.6
×	21	21	6	22.8	22.8
×	22	22	6	22.2	22.2
×	23	23	6	23.2	23.2
×	24	24	6	23	23
	25	25	7	19	19
	26	26	7	20.5	20.5

Individual Measurement and Moving Range Charts

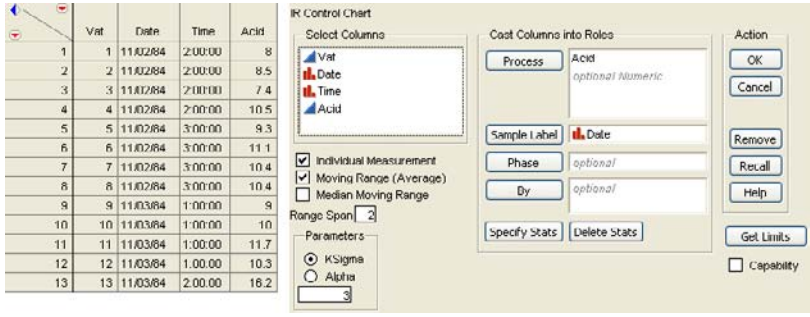
If you are charting individual measurements, the Individual Measurement chart shows above its corresponding Moving Range chart. Follow the example dialog in **Figure 19.4** to see these kinds of control charts.

- Open the Pickles.jmp data table.

The data show the acid content for vats of pickles. The pickles are produced in large vats, and high acidity can ruin an entire pickle vat. The acidity in four randomly selected vats was measured each day at 1:00, 2:00, and 3:00 PM. The data table records day (Date), time (Time), and acidity (Acid) measurements.

- Choose **Graph > Control Charts** and select IR from the Chart Type menu. Use Acid as the process variable, Date as the Sample Label, and enter a range span of 2 for the Moving Range chart (see **Figure 19.4**).

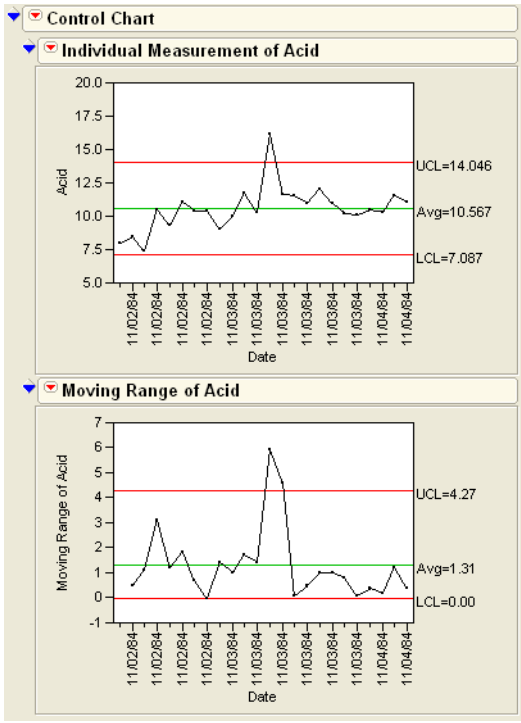
Figure 19.4 Data and Control Chart Dialog for Individual Measurement Chart



- Click the **OK** button on the dialog to see the Individual Measurement and Moving Range charts shown in **Figure 19.5**.

In the pickle example, the Date variable labels the horizontal axis, which has been modified to better display the values. Tailoring axes is covered later in the chapter.

Figure 19.5 Individual Measurement Charts for Pickles Data

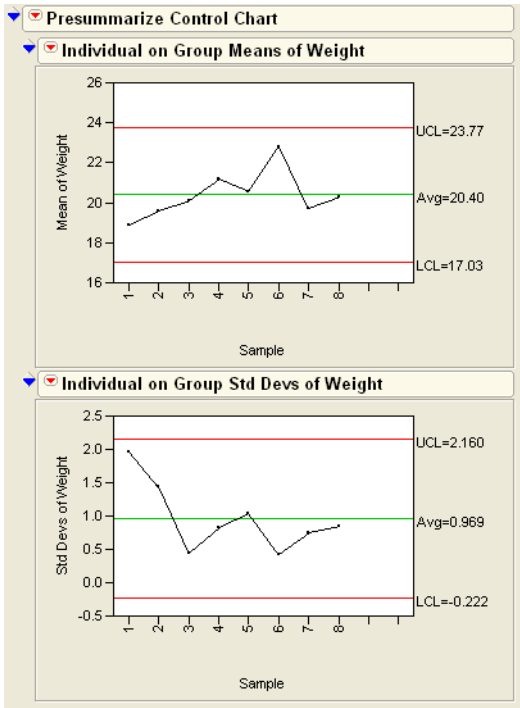


Presummarize charts summarize the process column before charting it. For an example, using the Coating.jmp data table,

- Choose **Graph > Control Chart** > **Presummarize**.
- Choose Weight as the **Process** variable and Sample as the **Sample Label**.
- Check only **Individual on Group Means** and **Individual on Group Std Devs**.

The **Group by Sample Label** button is automatically selected when you choose a Sample Label variable.

Figure 19.6 Example of Charting Pre-Summarized Data



Although the points for $\bar{X}$-and s- charts are the same as the **Indiv of Mean** and **Indiv of Std Dev** charts, the limits are different because they are computed as Individual charts.

Here is another way to generate the pre-summarized charts:

- Open Coating.jmp.
- Choose **Tables > Summary**.
- Assign Sample as the Group variable, then Mean(Weight) and Std Dev(Weight) as statistics, and click **OK.**
- Select **Graph > Control Chart**.
- Set the chart type again to **IR** and choose both Mean(Weight) and Std Dev(Weight) as process variables.
- Click **OK**.

These new charts match the pre-summarized charts.

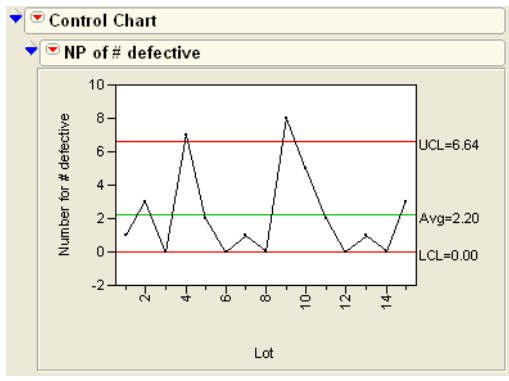
Types of Control Charts for Attributes

Attributes charts, like variables charts, are classified according to the subgroup sample statistic plotted on the chart:

- *p-charts* display the proportion of nonconforming (defective) items in a subgroup sample.
- *np-charts* display the number of nonconforming (defective) items in a subgroup sample.
- *c-charts* display the number of nonconformities (defects) in a subgroup sample that usually consists of one inspection unit.
- *u-charts* display the number of nonconformities (defects) per unit in a subgroup sample with an arbitrary number of inspection units.

p- and *np*-Charts

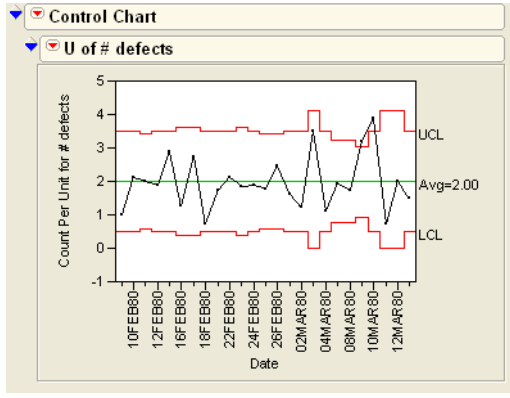
The Washers.jmp data table contains defect counts of 15 lots of 400 galvanized washers. The washers were inspected for finish defects such as rough galvanization and for exposed steel. The chart to the right illustrates an *np*-chart for the number of defects in the Washers data. A corresponding *p*-chart is identical except the vertical axis scale shows proportions instead of counts.



u-Charts

The Braces.jmp data records the defect count in boxes of automobile support braces. A box of braces is one inspection unit. The number of defective braces found in a day is the process variable. The subgroup sample size is the number of boxes inspected in a day, which can vary.

The *u*-chart shown to the right is monitoring the number of brace defects per unit. The upper and lower bounds vary according to the number of units (boxes of braces) inspected.



Moving Average Charts

The control charts previously discussed plot each point based on information from a single subgroup sample. The Moving Average chart is different from other types because each point combines information from the current sample and from past samples. As a result, the Moving Average chart is more sensitive to small shifts in the process average. On the other hand, it is more difficult to interpret patterns of points on a Moving Average chart because consecutive moving averages can be highly correlated (Nelson 1984).

In a Moving Average chart, the quantities that are averaged can be individual observations instead of subgroup means. However, a Moving Average chart for individual measurements is not the same as a control (Shewhart) chart for individual measurements or moving ranges with individual measurements plotted.

Uniformly Weighted Moving Average (UWMA) Charts

Each point on a Uniformly Weighted Moving Average (UWMA) chart is the average of the w most recent subgroup means, including the present subgroup mean. When you obtain a new subgroup sample, the next moving average is computed by dropping the oldest of the previous w subgroup means and including the newest subgroup mean. The constant w is called the *span* of the moving average. There is an inverse relationship between w and the magnitude of the shift that can be detected. Thus, larger values of w allow the detection of smaller shifts. Complete the following steps to see an example.

- Open the data table called Clips1.jmp.

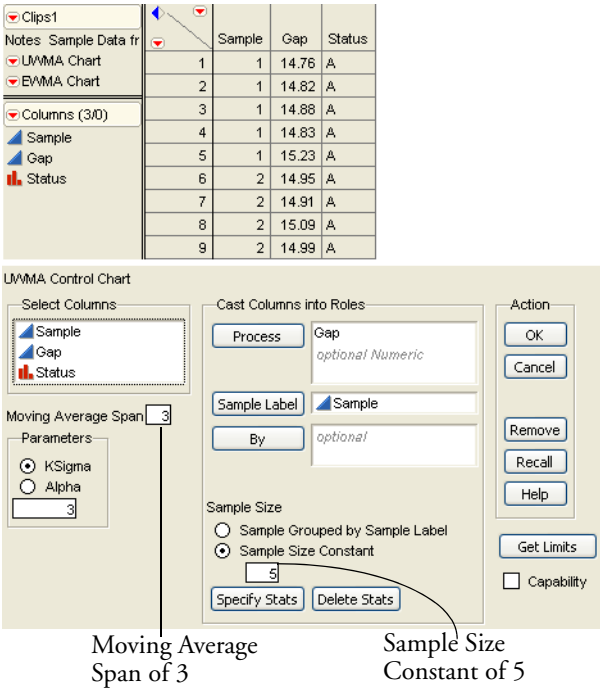
A partial listing of the data is shown in **Figure 19.7**.

The measure of interest is the gap between the ends of manufactured metal clips. To monitor the process for a change in average gap, subgroup samples of five clips are selected daily, and a UWMA chart with a moving average span of three samples is examined. To see the UWMA chart, complete the Control Chart dialog.

- Choose **Graph > Control Charts** and select **UWMA** from the Chart Type drop down menu.
- Use Gap as the **Process** variable and Sample as the **Sample Label**.
- Enter 3 as the moving average span and 5 as the Constant sample size.

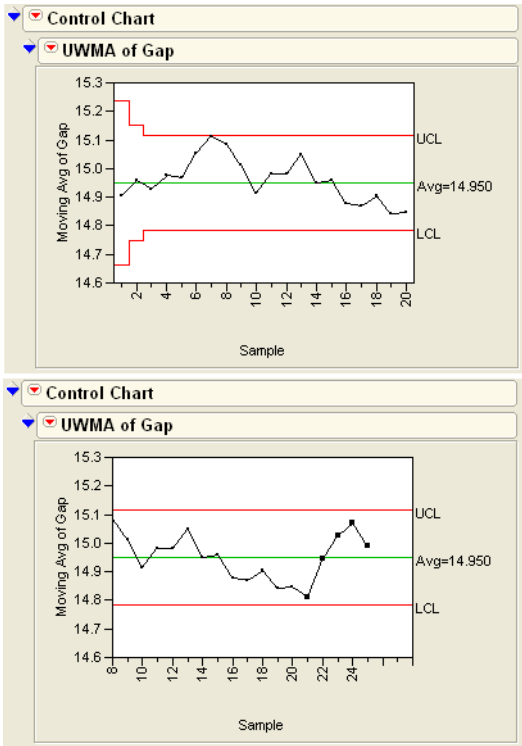
The completed Control Chart dialog should look like the one shown in **Figure 19.7**.

Figure 19.7 Specification for UWMA Charts of Clips1 Data



Click the **OK** button to see the top chart in **Figure 19.8**.

Figure 19.8 UWMA Charts for the Clips1 Data



The point for the first day is the mean of the first subsample only, which consists of the five sample values taken on the first day. The plotted point for the second day is the average of subsample means for the first and second day. The points for the remaining days are the average of subsample means for each day and the two previous days.

Like all control charts, the UWMA chart updates dynamically when you add rows to the current data table.

Add rows to the Clips1 data table as follows:

- Open the clipsadd.jmp data table.
- Shift-click at the top of both columns, Sample and Gap, to highlight them.
- Choose **Edit > Copy** to copy the two columns to the clipboard.
- Click on the Clips1 data table to make it the active table.
- Click at the top of the Sample and Gap columns to highlight them in the Clips1 table.

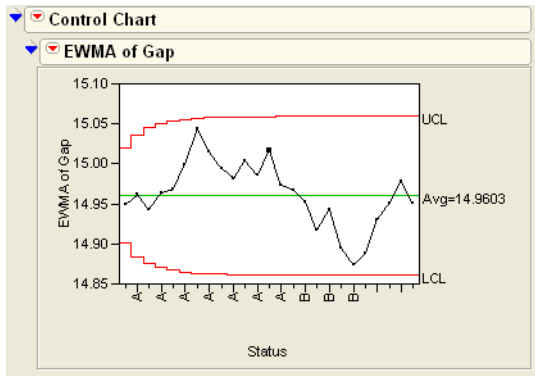
- Click in the cell immediately below the Sample measurement in the last row (in the first column that will become row 101).
- Choose **Edit > Paste** to append the contents of the clipboard to the Clips1 table.

When you paste the new data into the table, the chart immediately updates, as shown in the bottom chart in **Figure 19.8**.

Exponentially Weighted Moving Average (EWMA) Chart

Each point on an Exponentially Weighted Moving Average (EWMA) chart, also referred to as a Geometric Moving Average (GMA) chart, is the weighted average of all the previous subgroup means, including the mean of the present subgroup sample.

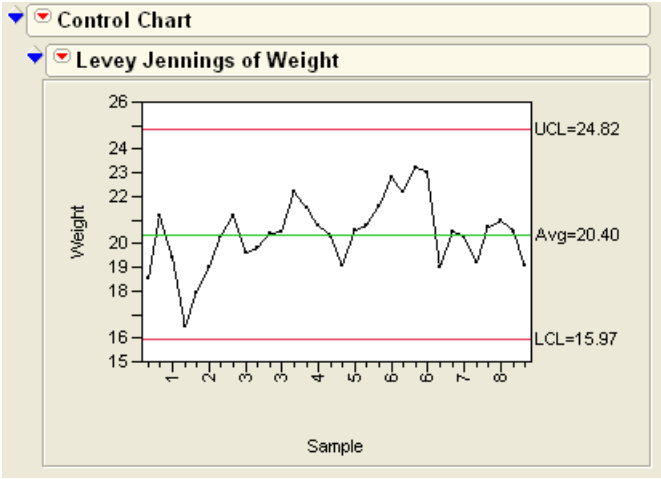
The weights decrease exponentially going backward in time. The weight ($0 < r < 1$) assigned to the present subgroup sample mean is a parameter of the EWMA chart. Small values of r are used to guard against small shifts. If $r = 1$, the EWMA chart reduces to a Mean control (Shewhart) chart, previously discussed. The figure shown here is an EWMA chart for the same data used for **Figure 19.8**.



Levey-Jennings Plots

Levey-Jennings plots show a process mean with control limits based on a long-term sigma. Control limits are placed at a distance $3s$ from the control line. See the *JMP Statistics and Graphics Guide* for details on the computation of s.

Figure 19.9 Levey-Jennings Plot Using the Coating Data

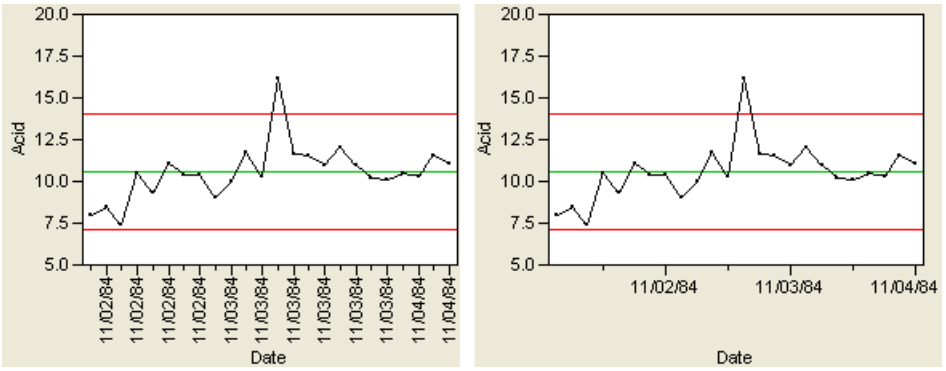


Tailoring the Horizontal Axis

When you double-click the *x*-axis, a dialog appears that allows you to specify the number of ticks to be labeled.

For example, the Pickles.jmp example, seen previously, lists eight measures a day for three days. **Figure 19.10** shows Individual Measurement charts for the Pickles data. The *x*-axis is labeled at every tick. Sometimes this gives hard-to-decipher labels, as shown to the left in **Figure 19.10**. If you specify a label for every eighth tick mark, and uncheck the option for rotated labels, the *x*-axis is labeled once for each day, as shown in the plot on the right.

Figure 19.10 Example of Modified *x*-Axis Tick Marks



Tests for Special Causes

You can select one or more tests for special causes (often called the *Western Electric rules*) in the Control Chart dialog or with the options popup menu above each plot. Nelson (1984, 1985) developed the numbering notation to identify special tests on control charts.

Table 19.1 lists and interprets the eight tests, and **Figure 19.11** illustrates the tests.

The following rules apply to each test:

- The area between the upper and lower limits is divided into six zones, each with a width of one standard deviation.
- The zones are labeled A, B, C, C, B, A with Zone C nearest the center line.
- A point lies in Zone B or beyond if it lies beyond the line separating zones C and B.
- Any point lying on a line separating two zones is considered to belong to the outermost zone.

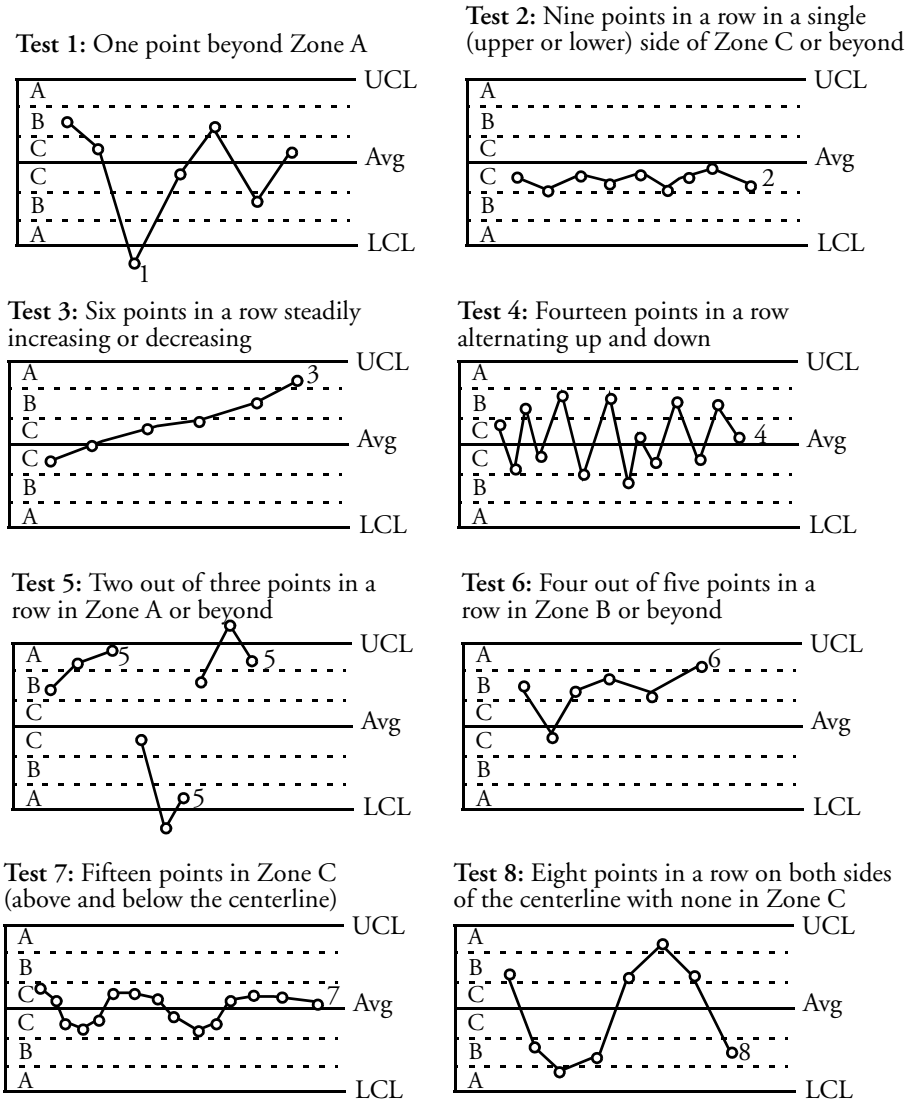
Tests 1 through 8 apply to Mean and Individual Measurement charts. Tests 1 through 4 can also apply to *p*-, *np*-, *c*-, and *u*- charts.

Table 19.1. Description of Special Causes Tests (from Nelson, 1984,1985)

Test 1	One point beyond Zone A	Detects a shift in the mean, an increase in the standard deviation, or a single aberration in the process. For interpreting Test 1, the *R* chart can be used to rule out increases in variation.
Test 2	Nine points in a row in Zone C or beyond	Detects a shift in the process mean.
Test 3	Six points in a row steadily increasing or decreasing	Detects a trend or drift in the process mean. Small trends are signaled by this test before Test 1.
Test 4	Fourteen points in a row alternating up and down	Detects systematic effects such as two alternately used machines, vendors, or operators.
Test 5	Two out of three points in a row in Zone A or beyond	Detects a shift in the process average or increase in the standard deviation. Any two out of three points provide a positive test.
Test 6	Four out of five points in Zone B or beyond	Detects a shift in the process mean. Any four out of five points provide a positive test.
Test 7	Fifteen points in a row in Zone C, above and below the center line	Detects stratification of subgroups when the observations in a single subgroup come from various sources with different means.
Test 8	Eight points in a row on both sides of the center line with none in Zone C	Detects stratification of subgroups when the observations in one subgroup come from a single source, but subgroups come from different sources with different means.

Tests 1, 2, 5, and 6 apply to the upper and lower halves of the chart separately. Tests 3, 4, 7, and 8 apply to the whole chart.

Figure 19.11 Illustration of Special Causes Tests (from Nelson 1984, 1985)



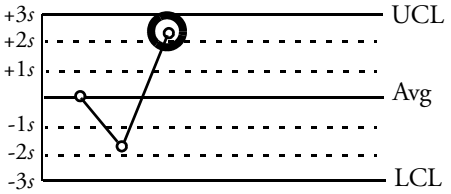
Westgard Rules

Westgard rules are implemented under the **Westgard Rules** submenu of the Control Chart platform. The different tests are abbreviated with the decision rule for the particular test. For example, **1 2s** refers to a test that one point is two standard deviations away from the mean.

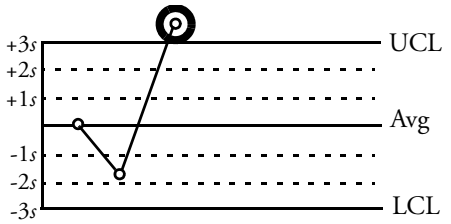
Because Westgard rules are based on sigma and not the zones, they can be computed without regard to constant sample size.

Table 19.2. Westgard Rules

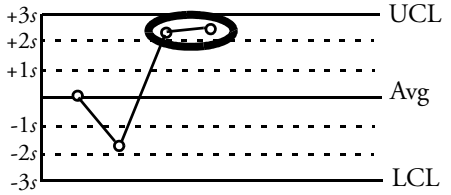
Rule 1 2s is commonly used with Levey-Jennings plots, where control limits are set 2 standard deviations away from the mean. The rule is triggered when any one point goes beyond these limits.



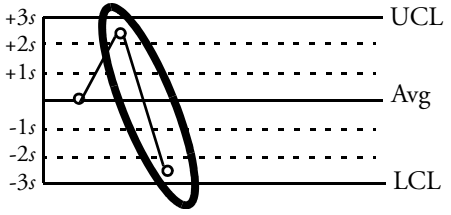
Rule 1 3s refers to a rule common to Levey-Jennings plots where the control limits are set 3 standard deviations away from the mean. The rule is triggered when any one point goes beyond these limits.



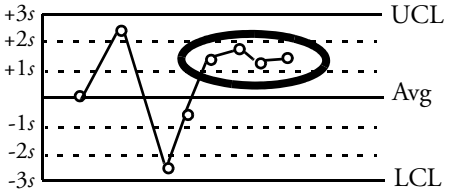
Rule 2 2s is triggered when two consecutive control measurements are farther than two standard deviations from the mean.



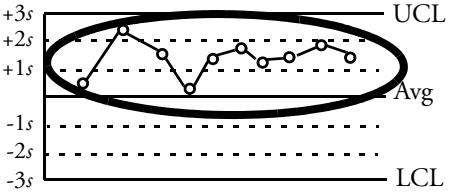
Rule 4s is triggered when one measurement in a group is two standard deviations above the mean and the next is two standard deviations below.



Rule 4 1s is triggered when four consecutive measurements are more than one standard deviation from the mean.



Rule 10 X is triggered when ten consecutive points are on one side of the mean.



Multivariate Control Charts

JMP has a Multivariate Control Chart platform that is useful when there are several related variables of interest. Multivariate control charts are used in two stages. Initially, they are used with historical data to establish control limits. Then, with control limits established, they are used to monitor a process. Details of the multivariate control chart are found in the *JMP Statistics and Graphics Guide*.

20 Time Series

Overview

When you have time series data, the best future prediction is usually some form of extrapolation of the historical behavior. For short-term forecasts, the behavior of interest is the short-term pattern of fluctuations. These fluctuations can be efficiently modeled by one of two methods:

- regression on recently-past values
- regression on past random errors

This chapter focuses mainly on these ARIMA (also known as Box-Jenkins) models, those that model a series based on lagged values of itself. Other possibilities for modeling this type of data include frequency analysis and extensions of standard regression techniques.

JMP also includes the ability to do transfer function models, where a time series is modeled using an input series. They are beyond the scope of this book, but are discussed in the *JMP Statistics and Graphics Guide*.

Introduction

When you take data from a process in the real world, the values can be a product of everything else that happened until the time of measurement. Time series data usually are non-experimental: they are the work of the world rather than the product of experimentally-controlled conditions. Furthermore, the data are not presented with all relevant covariates; they are often taken alone, without any other variables to co-analyze.

Time series methods have been developed to characterize how variables behave across time and how to forecast them into the future.

There are three general approaches to fitting a time series.

- Model the data as a function of time itself.
- Model the data as a function of its past (lagged) values.
- Model the data as a function of random noise.

The first approach emphasizes the structural part of the model, whereas the second and third approaches emphasize the random part of the model. The first approach is generally modeled using regression, which is covered in other chapters of this book. The second and third methods are the focus of this chapter.

Techniques of time series analysis implemented in JMP assume that the series is made up of points that are equally spaced in time. This doesn't mean that some of the values cannot be missing, but only that the data is collected on an regular schedule.

Models in this chapter are most useful for short-term predictions. For example, suppose you need to predict whether it will be raining one minute from now. Faced with this question, which information is more useful, what time it is, or if it is currently raining? If rain were structurally predictable (like a television schedule) you would rather know what time it is. Weather, however, does not behave in this fashion, so to predict if it will rain in a minute, it is more helpful to know if it is currently raining.

Lagged Values

This idea of looking into the past to forecast the future is the idea behind a *lagged variable*. Put simply, a lagged variable is formed by looking at values of the variable that occurred in the past.

As an example, examine a series of chemical concentration readings taken every two hours.

- Open the data file Seriesa.jmp[1].

The data set consists of only one variable. Make sure the column is named Cn, If it isn't,

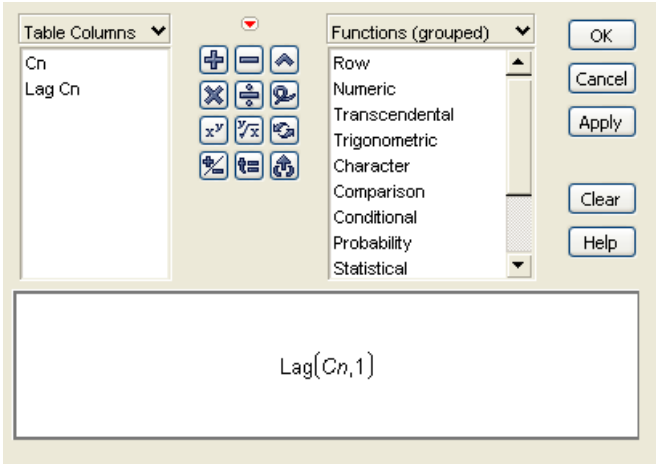
- Click in the column header to highlight it, then type the name Cn.

A lagged variable of Cn takes these values and shifts them forward or backward. Although it's not necessary for most time series analyses, we actually construct a variable that is lagged by one period for illustration.

- Add a new column to the data table and call it Lag Cn.
- Right-click (Control-click on the Macintosh) and select **Formula** from the menu that appears.
- Select **Row > Lag** from the list of functions.
- Click on the Cn column in the **Table Columns** list to specify that we are lagging the Cn variable.

The Formula Editor should now appear as in **Figure 20.1**.

Figure 20.1 Formula Editor for Lagged Variable

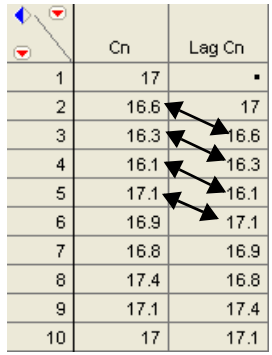


- Click **OK**.

1. The file is named SeriesA because it is the first data set presented in the classic time series book by Box and Jenkins. The book was so influential that time series modeling is often called Box-Jenkins analysis in their honor.

Figure 20.2 shows the results of the formula. Lag Cn is merely Cn shifted by one time period. We could have specified another lag period by adjusting the second argument of the `Lag()` function in the formula.

Figure 20.2 Illustration of a Lagged Variable



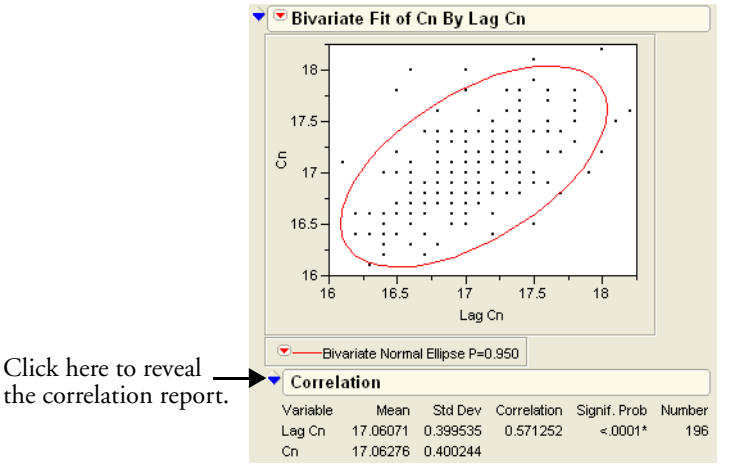
	Cn	Lag Cn
1	17	.
2	16.6	17
3	16.3	16.6
4	16.1	16.3
5	17.1	16.1
6	16.9	17.1
7	16.8	16.9
8	17.4	16.8
9	17.1	17.4
10	17	17.1

We can now examine whether Cn is correlated with its lagged values. A variable that is correlated with its own lagged values is said to be *autocorrelated*.

- Select **Analyze > Fit Y by X**.
- Assign Cn to **Y** and Lag Cn to **X**.
- Click **OK**.
- When the bivariate plot appears, select **Density Ellipse > 0.95**.
- Click the blue disclosure icon to open the correlation report.

The correlation coefficient is shown to be 0.5712. If there were no autocorrelation present, we would expect this value to be near zero. The presence of autocorrelation is an indicator that time series methods are necessary.

Figure 20.3 Bivariate Report Showing Autocorrelation



We are also interested in autocorrelations of lags greater than one. Rather than computing multiple lag columns, JMP provides a report that shows the correlation of many lagged values. To see this report,

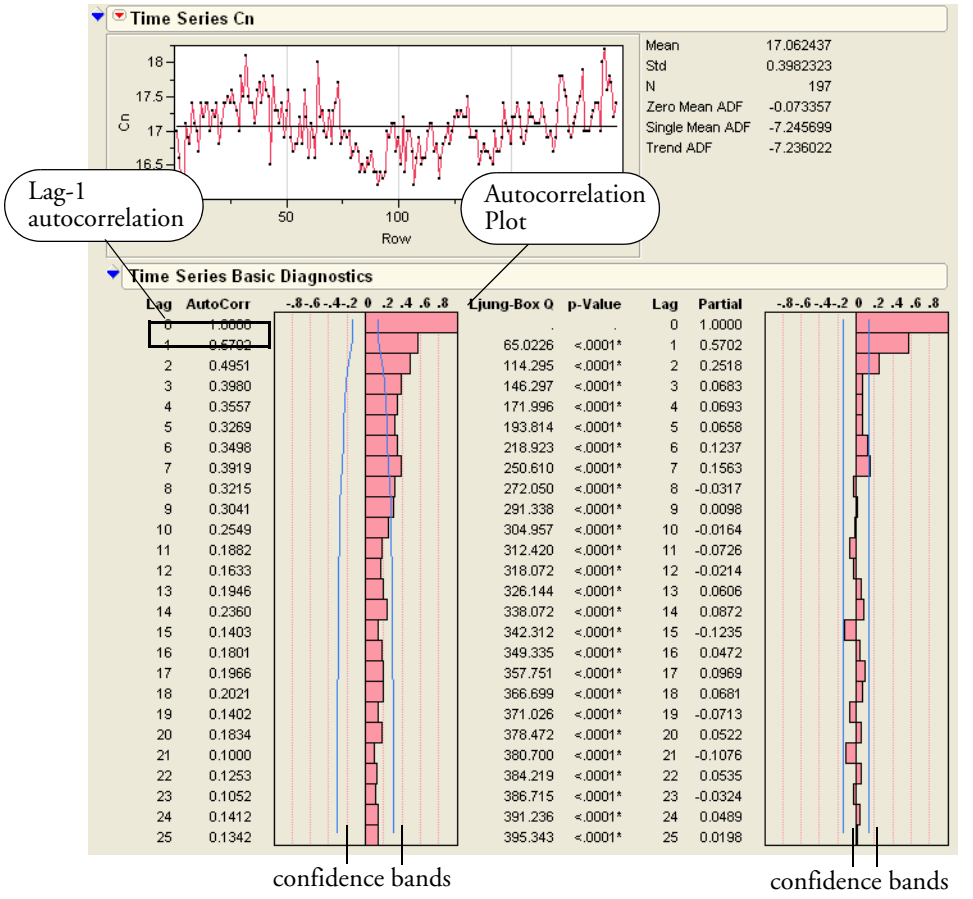
- Select **Analyze > Modeling > Time Series**.
- Assign Cn to the **Y, Time Series** role.
- Click **OK**.

The lag-1 correlation (shown in **Figure 20.4**) is close to the one found in the bivariate platform[1]. Values for higher lags are also shown. Note that the lag-0 value (each value of the variable correlated with itself) is exactly one. This plot is the fundamental tool used to determine the appropriate model for the data.

Notice the blue lines on the autocorrelation report. These are confidence intervals on the null hypothesis that the autocorrelation at each lag is zero. Bars that extend beyond these lines are evidence that the autocorrelation at that lag is not zero.

1. The correlation (called autocorrelation) in the Time Series platform is slightly different from the one reported in the Bivariate platform because the Time Series tradition uses n instead of $n–k$ as a divisor in calculating autocorrelations, where n is the length of the series, and k is the lag.

Figure 20.4 Initial Time Series Report



Testing for Autocorrelation

The recommended method for observing autocorrelation is to examine the autocorrelation plots from the Time Series platform. However, there is a statistical test for lag-1 autocorrelation known as the *Durbin-Watson* test. It is available in the Fit Model platform.

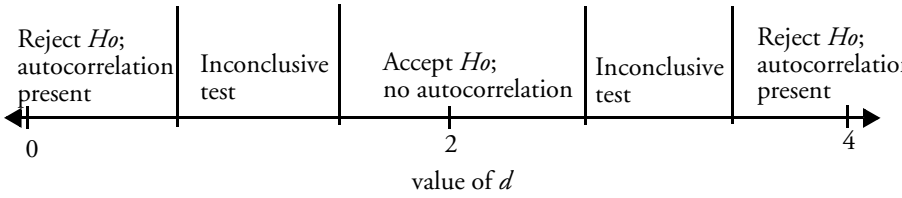
The Durbin-Watson test examines the difference between consecutive errors compared to the values of the errors themselves. Specifically, the test statistic is

$$d = \frac{\sum_t (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_t \hat{e}_t^2}$$

If you expand the numerator, this statistic becomes

$$d = \frac{\sum_t (e_t^2 - 2e_t\hat{e}_{t-1} + \hat{e}_{t-1}^2)}{\sum_t \hat{e}_t^2}$$

which allows you to see its value when there is no autocorrelation. With no autocorrelation, the middle term $2\hat{e}_t\hat{e}_{t-1}$ becomes zero, and the terms $\hat{e}_t^2$ and $\hat{e}_{t-1}^2$ combine to form $2\hat{e}_t^2$, so $d = 2$. If there is significant positive autocorrelation, d becomes quite small. Similarly, significant negative autocorrelation causes d to approach the value 4. With the null hypothesis that there is no correlation, then, we have a situation similar to the following.



The particular boundaries for each region depend on the values of the variables under consideration. Luckily, JMP can calculate the value and significance of the test automatically. Use the Seriesa.jmp data table as an example.

- If the Seriesa data table is not front-most, bring it to the front using the **Window** menu.
- Select **Analyze > Fit Model** and assign Cn to the **Y** role.
- Click **Run Model**.

When the report appears,

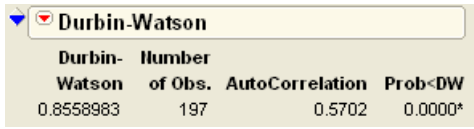
- Select **Row Diagnostics > Durbin Watson** from the platform popup menu.

When the Durbin-Watson report appears,

- Select **Significance P Value** from the Durbin Watson popup menu.
- When JMP warns you that this may take a long time (it won't), click **OK**.

You should now have the report in **Figure 20.5**.

Figure 20.5 Durbin-Watson Report



Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
0.8558983	197	0.5702	0.0000*

The low *p*-value indicates that there is significant autocorrelation in this data. Note, however, that the Durbin-Watson test examines only first-order (lag-1) autocorrelation, so if there is autocorrelation beyond lag 1, this test provides no information.

White Noise

The most fundamental of all time series is called *white noise* (so termed because values tend to enter at all frequencies, similar to white light). The white noise series is a series composed of values chosen from a random Normal distribution with mean zero and variance one. It should also have autocorrelation of (near) zero for all lags. To construct this series in JMP,

- Select **File >New > Data Table**.

In the data table that appears,

- Right-click (Control-click on the Macintosh) on the column header and select **Formula**.

In the Formula Editor,

- Select **Random > Random Normal**.
- Click **OK.**
- Select **Rows > Add Rows** and add 100 rows to the data table.

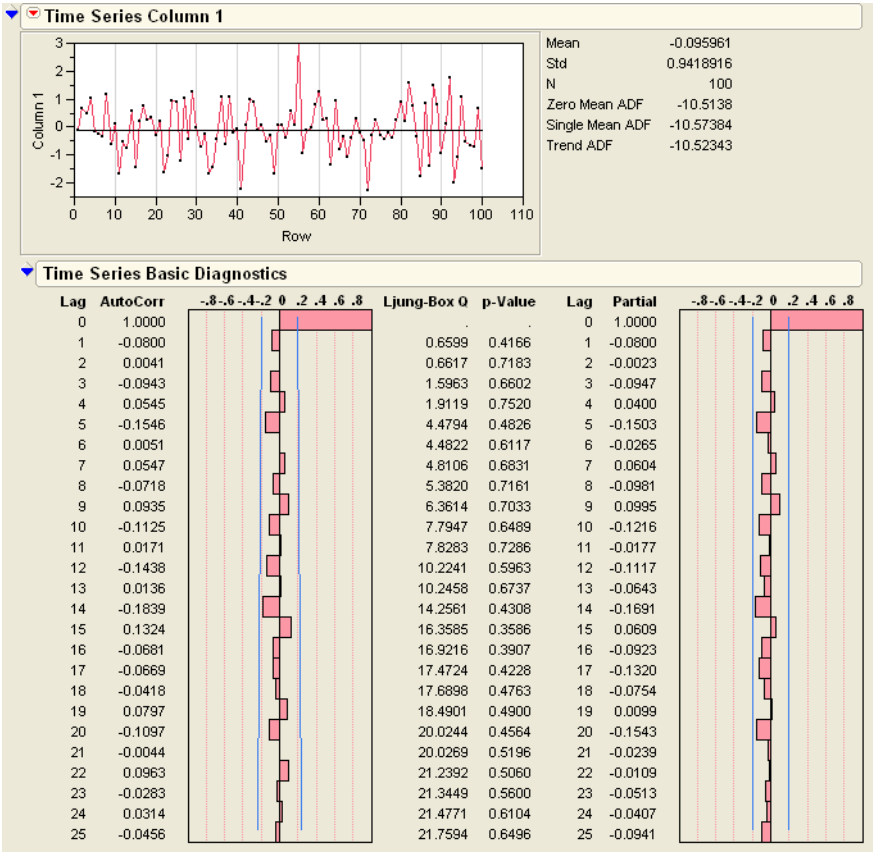
You should now have a single column that contains 100 random values.

- Select **Analyze > Modeling > Time Series.**
- Assign the column to the **Y, Time Series** role.
- Click **OK**.

None of the bars in the autocorrelation report extend beyond the blue confidence lines and the mean and standard deviation are as expected. None of the correlations appear to be

different from zero. You should remember the look of the correlation plots for future reference, since reducing a time series to white noise is part of the modeling procedure.

Figure 20.6 White Noise Report



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.0800	0.6599	0.4166	1	-0.0800
2	0.0041	0.6617	0.7183	2	-0.0023
3	-0.0943	1.5963	0.6602	3	-0.0947
4	0.0545	1.9119	0.7520	4	0.0400
5	-0.1546	4.4794	0.4826	5	-0.1503
6	0.0051	4.4822	0.6117	6	-0.0265
7	0.0547	4.8106	0.6831	7	0.0604
8	-0.0718	5.3820	0.7161	8	-0.0981
9	0.0935	6.3614	0.7033	9	0.0995
10	-0.1125	7.7947	0.6489	10	-0.1216
11	0.0171	7.8283	0.7286	11	-0.0177
12	-0.1438	10.2241	0.5963	12	-0.1117
13	0.0136	10.2458	0.6737	13	-0.0643
14	-0.1839	14.2561	0.4308	14	-0.1691
15	0.1324	16.3585	0.3586	15	0.0609
16	-0.0681	16.9216	0.3907	16	-0.0923
17	-0.0669	17.4724	0.4228	17	-0.1320
18	-0.0418	17.6898	0.4763	18	-0.0754
19	0.0797	18.4901	0.4900	19	0.0099
20	-0.1097	20.0244	0.4564	20	-0.1543
21	-0.0044	20.0269	0.5196	21	-0.0239
22	0.0963	21.2392	0.5060	22	-0.0109
23	-0.0283	21.3449	0.5600	23	-0.0513
24	0.0314	21.4771	0.6104	24	-0.0407
25	-0.0456	21.7594	0.6496	25	-0.0941

Autoregressive Processes

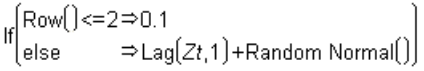
Suppose you construct a series like the following.

$$x_1 = 1$$
$$x_n = x_{n-1} + \varepsilon$$

where ε is random white noise. Such a series is called a *random walk*.

To create this simple random walk,

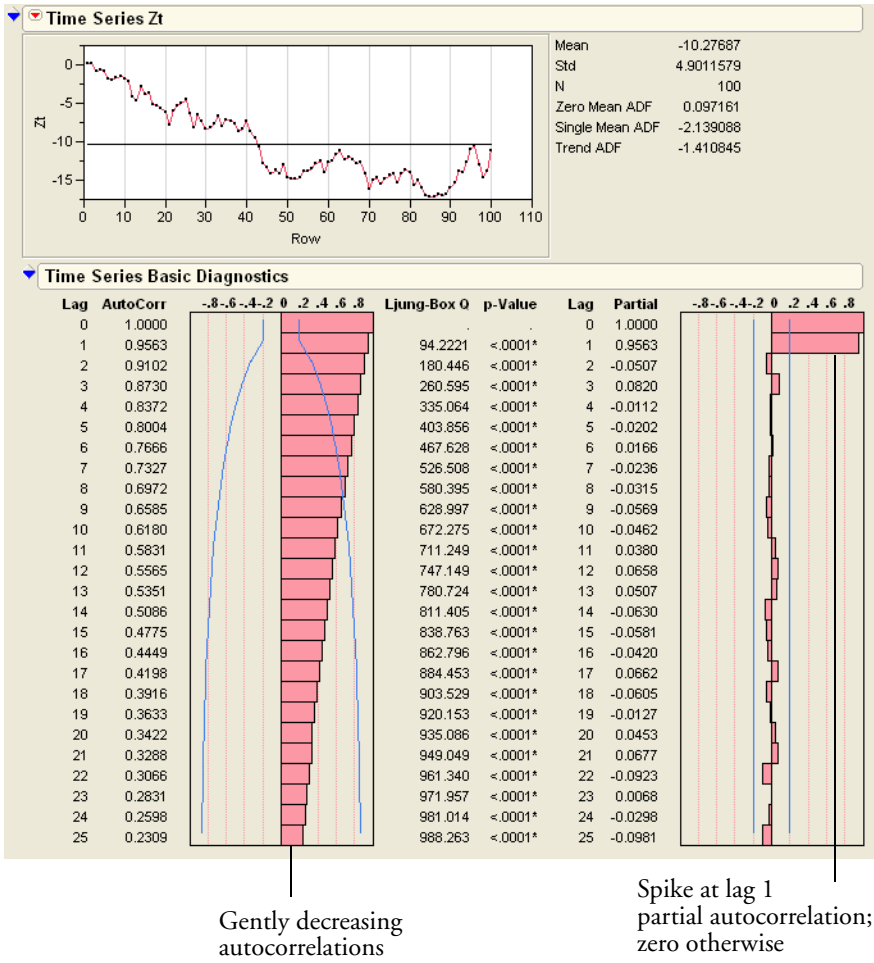
- Select **File > New > New Data Table**.
- Name the data table's column Zt.
- Enter the following formula into this new column:



- Add 100 rows to the column.

Since each value is just a slight perturbation of its previous value, we expect to see significant lag-1 autocorrelation, slightly less lag-2 correlation, even less lag-3 correlation, and so on. **Figure 20.7** shows the time series report from this series.

Figure 20.7 Random Walk Report



The random walk is the simplest example of an *autoregressive process*. Each value is the sum of the previous value and random noise. Written mathematically, if z_t represents the value of the series at time t,

$$z_t = \varphi z_{t-1} + \varepsilon$$

where φ represents the influence of the lag value on the present value and ε represents random (white noise) error. In other words, the series is composed of two parts—a fraction of its lag-1 values and noise. As shorthand, we refer to this as an AR(1) ("autoregressive lag-1") process.

More generally, if we use *p* to represent the number of lags, autoregressive processes of order *p* are written AR(*p*).

The plot to the right of the autocorrelation plot in **Figure 20.7** is called the *partial autocorrelation plot*.This plot shows the autocorrelations at several lag values *after all lower-valued lagged autocorrelations are taken into account*. In this example, there is a spike at lag 1, indicating that the model should contain a lag-1 term. All other partial autocorrelations aren't significantly different than zero, indicating that we don't need further autocorrelation terms. None of this is a surprise since we constructed this series out of lag-1 values.

Correlation Plots of AR Series

The pattern in **Figure 20.7** is a typical example of an AR(*p*) process. The partial autocorrelation plot has spikes at lags 1 to *p* and is zero for lags greater than *p*. The autocorrelation plot shows a gently decreasing pattern.

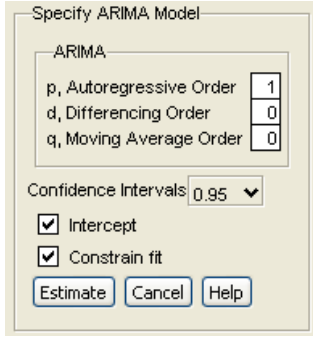
Estimating the Parameters of an Autoregressive Process

Our objective is to find the value of φ that generated the model. To estimate this parameter in JMP,

- Select **ARIMA** from the platform popup menu.

This produces the dialog shown in **Figure 20.8**.

Figure 20.8 Specify ARIMA Dialog



- Enter a 1 beside **p, Autoregressive Order**.

Click **Estimate**.

This produces a report similar to the one in **Figure 20.9** (remember, this is random data, so your results certainly vary from those here).

Figure 20.9 AR(1) Report

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate
AR1	1	0.984439	0.014568	67.57	<.0001*	-0.1185463
Intercept	0	-7.618137	4.448756	-1.71	0.0900	

The estimated lag-1 autoregressive coefficient is 0.98, very close to the actual value of 1. After accounting for this AR(1) term, all that should be left is white noise. To verify this,

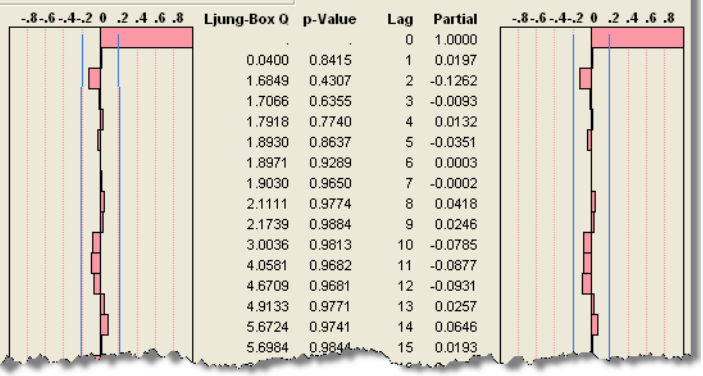
Click on the disclosure icon beside **Residuals**.

This reveals a time series report of the residuals of the model. Since there are no significant autocorrelations or partial autocorrelations, we conclude that our residuals are indistinguishable from white noise, so the AR(1) model is adequate.

Figure 20.10 Residuals Report

Residuals

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.0197	0.0400	0.8415	1	0.0197
2	-0.1257	1.6849	0.4307	2	-0.1262
3	-0.0144	1.7066	0.6355	3	-0.0093
4	0.0283	1.7918	0.7740	4	0.0132
5	-0.0307	1.8930	0.8637	5	-0.0351
6	-0.0061	1.8971	0.9289	6	0.0003
7	0.0073	1.9030	0.9650	7	-0.0002
8	0.0433	2.1111	0.9774	8	0.0418
9	0.0237	2.1739	0.9884	9	0.0246
10	-0.0856	3.0036	0.9813	10	-0.0785
11	-0.0959	4.0581	0.9682	11	-0.0877
12	-0.0727	4.6709	0.9681	12	-0.0931
13	0.0455	4.9133	0.9771	13	0.0257
14	0.0800	5.6724	0.9741	14	0.0646
15	0.0147	5.6984	0.9844	15	0.0193



Moving Average Processes

A *moving average* (MA) process models a time series on lagged values of a white noise series rather than lagged values of the series itself. In other words, MA series listen to white noise and derive their values from what they hear. Using q to designate the number of lags involved in the moving average, we abbreviate "moving average of q lags" as MA(q).

To see a simple example of an MA series,

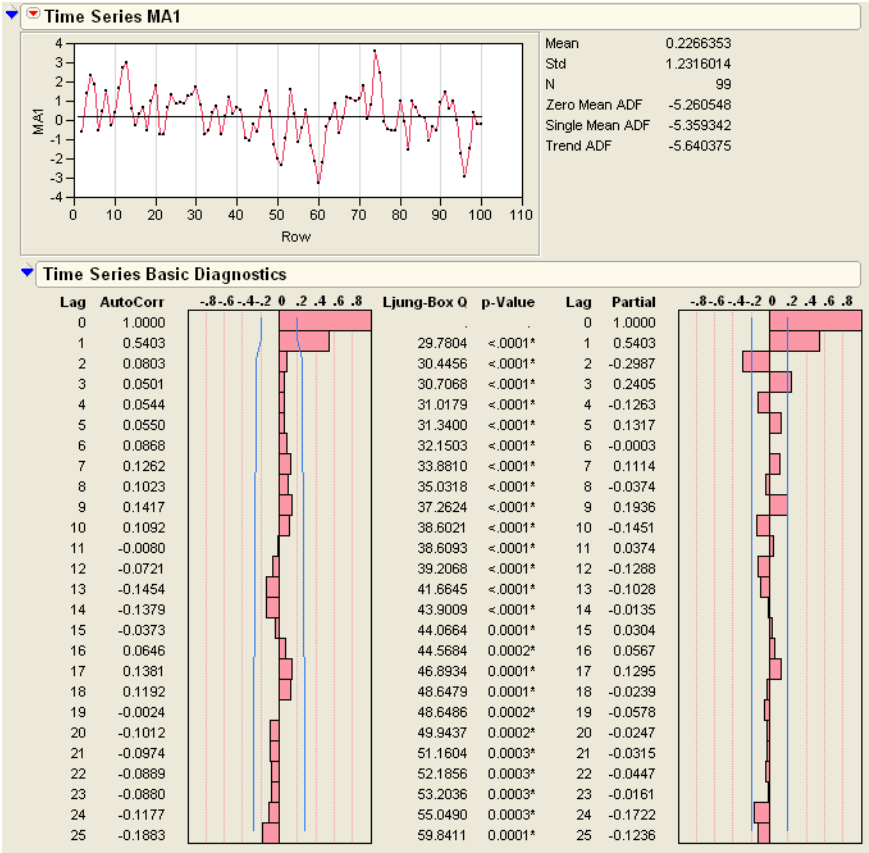
- Open Timema1.jmp.
- Add 100 rows to the data table.

There are two columns in the data table. The Noise column holds simple white noise, generated with a formula. The MA1 column is computed using lagged values of the white noise. Examine a simple time series report of the value of MA1.

- Select **Analyze > Modeling > Time Series** and use MA1 as the **Y, Time Series** variable.

This produces a report similar to the one in **Figure 20.11**.

Figure 20.11 MA(1)Report



Correlation Plots of MA Series

Contrast the autocorrelation plot of this moving average process with the one from the autoregressive process of the last section. For an MA(q) series, the autocorrelation plot goes to zero after q lags, and the partial autocorrelation plot drops off slowly. In this case, "drops off slowly" means the *magnitudes* of the autocorrelations drop off slowly, even though they alternate in sign. Since this is an MA(1) series, we see a spike in the autocorrelation plot at lag 1, and spikes that are (essentially) zero everywhere else.

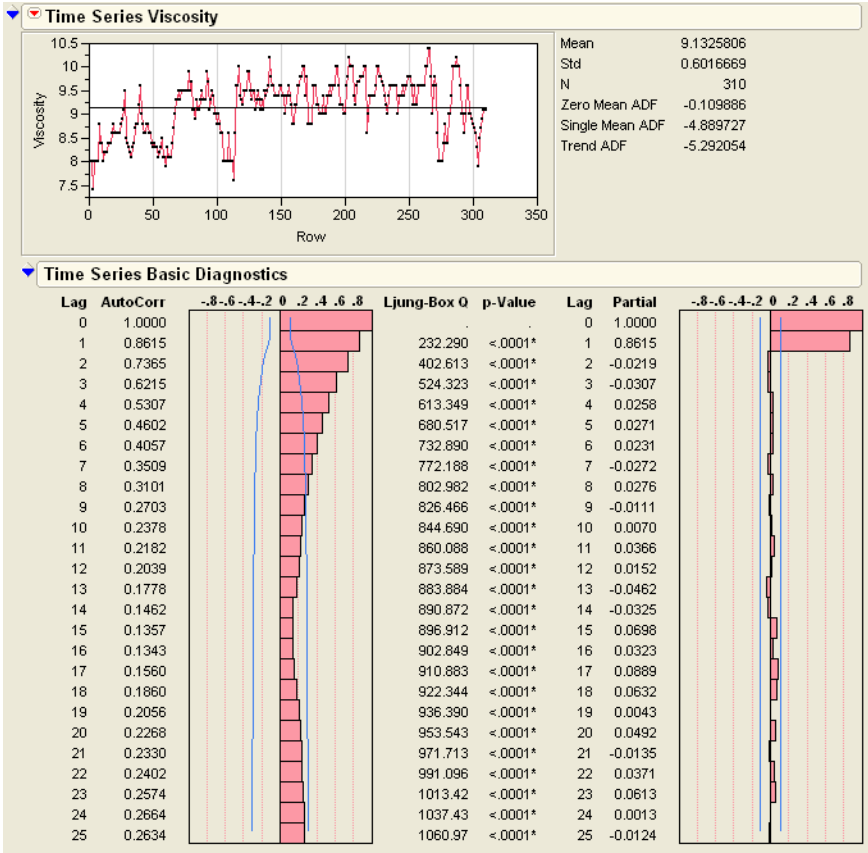
Example of Diagnosing a Time Series

The data set SeriesD.jmp (Box, Jenkins, and Reinsel 1994) contains viscosity readings from a chemical process. Using the autocorrelation and partial autocorrelation functions, we try to determine the nature of the model.

- Open the data file SeriesD.jmp.
- Select **Analyze > Modeling > Time Series**.
- Assign Viscosity to the **Y, Time Series** role.
- Click **OK**.

Examining the plots, it is easy to see that the autocorrelation plot decreases slowly, while the partial autocorrelation plot has a spike at lag 1, but is (essentially) zero everywhere else. This is the condition we saw for an AR(p) process ("Correlation Plots of AR Series" on page 522). Therefore, we guess that an AR(1) model is appropriate.

Figure 20.12 SeriesD Report

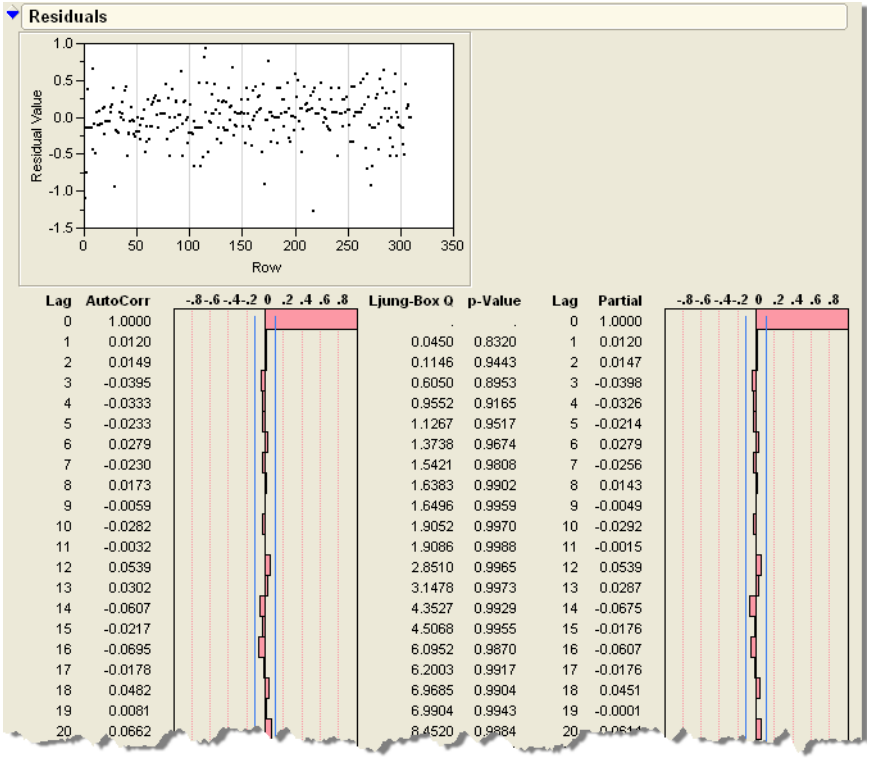


To estimate the parameters of the model,

- Select **ARIMA** from the platform menu.
- Enter a 1 in the **p, Autoregressive Order** box.
- Click **Estimate**.

The resulting residuals (and their correlation plots) look like white noise, so we are satisfied with the model we have chosen.

Figure 20.13 SeriesD Residual Report



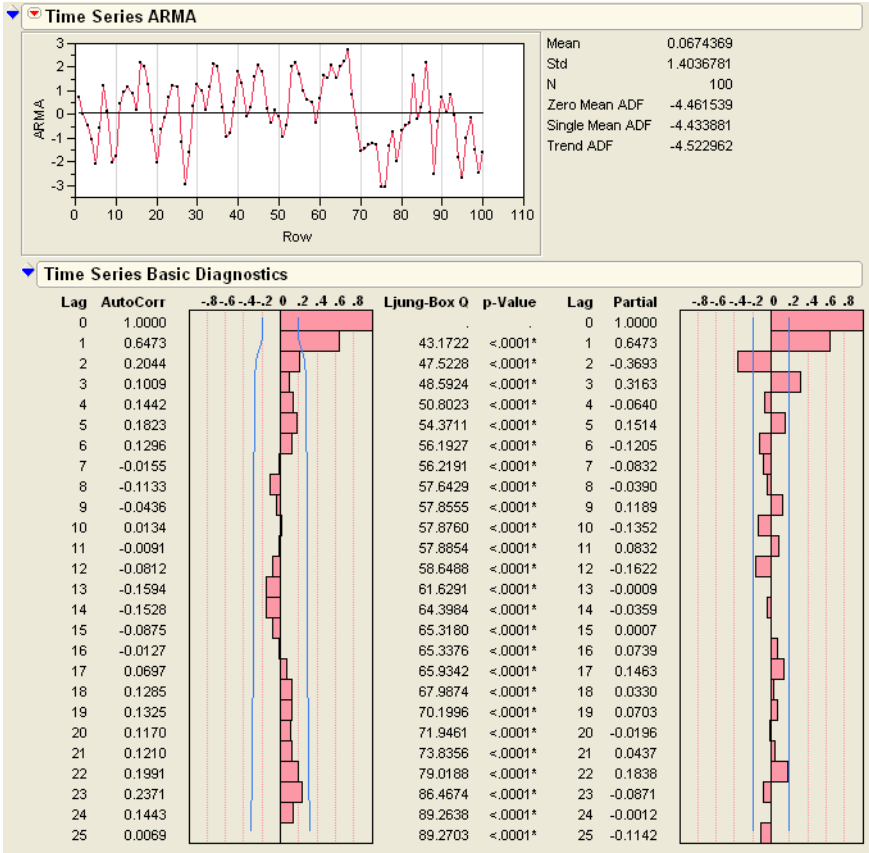
Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.0120	0.0450	0.8320	1	0.0120
2	0.0149	0.1146	0.9443	2	0.0147
3	-0.0395	0.6050	0.8953	3	-0.0398
4	-0.0333	0.9552	0.9165	4	-0.0326
5	-0.0233	1.1267	0.9517	5	-0.0214
6	0.0279	1.3738	0.9674	6	0.0279
7	-0.0230	1.5421	0.9808	7	-0.0256
8	0.0173	1.6383	0.9902	8	0.0143
9	-0.0059	1.6496	0.9959	9	-0.0049
10	-0.0282	1.9052	0.9970	10	-0.0292
11	-0.0032	1.9086	0.9988	11	-0.0015
12	0.0539	2.8510	0.9965	12	0.0539
13	0.0302	3.1478	0.9973	13	0.0287
14	-0.0607	4.3527	0.9929	14	-0.0675
15	-0.0217	4.5068	0.9955	15	-0.0176
16	-0.0695	6.0952	0.9870	16	-0.0607
17	-0.0178	6.2003	0.9917	17	-0.0176
18	0.0482	6.9685	0.9904	18	0.0451
19	0.0081	6.9904	0.9943	19	-0.0001
20	0.0662	8.4520	0.9884	20	0.0614

ARMA Models and the Model Comparison Table

We cannot always model time series using AR or MA models alone. There may be some terms from both types of models. Notationally, we call these *ARMA* models and designate them as ARMA(p, q) where p represents the autoregressive order and q represents the moving average order. When either of these numbers are zero, we drop their corresponding letters from the ARMA designation. So, for example, and ARMA(1,0) model is written merely as AR(1). As an example,

- Open the data table TimeARMA.jmp.
- Add 100 rows to the data table.
- Produce a time series report of the ARMA variable.

Figure 20.14 TimeARMA Report



It is not immediately clear what order AR and MA coefficients to use, since neither autocorrelation plot has a familiar pattern. Both appear to drop to zero after a few lags, yet they both have spikes at later lags as well. If you ignore the later lags, the partial autocorrelation plot suggests an AR(2). Similarly, the autocorrelation plot suggests an MA(1). Since there is no clear-cut model, we fit several models and examine their residuals and fit statistics to find a suitable model.

- Fit an MA(1), MA(2), AR(1), AR(2), and ARMA(1,1) model to the data.

At the top of the report, you find a model comparison table, as shown in **Figure 20.15**. As each model is fit, its corresponding statistics are appended to this table. These statistics give clues as to which model is best-fitting.

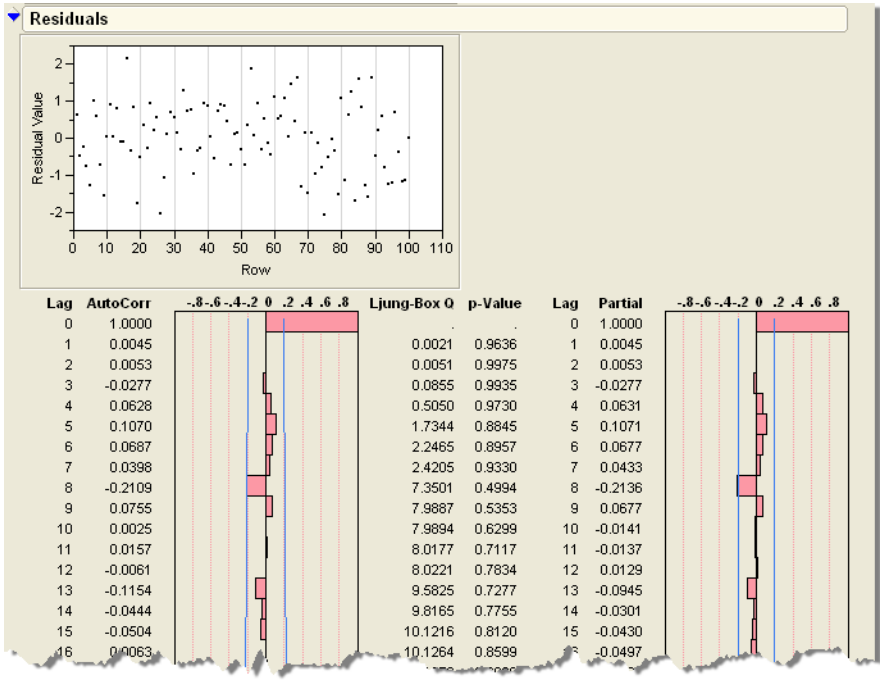
Figure 20.15 Model Comparison Table

Model Comparison

Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	AIC Rank	SBC Rank	MAPE	MAE
MA(1)	98	0.9350051	280.42180	285.63214	0.532	276.4218	3	3	250.63092	0.798975
MA(2)	97	0.8873211	276.15428	283.96979	0.561	270.15428	2	2	297.59263	0.764022
AR(1)	98	1.1541826	300.65899	305.86933	0.425	296.65899	5	5	312.16332	0.877487
AR(2)	97	1.008064	288.38276	296.19827	0.503	282.38276	4	4	326.46408	0.809075
ARMA(1,1)	97	0.8782015	275.17795	282.99346	0.565	269.17795	1	1	307.85674	0.757383

In general, higher values of R^2 are desirable, and lower values of AIC and SBC are desirable. The rank of each model is listed under the AIC Rank and SBC rank columns. So, for this example, the ARMA(1,1) model is the best choice among the candidates. This decision is supported by the residual plot for the ARMA(1,1) model, which looks fairly close to white noise.

Figure 20.16 TimeARMA Residuals



Stationarity and Differencing

In all the examples so far, the series have had no noticeable linear trend. They seem to hover around a constant mean, and they do not show an overall increase or decrease as time

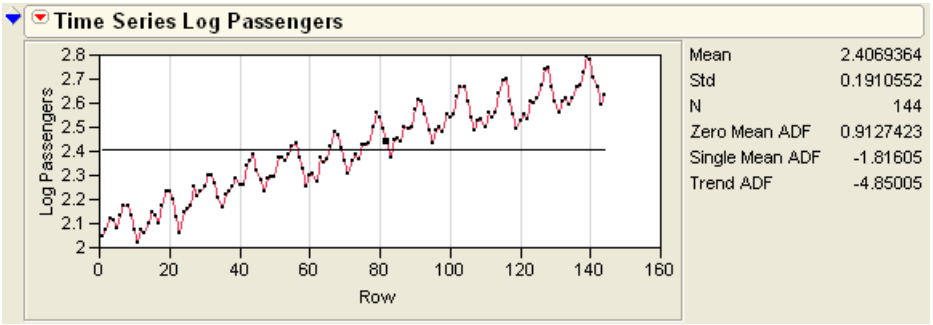
progresses. The methods illustrated so far apply only to these *stationary* models. If a model exhibits a trend over time, it is termed *non-stationary*, and we must transform the data to attempt to remove the trend.

As an example,

- Open the data table Seriesg.jmp.
- Select **Analyze > Modeling > Time Series** and designate log passengers as the **Y, Time Series** variable.

The time series plot (**Figure 20.17**) shows a definite trend of increasing passengers over time.

Figure 20.17 Seriesg Plot



One approach to compensating for a trend is to *difference* the series. We compute a new series that is the difference between each value and its lag-1 value.

- Create a new column in the data table called Difference.
- Enter the following formula into the column.

$$Log\ Passengers - Log\ Passengers_{Row()-1}$$

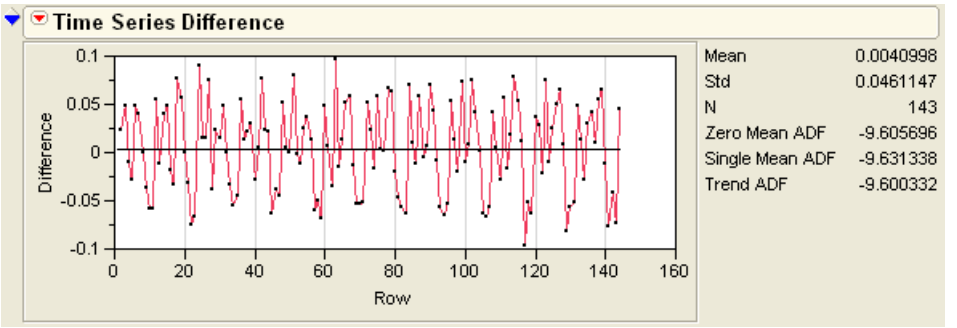
Note: To enter a subscript, select **Subscript** from the **Row** group of formulas.

When you apply this formula, an error message appears because we are attempting to difference the first row with the (non-existent) zero-th row. Click **OK** to dismiss the error and the rest of the column fills in as desired.

- Produce a time series report of the new Difference variable.

As shown in **Figure 20.18**, the trend has disappeared.

Figure 20.18 Differenced Variable



Differencing is a common task in time series analysis, so JMP contains an automatic **Difference** command in the report drop-down menu.

- From the original Log Passengers report, select **Difference**.
- In the dialog that appears, set the differencing order to 1.
- Click **OK**.

You can see from the plot that this command performs the same function as our formula. As seen from the choices in the menu, differencing can be performed to several levels of lag, not just lag-1 as in this example.

When a model requires differencing, it is called an ARIMA model (where the I stands for *Integrated*). Symbolically, we write the model as ARIMA (p, d, q), where p and q have the same functions as before, and d represents the order of differencing.

With differencing, the Seriesg process becomes stationary, and the other methods of this chapter can be used to determine the form of the model.

Seasonal Models

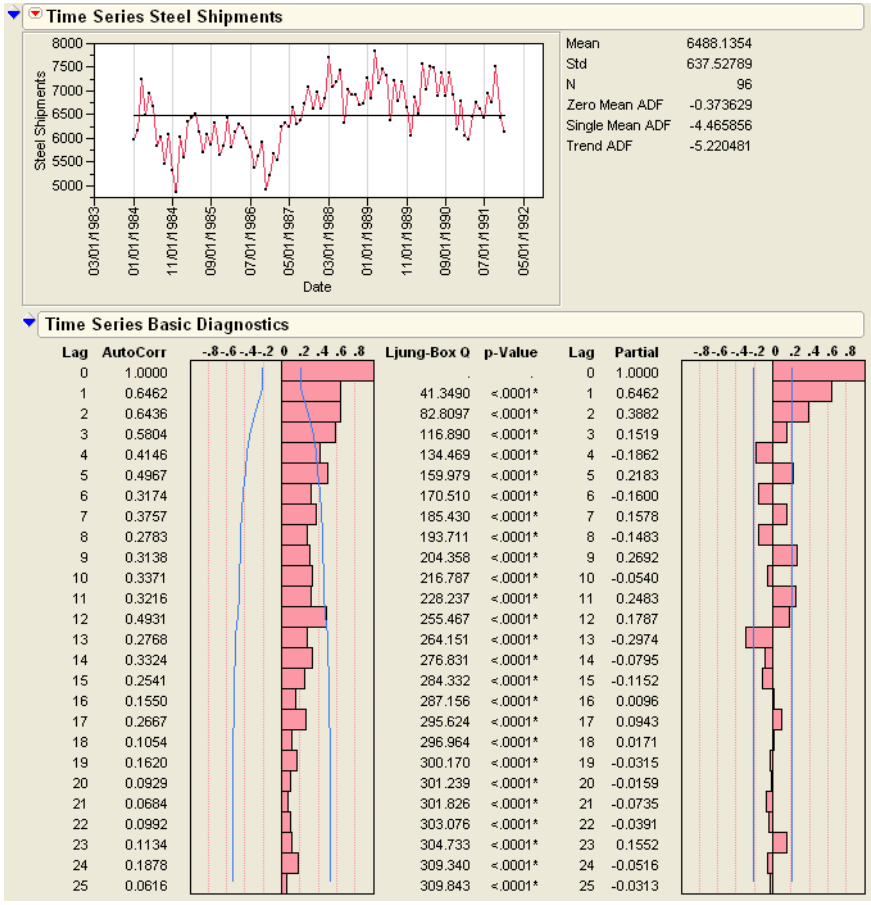
Frequently, time series data are affected by a season—often, corresponding to seasons of the year or to business cycles. To model these seasonal factors, autoregressive and moving average factors of large lags are examined. For example, for yearly data, lags of length 12, 24, 36, and so on are often examined in addition to the nonseasonal terms.

As an example of a seasonal process, examine the file Steel Shipments.jmp. This file contains data on the monthly amount of steel shipped from U.S. steel mills from 1984 to 1991. From past experience, the researchers who collected this data expect a yearly cycle.

- Open the file Steel Shipments.jmp.
- Select **Analyze > Modeling > Time Series**.
- Assign Steel Shipments to the **Y, Time Series** role and Date as the **X, Time ID** role.
- Click **OK**.

You should see the report in **Figure 20.19**.

Figure 20.19 Steel Shipments Initial Report

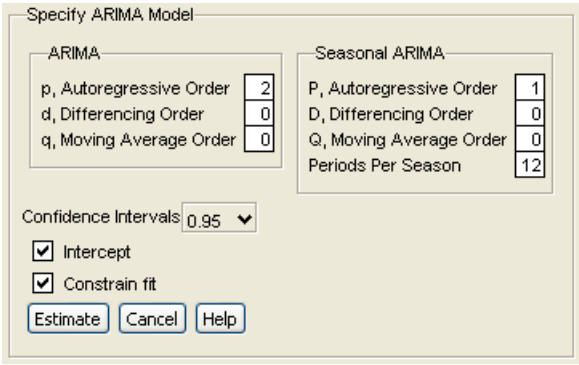


Note there is a spike in the autocorrelation plot near lag 12. This is a clue that there is a yearly seasonal element to the data. Since there is no corresponding spike at lag 24, the series may be seasonally stationary. The partial autocorrelation plot indicates that the nonseasonal component may be AR(2).

JMP designates seasonal components of an ARIMA model in a second set of parentheses listed after the nonseasonal components. Generally, the model is designated as ARIMA(p, d, q)(P, D, Q) where the capital letters indicate the order of the seasonal terms. In addition, the order of the seasonal terms (12 for a year-long monthly season, as in this example) is written after the model. So, for this example, we guess that an ARIMA(2, 0, 0)(1, 0, 0)12 model is appropriate in this case. To fit this model,

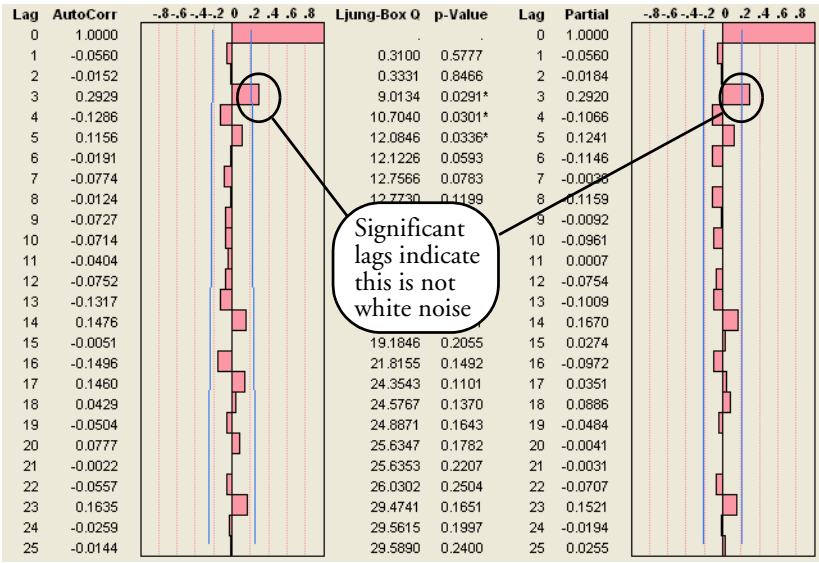
- Select **Seasonal ARIMA** from the platform drop-down menu.
- Fill out the resulting dialog so that it appears as in **Figure 20.20**.

Figure 20.20 Seasonal ARIMA Dialog



The resulting correlation plots (**Figure 20.21**) show that we do not have an adequate model, since the plots do not have the characteristics of white noise.

Figure 20.21 Correlation Plots



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.0560	0.3100	0.5777	1	-0.0560
2	-0.0152	0.3331	0.8466	2	-0.0184
3	0.2929	9.0134	0.0291*	3	0.2920
4	-0.1286	10.7040	0.0301*	4	-0.1066
5	0.1156	12.0846	0.0336*	5	0.1241
6	-0.0191	12.1226	0.0593	6	-0.1146
7	-0.0774	12.7566	0.0783	7	-0.0038
8	-0.0124	12.7730	0.1199	8	0.1159
9	-0.0727			9	-0.0092
10	-0.0714			10	-0.0961
11	-0.0404			11	0.0007
12	-0.0752			12	-0.0754
13	-0.1317			13	-0.1009
14	0.1476			14	0.1670
15	-0.0051	19.1846	0.2055	15	0.0274
16	-0.1496	21.8155	0.1492	16	-0.0972
17	0.1460	24.3543	0.1101	17	0.0351
18	0.0429	24.5767	0.1370	18	0.0886
19	-0.0504	24.8871	0.1643	19	-0.0484
20	0.0777	25.6347	0.1782	20	-0.0041
21	-0.0022	25.6353	0.2207	21	-0.0031
22	-0.0557	26.0302	0.2504	22	-0.0707
23	0.1635	29.4741	0.1651	23	0.1521
24	-0.0259	29.5615	0.1997	24	-0.0194
25	-0.0144	29.5890	0.2400	25	0.0255

The autocorrelation plot of the seasonal model has a spike at lag 3, indicating a possible MA(3) portion of the model.

- As another candidate, try an ARIMA(2, 0, 3)(1, 0, 0)12 model, resulting in the report shown in **Figure 20.22**.

Figure 20.22 ARIMA(2,0,3)(1,0,0)12 Model

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.0005	0.0000	0.9962	1	0.0005
2	-0.0257	0.0663	0.9674	2	-0.0257
3	0.0539	0.3597	0.9484	3	0.0539
4	-0.0105	0.3710	0.9848	4	-0.0113
5	0.0792	1.0192	0.9610	5	0.0823
6	0.0184	1.0547	0.9835	6	0.0146
7	-0.0665	1.5221	0.9815	7	-0.0615
8	0.0024	1.5227	0.9923	8	-0.0053
9	-0.0680	2.0221	0.9911	9	-0.0726
10	-0.0351	2.1571	0.9950	10	-0.0346
11	-0.0646	2.6191	0.9949	11	-0.0735
12	-0.0796	3.3289	0.9927	12	-0.0659
13	-0.0737	3.9444	0.9918	13	-0.0755
14	0.1303	5.8935	0.9691	14	0.1432
15	0.0010	5.8936	0.9815	15	0.0128
16	-0.1409	8.2290	0.9418	16	-0.1279
17	0.1112	9.7003	0.9157	17	0.1141
18	0.0499	10.0006	0.9319	18	0.0494
19	0.0090	10.0105	0.9527	19	-0.0080
20	-0.0092	10.0209	0.9678	20	-0.0501
21	0.0088	10.0306	0.9785	21	0.0338
22	-0.0569	10.4422	0.9818	22	-0.0987
23	0.1634	13.8849	0.9301	23	0.1591
24	-0.0367	14.0608	0.9453	24	-0.0457
25	0.0026	14.0617	0.9606	25	0.0127

These plots appear to be white noise, so we are satisfied with the fit. Notice how we tried a model, evaluated it, then tried another. This iterative method of refining the model is common; don't be surprised if several stages are required to find an adequate model.

Spectral Density

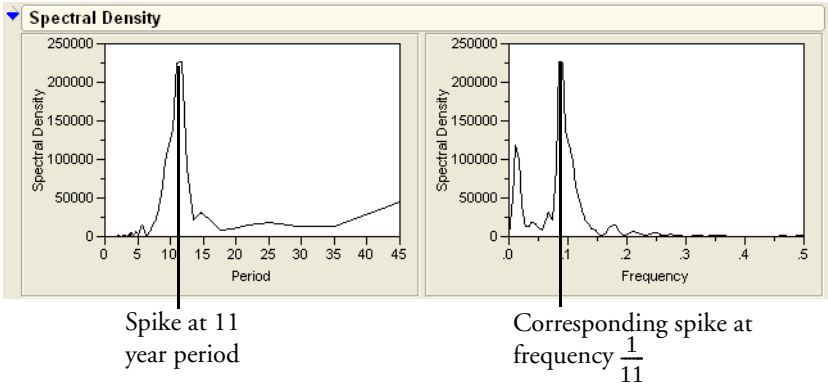
The *spectral density plot* decomposes a series into frequencies, much like a prism does to a spectrum of white light. Spikes at certain frequencies of the spectral density plot give an indication of repeating cycles.

As an example,

- Open Wolfer Sunspot.jmp.
- Generate a time series report on the wolfer variable.
- Select **Spectral Density** from the platform drop-down menu.

You should see plots as in **Figure 20.23**.There is an obvious spike at a period of 11 years (corresponding to the now-known 11-year sunspot cycle). Since frequency is the reciprocal of period, there is a corresponding spike at frequency 1/11.

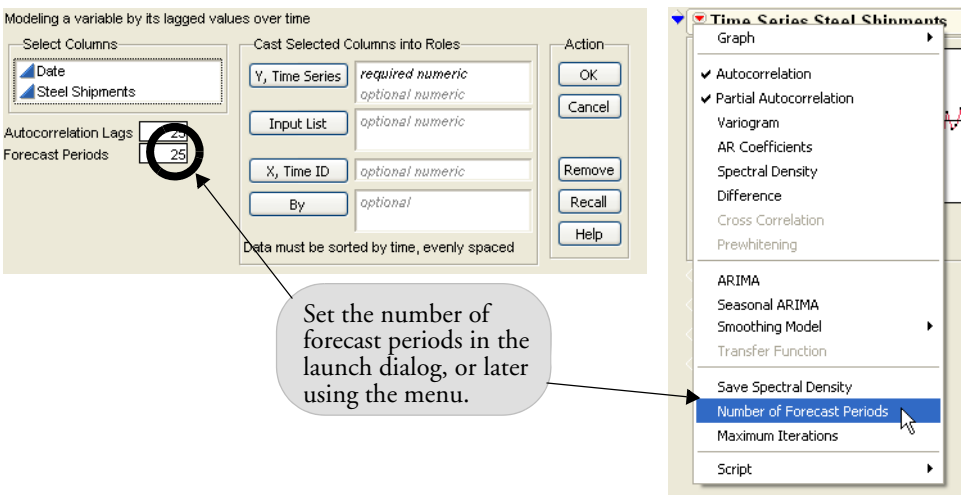
Figure 20.23 Spectral Density Plots



Forecasting

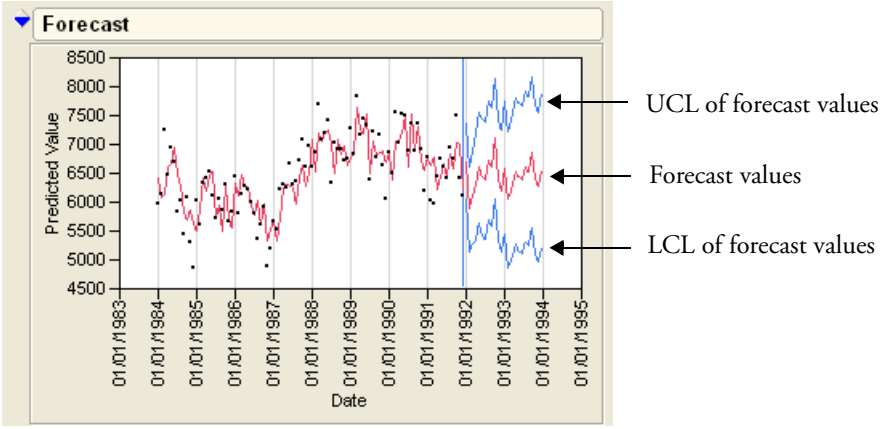
Frequently, the purpose of fitting an ARIMA model to a time series is to forecast values into the future. In JMP, you must specify the number of periods into the future that you want forecasted. This number can be specified in the Time Series launch dialog (before the report is produced) or from the menu at the top level of the Time Series report (after the report is produced).

Figure 20.24 Changing the Number of Forecast Periods



For example, the Steel Shipments.jmp data with the seasonal ARIMA(2,0,3)(1,0,0)12 model has the following Forecast graph. Because the defaults were used when generating this report, forecasts extend 25 periods into the future. Confidence limits on the forecasts are also shown.

Figure 20.25 Forecast Graph

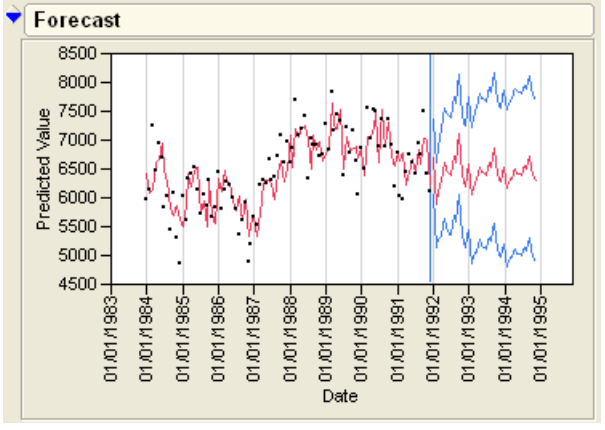


To add more periods to the forecast,

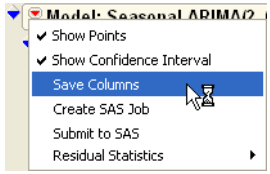
- Select **Number of Forecast Periods** from the platform drop-down as shown in **Figure 20.24**.
- Enter 36 in the dialog that appears.

The report now forecasts 36 periods into the future.

Figure 20.26 Forecast with 36 Periods



Aside from these graphical presentations of the forecasts, JMP can save the values to a new data table.



Select **Save Columns** from the drop down menu on the model's outline bar.

This produces a new JMP table containing values for

- the time series itself
- the time variable, expressed in seconds since January 1, 1904
- predicted values for the time series and their standard errors
- residuals, representing the difference between the actual and predicted values
- upper and lower confidence intervals for each point

This data table is often used for follow-up analyses on the series. For example, you can check the distribution of residuals (using the **Analyze > Distribution** command) or make overlay plots of the actual values with their confidence limits.

Exercises

1. The Wolfer Sunspot.jmp data table was used as an example earlier in this chapter. Suppose two analysts conduct a time series analysis of this data, with one settling on an AR(2) model, while the other settles on an AR(3) model. Comment on their conclusions.
2. SeriesF.jmp is from the classic Box and Jenkins text. It can be sufficiently modeled with an ARMA model. Determine the AR and MA coefficients.
3. Simulated.jmp is a simulated time series. Find the ARIMA model used to generate it.
4. In order to test an automatic atomizer, Antuan Negiz (1994, "Statistical Monitoring and Control of Multivariate Continuous Processes") collected data from an aerosol mini-spray dryer device. The dryer injects a slurry at high speed. A by-product of his analysis was small dried particles left in his machine. The data from his study is stored in Particle Size.jmp. Fit an ARIMA model to the data.
5. The file RaleighTemps.jmp contains the average daily temperature for Raleigh, North Carolina, from January 1980 to December 1990. Temperature data have an obvious seasonal component to them. Determine this seasonal component and fit an ARIMA model to the data.
6. Nile.jmp consists of annual flow volumes of the Nile river at Aswan from 1871 to 1970.
   (a) Does there appear to be a trend in this data?

(b) Fit an appropriate ARIMA model.

7. The Southern Oscillation Index is defined as the barometric pressure difference between Tahiti and the Darwin Islands at sea level. It is considered as an indicator of the El Niño effect, which affects fish populations and has been blamed for floods in the midwestern United States. The data are stored in Southern Oscillation.jmp. Use a Spectral Density plot to determine if there are cycles in the El Niño effect. (Scientists recognize one at around 12 months, as well as one around 50 months—slightly more than four years).

8. This exercise examines moving average processes stored in Moving Average.jmp.

   (a) Examine the formulas in the columns Y1 and Y2 and determine the MA(1) coefficients.

   (b) Conduct a time series analysis on Y1 and Y2 and comment on what you see. Examine a time series text for a section on invertability to explain your findings.

21 Machines of Fit

Overview

This chapter is an essay on fitting for those of you who are mechanically inclined. If you have any talent for imagining how springs and tire pumps work, you can put it to work here in a fantasy in which all the statistical methods are visualized in simple mechanical terms.

The goal is to not only remember how statistics works, but also train your intuition so you will be prepared for new statistical issues.

Here is an illuminating trick that will help you understand and remember how statistical fits really work. It involves pretending that statistical fitting is performed by machines. If we can figure out the right machines and visualize how they behave, we can reconstruct all of statistics by putting together these simple machines into arrangements appropriate to the situation. We need only two machines of fit, the spring for fitting continuous Normal responses and the pressure cylinder for fitting categorical responses.

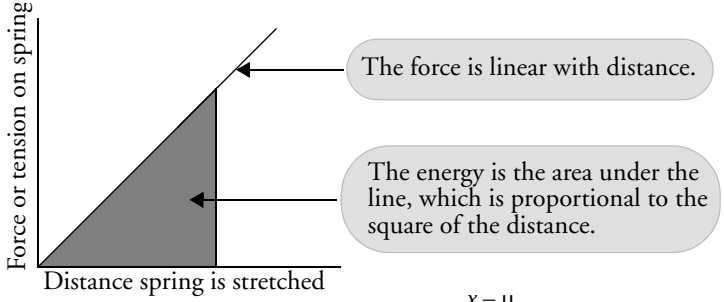
Readers interested in this approach should consult Farebrother (2002), who covers physical models of statistical concepts extensively.

Springs for Continuous Responses

How does a spring behave? As you stretch the spring, the tension increases linearly with distance. The energy that you need to pull a spring a given distance is the integral of the force over the distance, which is proportional to the square of the distance.

Take $1/\sigma^2$ as the measure of how stiff a spring is. Then the graph and equations for the spring are as shown in **Figure 21.1**.

Figure 21.1 Behavior of Springs



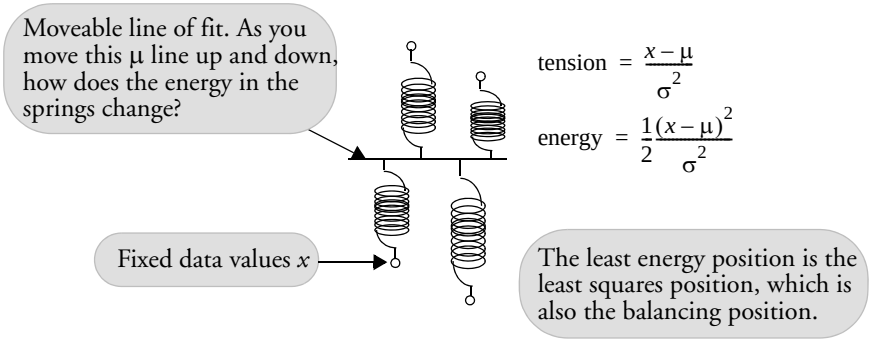
In this way, springs will help us visualize least squares fits. They also help us do maximum likelihood fits when the response has a Normal distribution.

The formula for the log of the density of a Normal distribution is identical to the formula for energy of a spring centered at the mean, with a spring constant equal to the reciprocal of the variance. A spring stores and yields energy in exactly the way that Normal deviations get and give log-likelihood. So, maximum likelihood is equivalent to least squares, which is equivalent to minimizing energy in springs.

Fitting a Mean

How do you fit a mean by least squares? Imagine stretching springs between the data points and the line of fit (see **Figure 21.2**). Then you move the line of fit around until the forces acting on it from the springs balance. That will be the point of minimum energy in the springs. For every minimization problem, there is an equivalent balancing (or orthogonality) problem, in which the forces (tensions, relative distances, residuals) add up to zero.

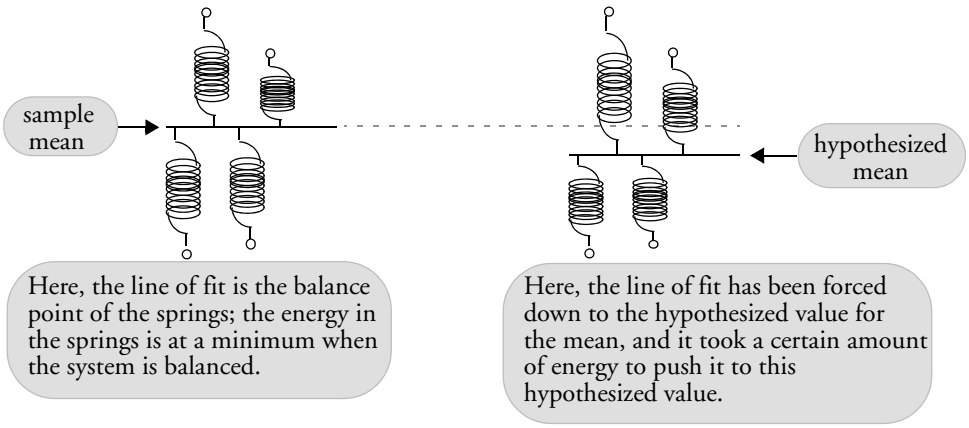
Figure 21.2 Fitting a Mean by Springs



Testing a Hypothesis

If you want to test a hypothesis that the mean is some value, you force the line of fit to be that value and measure how much more energy you had to add to the springs (how much more the sum of squared residuals was) to constrain the line of fit. This is the sum of squares that is the main ingredient of the *F*-test. To test that the mean is (not) the same as a given value, find out how hard it is to move it there (see **Figure 21.3**).

Figure 21.3 Compare a Mean to a Given Value

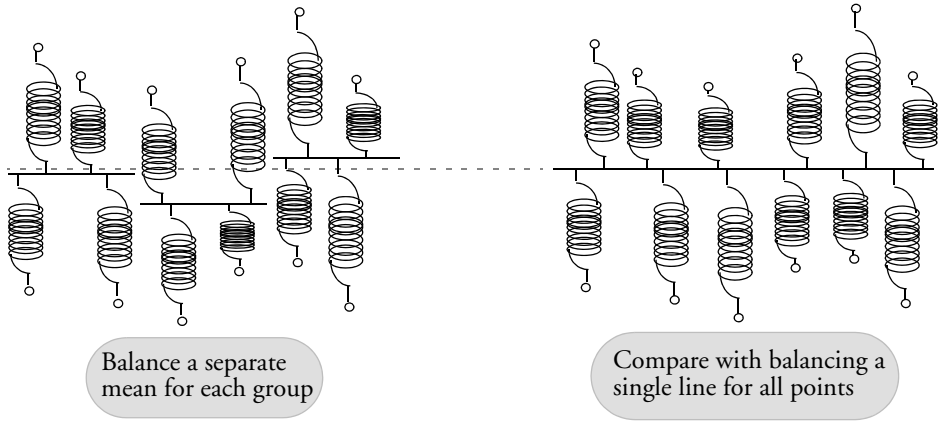


One-Way Layout

If you want to fit several means, you can do so by balancing a line of fit with springs for each group. To test that the means are the same, you force the lines of fit to be the same, so that

they balance as a single line, and measure how much energy you had to add to the springs to do this (how much greater the sum of squared residuals was). See **Figure 21.4.**

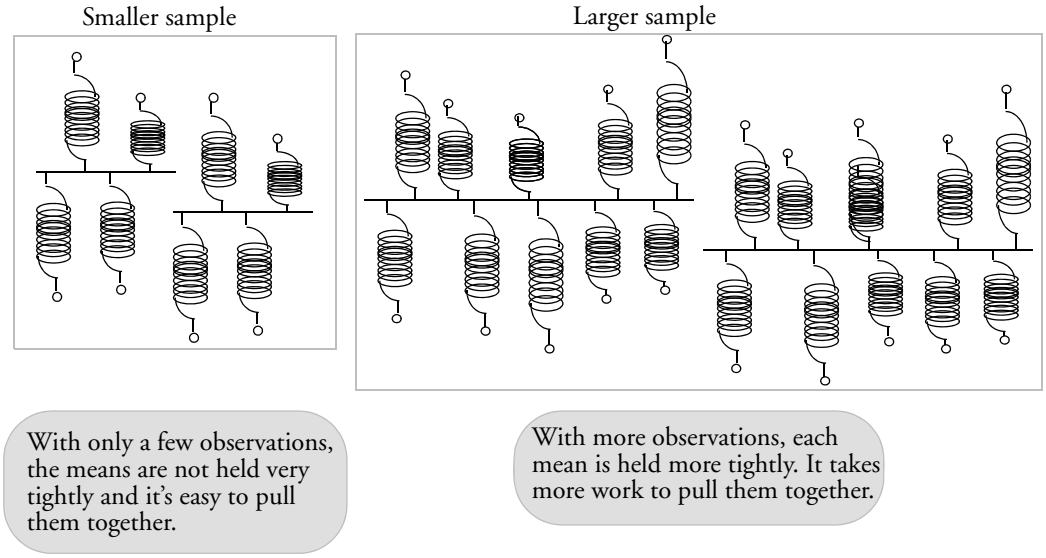
Figure 21.4 Means and the One-Way Analysis of Variance



Effect of Sample Size Significance

When you have a larger sample, there are more springs holding on to each mean estimate, and it is harder to pull them together. Larger samples lead to a greater energy expense (sum of squares) to test that the means are equal. The spring examples in **Figure 21.5** show how sample size affects the sensitivity of the hypothesis test.

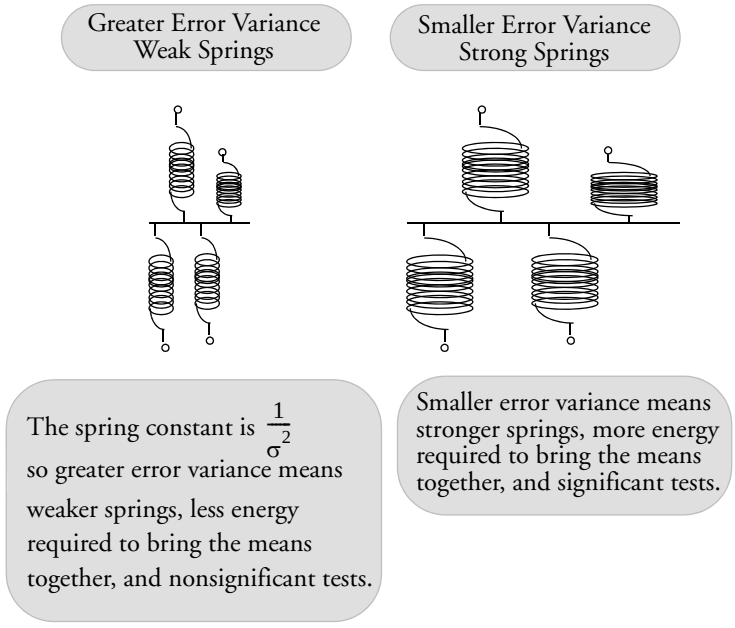
Figure 21.5 A Larger Sample Helps Make Hypothesis Tests More Sensitive



Effect of Error Variance on Significance

The spring constant is the reciprocal of the variance. Thus, if the residual error variance is small, the spring constant is bigger, the springs are stronger, it takes more energy to bring the means together, and the test is therefore more significant. The springs in **Figure 21.6** illustrate the effect of variance size.

Figure 21.6 Reduced Residual Error Variance Makes Hypothesis Tests More Sensitive



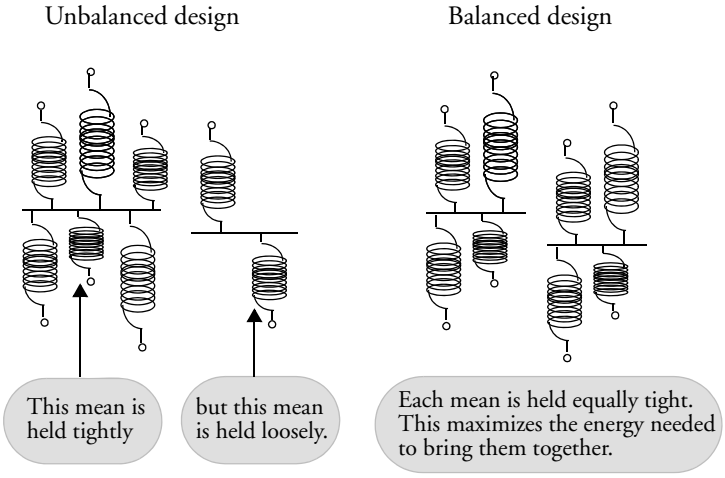
Experimental Design's Effect on Significance

If you have two groups, how do you arrange the points between the two groups to maximize the sensitivity of the test that the means are equal? Suppose that you have two sets of points loading two lines of fit, as in the one-way layout shown previously in **Figure 21.4**. The test that the true means are equal is done by measuring how much energy it takes to force the two lines together.

Suppose that one line of fit is suspended by a lot more points than the other. The line of fit that is suspended by few points will be easily movable and can be stretched to the other mean without much energy expenditure. The lines of fit would be more strongly separated if you had more points on this loosely sprung side, even at the expense of having fewer points on the more tightly sprung side. It turns out that to maximize the sensitivity of the test for a given number of observations, it is best to allocate points in equal numbers between the two groups. In this way both means are equally tight, and the effort to bring the two lines of fit together is maximized.

So the power of the test is maximized in a statistical sense by a balanced design, as illustrated in **Figure 21.7**.

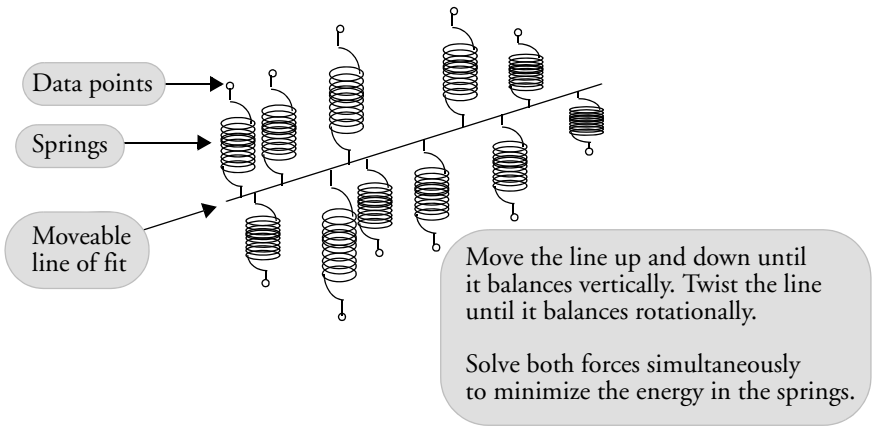
Figure 21.7 Design of Experiments



Simple Regression

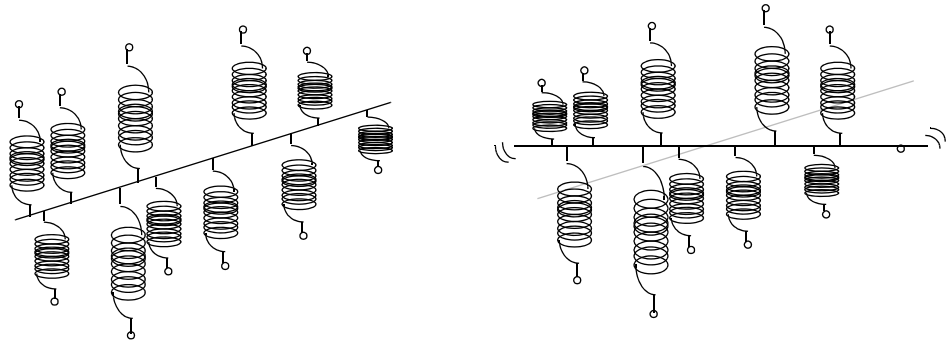
If you want to fit a regression line through a set of points, you fasten springs between the data points and the line of fit, such that the springs stay vertical. Then let the line be free so that the forces of the springs on the line balance, both vertically and rotationally (see **Figure 21.8**). This is the least-squares regression fit.

Figure 21.8 Fitting a Regression Line with Springs



If you want to test that the slope is zero, you force the line to be horizontal so that you're just fitting a mean and measure how much energy it took to constrain the line (the sum of squares due to regression). (See **Figure 21.9**.)

Figure 21.9 Testing the Slope Parameter for the Regression Line



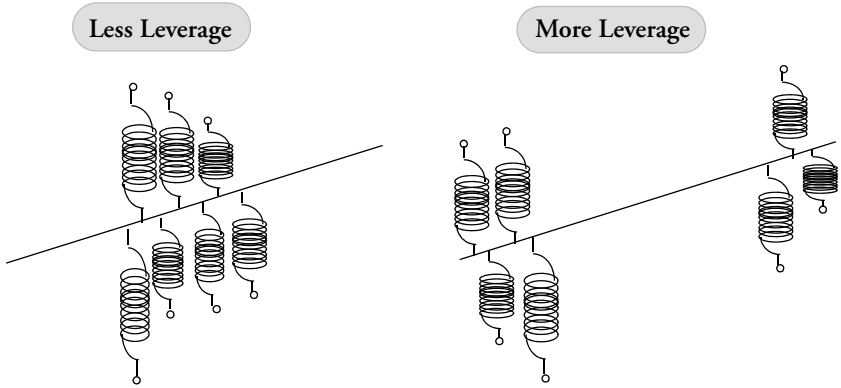
This line is where the forces governing the slope of the line balance. It is the minimum energy solution.

If you force the line to have a slope of zero, how much additional energy do you have to give the springs? How much work is it to move the line to be horizontal?

Leverage

If most of the points that are suspending the line are near the middle, then the line can be rotated without much effort to change the slope within a given energy budget. If most of the points are near the end, the slope of the line of fit is pinned down with greatest resistance to force. That is the idea of leverage in a regression model. Imagine trying to twist the line to have a different slope. Look at **Figure 21.10** and decide which line would be easier to twist.

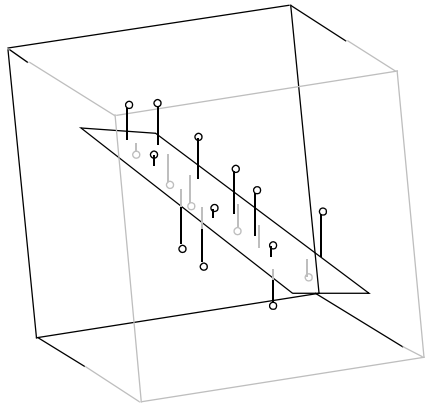
Figure 21.10 Leverage with Springs



Multiple Regression

The same idea works for fitting a response to two regressors; the difference is that the springs are attached to a plane rather than a line. Estimation is done by adjusting the plane so that it balances in each way. Testing is done by constraining the plane.

Figure 21.11 Three-Dimensional Plot of Two Regressors and Fitted Plane



Summary: Significance and Power

To get a stronger (more significant) fit, in which the line of fit is suspended more tightly, you must either have stiffer springs (have smaller variance in error), use more data (have more points to hang springs from), or move your points farther out on both ends of the *x*-axis (more leverage). The power of a test is how likely it is that you will be unable to move the line of fit given a certain energy budget (sum of squares) determined by the significance level.

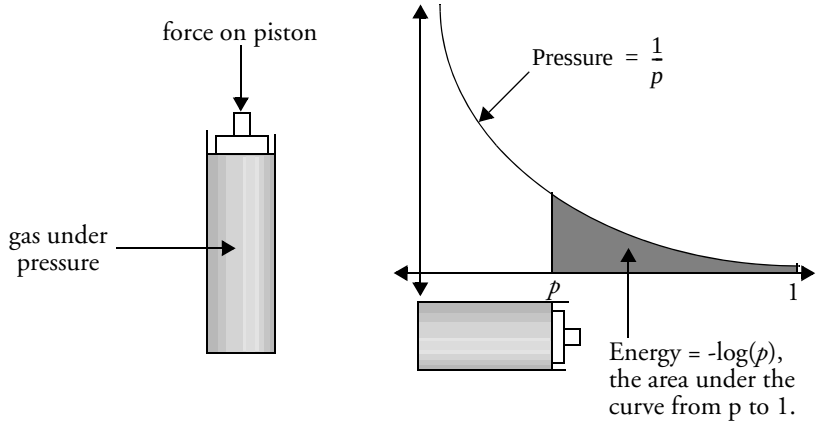
Machine of Fit for Categorical Responses

Just as springs are analogous to least squares fits, gas pressure cylinders are analogous to maximum likelihood fits for categorical responses (see **Figure 21.12**).

How Do Pressure Cylinders Behave?

Using Boyle's law of gases (pressure times volume is constant), the pressure in a gas cylinder is proportional to the reciprocal of the distance from the bottom of the cylinder to the piston. The energy is the force integrated over the distance (starting from a distance, p, of 1), which turns out to be $-\log(p)$.

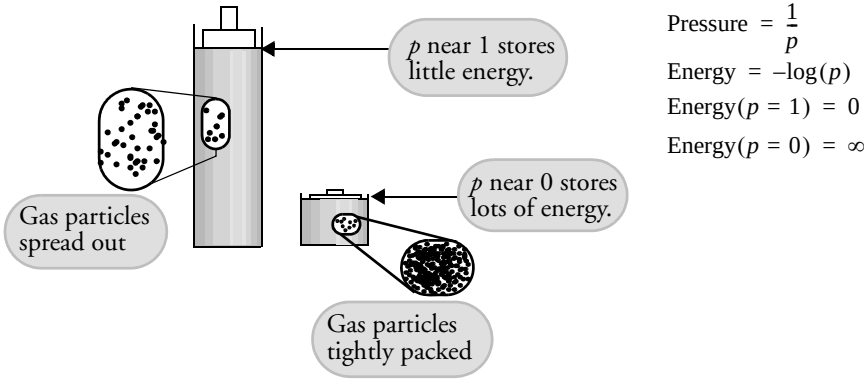
Figure 21.12 Gas Pressure Cylinders Equate –log(probability) to Energy



Now that you know how pressure cylinders work, start thinking of the distance from the bottom of the cylinder to the piston as the probability that some statistical model attributes to some response. The height of 1 will mean no stored energy, no surprise, a probability of 1. The height of zero will mean infinite stored energy, an impossibility, a probability of zero.

When stretching springs, we measured energy by how much work it took to pull a spring, which turned out to be the square of the distance. Now we measure energy by how much work it takes to push a piston from distance 1 to distance p, which turns out to be $-\log(p)$, the logarithm of the probability. We used the logarithm of the probability before in categorical problems when we were doing maximum likelihood. The maximum likelihood method estimates the response probabilities so as to minimize the sum of the negative logarithms of the probability attributed to the responses that actually occurred. This is the same as minimizing the energy in gas pressure cylinders, as illustrated in **Figure 21.13**.

Figure 21.13 Gas Pressure Cylinders Equate –log(probability) to Energy



Estimating Probabilities

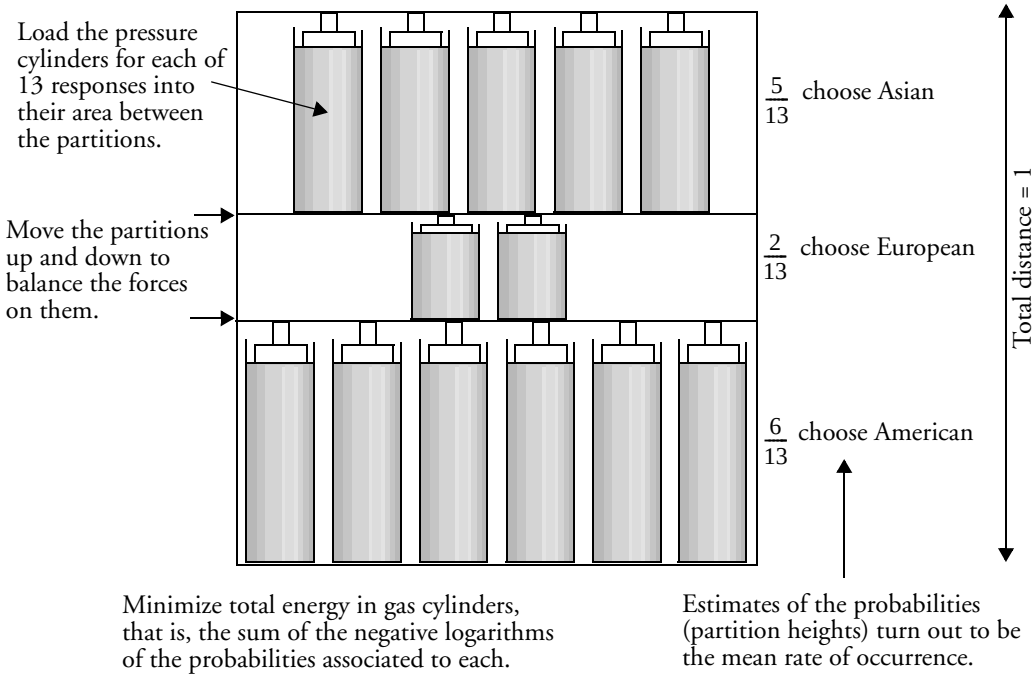
Now we want to estimate the probabilities by minimizing the energy stored in pressure cylinders. First, we need to build a partitioned frame with a compartment for each response category and add the constraint that the sum of the heights of the partitions is 1. We will be moving the partitions around so that the compartments for each response category can get bigger or smaller (see **Figure 21.14**).

For each observation on the categorical response, put a pressure cylinder into the compartment for that response. After you have all the pressure cylinders in the compartments, start moving the partitions around until the forces acting on the partitions balance out. This will be the solution to minimize the energy stored in the cylinders. It turns out that the solution for the minimization is to make the partition sizes proportional to the number of pressure cylinders in each compartment.

For example, suppose you did a survey in which you asked 13 people what brand of car they preferred, and 5 chose Asian, 2 chose European, and 6 chose American brands. Then you would stuff the pressure cylinders into the frame as in **Figure 21.14**, and the partition sizes that would balance the forces would work out to 5/13, 2/13, and 6/13, which sum to 1.

To test that the true probabilities are some specific values, you move the partitions to those values and measure how much energy you had to add to the cylinders.

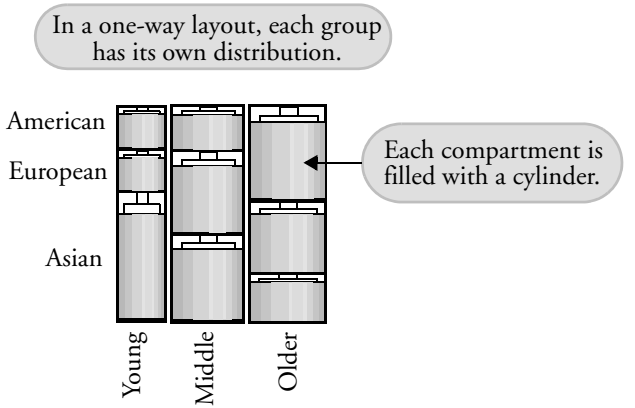
Figure 21.14 Gas Pressure Cylinders Estimate Probabilities for a Categorical Response



One-Way Layout for Categorical Data

If you have different groups, you can fit a different response probability to each group. The forces acting on the partitions balance independently for each group. The plot shown in **Figure 21.15** (which should remind you of a mosaic plot) helps maintain the visualization of pressure compartments. As an alternative to pressure cylinders, you can visualize with free gas in each cell.

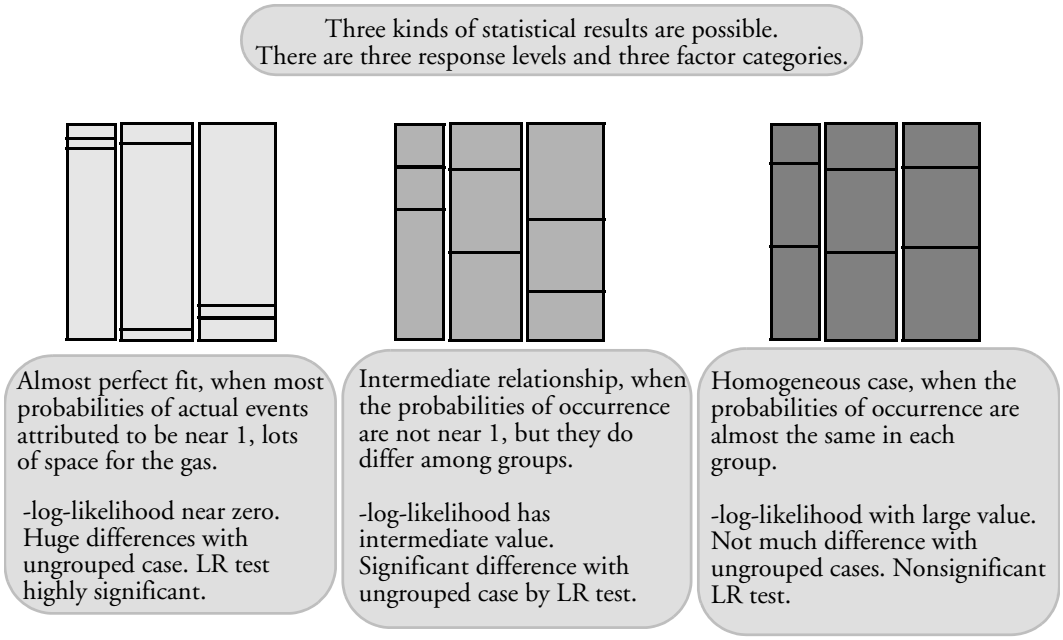
Figure 21.15 Gas Pressure Cylinder Estimate Probabilities for a Categorical Response



How do you test the hypothesis that the true rates are the same in each group and that the observed differences can be attributed to random variation? You just move the partitions so that they are in the position corresponding to the ungrouped population and measure how much more energy you had to add to the gas-pressure system to force the partitions to be in the same positions.

Figure 21.16 shows the three kinds of results you can encounter, corresponding to perfect fit, significant difference, and nonsignificant difference. To the observer, the issue is whether knowing which group you are in will tell you which response level you will have. When the fit is near perfect, you know with near certainty. When the fit is intermediate, you have more information if you know the group you are in. When the fit is inconsequential, knowing which group you are in doesn't matter. To a statistician, though, what is interesting is how firmly the partitions are held by the gases, how much energy it would take to move the partitions, and what consequences would result from removing boundaries between samples and treating it as one big sample.

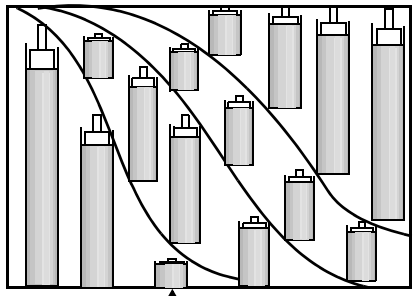
Figure 21.16 Degrees of Fit



Logistic Regression

Logistic regression is the fitting of probabilities over a categorical response to a continuous regressor. Logistic regression can also be visualized with pressure cylinders (see **Figure 21.17**). The difference with contingency tables is that the partitions change the probability as a continuous function of the *x*-axis. The distance between lines is the probability for one of the responses. The distances sum to a probability of 1. **Figure 21.18** shows what weak and strong relationships look like.

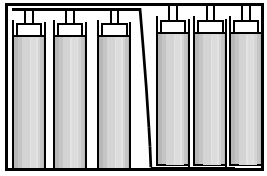
Figure 21.17 Logistic Regression as the Balance of Cylinder Forces



The pressure cylinders push on the logistic curves that partition the responses, moving them so that the total energy in the cylinders is minimized. The more cylinders in an area, the more probability room it pushes for.

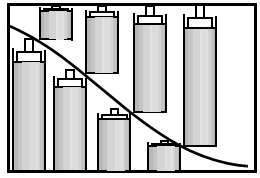
This is an outlier, a response that is holding a lot of energy, having been squeezed down to a small probability.

Figure 21.18 Strengths of Logistic Relationships



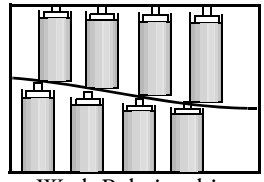
Near Perfect Fit

The probabilities are all near one. No cylinder is hot with energy.



Strong Relationship

Some cylinders must compete with nearby cylinders from a different response. More energy in some cylinders.



Weak Relationship

The probabilities are squeezed to near the horizontal case of homogeneity. All cylinders are warm with energy.

References and Data Sources

Agresti, A. (1984), *Analysis of Ordinal Categorical Data*, New York: John Wiley and Sons, Inc.

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley and Sons, Inc.

Allison, T., and Cicchetti, D. V. (1976), "Sleep in Mammals: Ecological and Constitutional Correlates." Science, November 12, 194, 732-734.

American Society for Testing and Materials. *ASTM Manual on Presentation of Data and Control Chart Analysis*. STP No. 15-D. Philadelphia, 1976.

Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, New York: John Wiley and Sons.

Anscombe, F.J. (1973), *American Statistician*, 27, 17-21.

Aviation Consumer Home Page. U.S. Government, Department of Transportation (http://www.dot.gov/airconsumer/).

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics*, New York: John Wiley and Sons.

Berger, R. L., and Hsu, J. C. (1996), "Bioequivalence Trails, Intersection-Union Tests, and Equivalence Confidence Sets," *Statistical Science*, 11, 283-319.

Box G.E.P., Jenkins G. M. and Reisnel, G.C. (1994), *Time Series Analysis: Forecasting and Control*, *Third Edition*. Prentice Hall: Englewood Cliffs, NJ.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (2005), *Statistics for Experimenters*, New York: John Wiley and Sons, Inc.

Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children," *Research Quarterly for Exercise and Sport*, 63, 418-424.

Cochran, W.G. and Cox, G.M. (1957), *Experimental Designs, 2nd edition*, New York: John Wiley and Sons.

Creighton, W. P. (2000), "Starch content's dependence on several manufacturing factors." Unpublished data.

Cryer, J. and Wittmer, J. (1999) Notes on the Theory of Inference. NCSSM Summer Statistics Institute, available at http://courses.ncssm.edu/math/Stat_Inst/Notes.htm.

Daniel, C. (1959), "Use of Half-Normal Plots in Interpreting Factorial Two-level Experiments," *Technometrics*, 1, 311-314.

Data and Story Library, http://lib.stat.cmu.edu/DASL/.

Ehrstein, James and Croarkin, M. Carroll. *Statistical Reference Datasets*. US Government, National Institute of Standards and Technology (http://www.nist.gov/itl/div898/strd/).

Eppright, E.S., Fox, H.M., Fryer, B.A., Lamkin, G.H., Vivian, V.M., and Fuller, E.S. (1972), "*Nutrition of Infants and Preschool Children in the North Central Region of the United States of America*," World Review of Nutrition and Dietetics, 14.

Eubank, R.L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker, Inc.

Farebrother, R. W. (2002), *Visualizing Statistical Models and Concepts*, New York: Marcel Dekker, Inc.

Fortune Magazine (1990), *The Fortune 500 List*, April 23, 1990.

Gabriel, K.R. (1982), "Biplot," *Encyclopedia of Statistical Sciences, Volume 1*, eds. N.L.Johnson and S. Kotz, New York: John Wiley and Sons, Inc., 263-271.

Gosset, W.S. (1908),"The Probable Error of a Mean," *Biometrika*, 6, pp 1-25.

Hajek, J. (1969), *A Course in Nonparametric Statistics*, San Francisco: Holden-Day.

Henderson, H. V. and Velleman, P. F. (1981), "Building Regression Models Interactively." *Biometrics*, 37, 391-411. Data originally collected from Consumer Reports.

Hosmer, D.W. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley and Sons.

Iman, R.L. (1995), *A Data-Based Approach to Statistics*, Belmont, CA: Duxbury Press.

Iman, R.L. and Conover, W.J. (1979), "The Use of Rank Transform in Regression," *Technometrics*, 21, 499-509.

Isaac, R. (1995) The Pleasures of Probability. New York: Springer-Verlag.

John, P.M. (1971), *Statistical Design and Analysis of Experiments*, New York: Macmillan.

Kaiser, H.F. (1958), "The varimax criterion for analytic rotation in factor analysis" *Psychometrika*, 23, 187-200.

Kemp, A.W. and Kemp, C.D. (1991),"Weldon's dice data revisited," *The American Statistician*, 45 216-222.

Klawiter, B. (2000), *An Investigation into the Potential for Geochemical / Geoarchaeological Provenance of Prairie du Chien Cherts*. Master's Thesis: University of Minnesota.

Koehler, G. and Dunn, J.D. (1988), "The Relationship Between Chemical Structure and the Logarithm of the Partition," *Quantitative Structure Activity Relationships*, 7.

Koopmans, L. (1987), *Introduction to Contemporary Statistical Methods*, Belmont, CA: Duxbury Press, p 86.

Ladd, T. E.(1980 and 1984) and Carle, R. H. (1996), Clerks of the House of Representatives. *Statistics of the Presidential and Congressional Elections*. US Government. Available at http://clerkweb.house.gov/histrecs/history/elections/elections.htm.

Larner, M. (1996), Mass and its Relationship to Physical Measurements. MS305 Data Project, Department of Mathematics, University of Queensland.

Lenth, R.V. (1989), "Quick and Easy Analysis of Unreplicated Fractional Factorials," *Technometrics*, 31, 469-473.

Linnerud (see Rawlings (1988)).

McCullagh, P. and Nelder, J. A. (1983), *Generalized Linear Models*. From *Monographs on Statistics and Applied Probability*, Cox, D. R. and Hinkley, D. V., eds. London: Chapman and Hall, Ltd.

Miller, A.J. (1990), *Subset Selection in Regression*, New York: Chapman and Hall.

Moore, D.S. and McCabe, G. P. (1989), *Introduction to the Practice of Statistics*, New York and London: W. H. Freeman and Company.

Myers and McCaulley pp 46-48 *Myers-Briggs test reference.*

Nelson, L. (1984), "The Shewhart Control Chart - Tests for Special Causes," *Journal of Quality Technology*, 15, 237-239.

Nelson, L. (1985), "Interpreting Shewhart X Control Charts," *Journal of Quality Technology*, 17, 114-116.

Perkiömäki, M. (1995) *Track and Field Statistics* (http://mikap.iki.fi/sport/index.html).

Smyth, G. (2000), "Selling Price of Antique Grandfather Clocks," *OzDASL* web site (http://www.maths.uq.edu.au/~gks/data/general/antiques.html).

Rasmussen, M. (1998), Observations on Behavior of Dolphins. University of Denmark, Odense, via OzDASL (http://www.maths.uq.edu.au/~gks/data/index.html).

Rawlings, J.O. (1988), *Applied Regression Analysis: A Research Tool*, Pacific Grove, CA: Wadsworth and Books/Cole.

Sall, J.P. (1990), "Leverage Plots for General Linear Hypotheses," *American Statistician*, 44, (4), 303-315.

SAS Institute (1986), *SAS/QC User's Guide, Version 5 Edition*, Cary, NC: SAS Institute Inc.

SAS Institute (1987), *SAS/STAT Guide for Personal Computers, Version 6 Edition*, Cary, NC: SAS Institute Inc.

SAS Institute (1988), *SAS/ETS User's Guide, Version 6 Edition*, Cary, NC: SAS Institute Inc.

SAS Institute (1989), *SAS/Technical Report P-188: SAS/QC Software Examples, Version 6 Edition*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1989), *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute Inc., 1165-1168.

Schiffman, A. (1982), *Journal of Counseling and Clinical Psychology*.

Schuirmann, D.L. (1981), "On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval," *Biometrics* 37, 617.

Snedecor, G.W. and Cochran, W.G. (1967), *Statistical Methods*, Ames, Iowa: Iowa State University Press.

Simpson, E.H. (1951), The interpretation of interaction in contingency tables, *JRSS* B13: 238-241.

Stichler, R.D., Richey, G.G. and Mandel, J.(1953), "Measurement of Treadwear of Commercial Tires," *Rubber* Age, 73:2.

Stigler, S.M. (1986), *The History of Statistics*, Cambridge: Belknap Press of Harvard Press.

Stigler, S. M. (1977), Do Robust Estimators Work with Real Data? *The Annals of Statistics* 5:4, 1075.

Swift, Jonathan (1735), *Gulliver's Travels*. Quote is from p. 44 of the *Norton Critical Edition*, (1961) Robert A. Greenwood, ed. New York: W.W. Norton & Co.

Negiz, A. (1994) "Statistical Monitoring and Control of Multivariate Continuous Responses". NIST/SEMATECH e-Handbook of Statistical Methods (http://www.itl.nist.gov/div898/handbook/pmc/section6/pmc621.htm).

Neter, J. and Wasserman, W. (1974), *Applied Linear Statistical Models*, Homewood, IL: Richard D Irwin, Inc.

Theil and Fiebig, (1984), *Exploiting Continuity*, Cambridge, Mass: Ballinger Publishing Co.

Third International Mathematics and Science Study, (1995). US Government: National Center for Educational Statistics and International Education Association.

Tukey, J. (1953), "A problem of multiple comparisons," Dittoed manuscript of 396 pages, Princeton University.

Tversky and Gilovich (1989), "The Cold Facts About the Hot Hand in Basketball," *CHANCE*, 2, 16-21.

Wardrop, Robert (1995), "Simpson's Paradox and the Hot Hand in Basketball", *American Statistician*, Feb 49:1, 24-28.

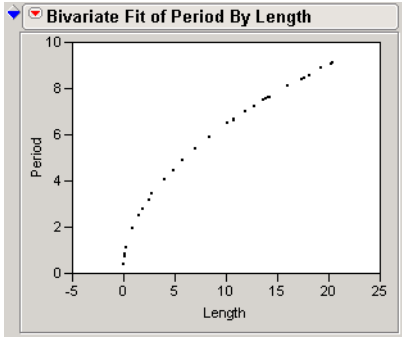
Westlake, W.J. (1981), "Response to R.B.L. Kirdwood: bioequivalence testing--a need to rethink", *Biometrics* 37, 589-594.

Yule, G.U. (1903), "Notes on the theory of association of attributes in statistics," *Biometrika* 2: 121-134.

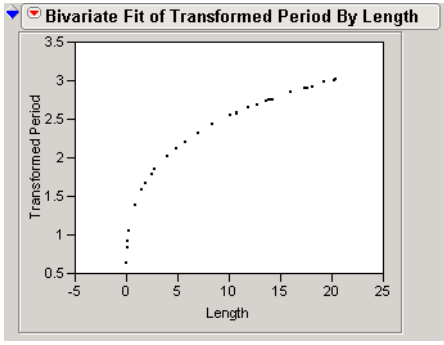
Answers to Selected Exercises

Chapter 4, "Formula Editor Adventures"

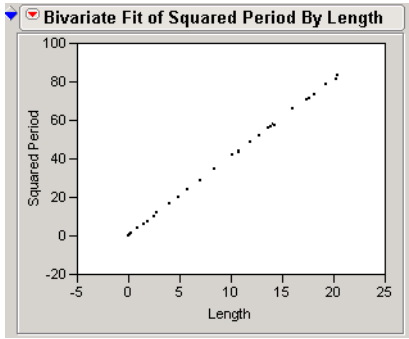
1. a.



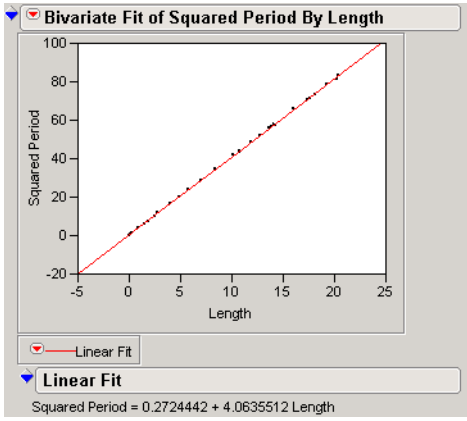
b.



c. The square of Period linearized the data.



d.



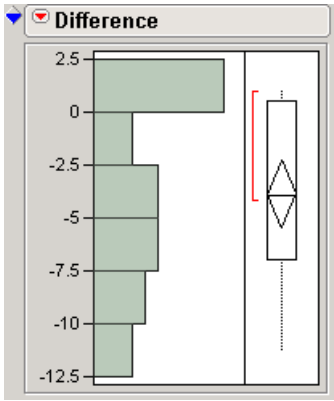
e. $\text{Period}^2 = 0.272 + 4.06 \times \text{Length}$, so $\text{Period} = \sqrt{0.272 + 4.06 \times \text{Length}}$

f. Enter the formula

$$\frac{2*Pi}{\sqrt{9.8}} * \sqrt{Length}$$

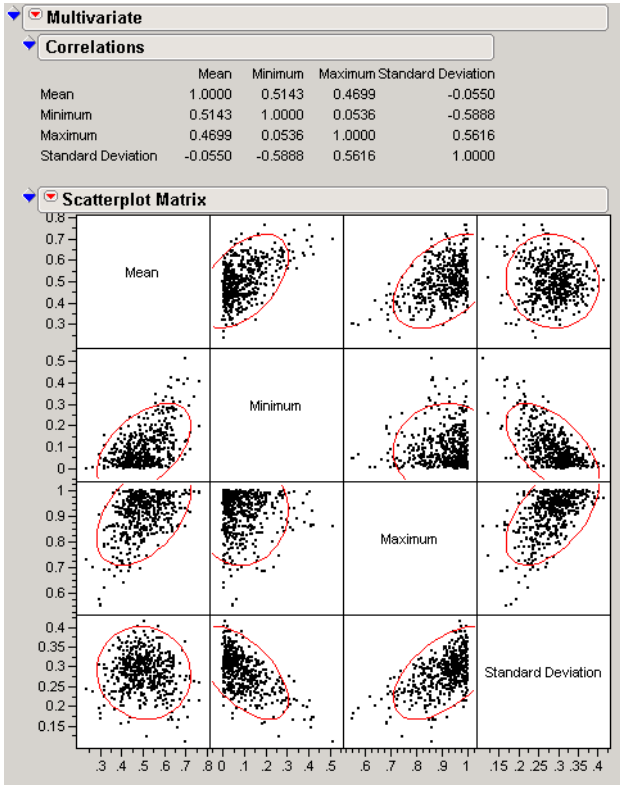
into a new column, then make yet another new column to find the difference between observed and theoretical values.

Use the Distribution platform to get the following histogram.

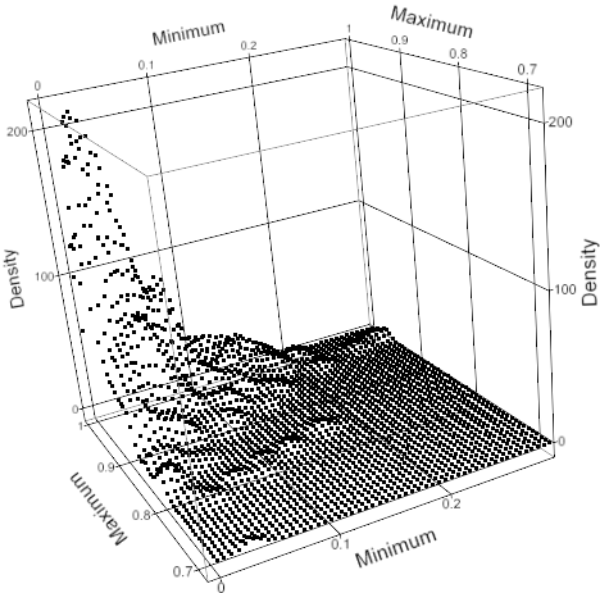


g. This histogram reveals that the students data is generally smaller than the theoretically predicted values.

2. c. The multivariate plot shows some correlation among the mean, minimum, and maximum, and among the standard deviation, minimum, and maximum.



e. Minimum and Maximum yield the following using **Scatterplot 3D**.



3. a. The value converges to $\frac{\text{Fib}}{\text{Fib}_{\text{Row()-1}}} \approx 1.618 \approx \frac{1 \pm \sqrt{5}}{2} = \phi$, the golden ratio.

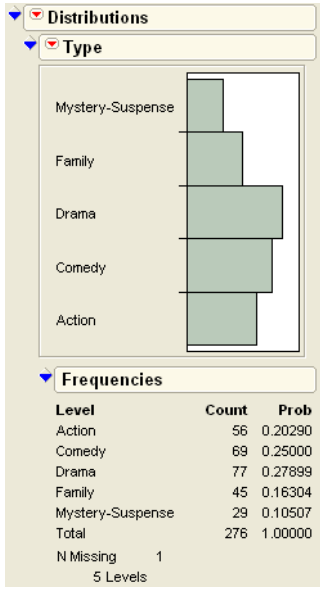
c. It converges to the same number.

d. Again, it converges to the same number.

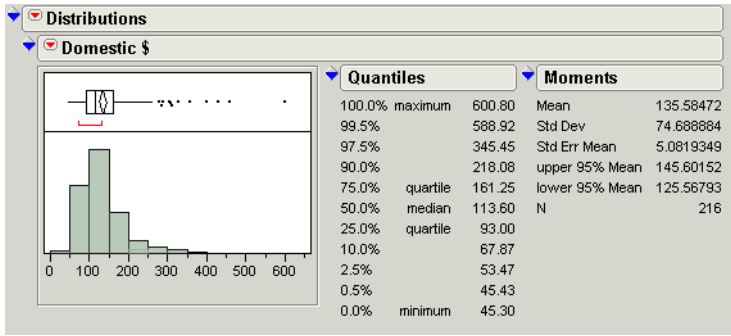
e. This time, the numbers converge to ¼.

Chapter 7, "Univariate Distributions: One Variable, One Sample"

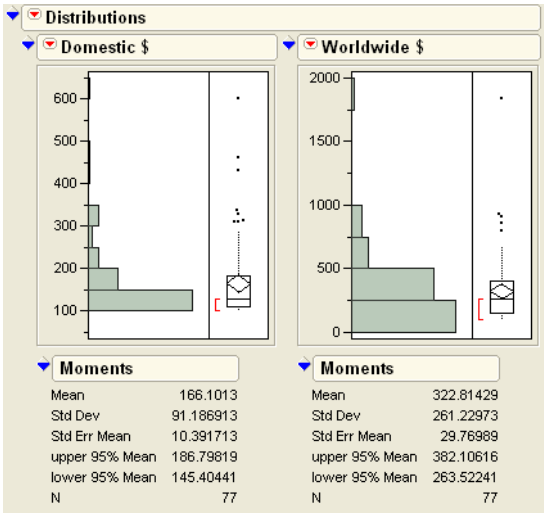
1. a. Levels and counts are shown in the **Frequencies** section of the report.



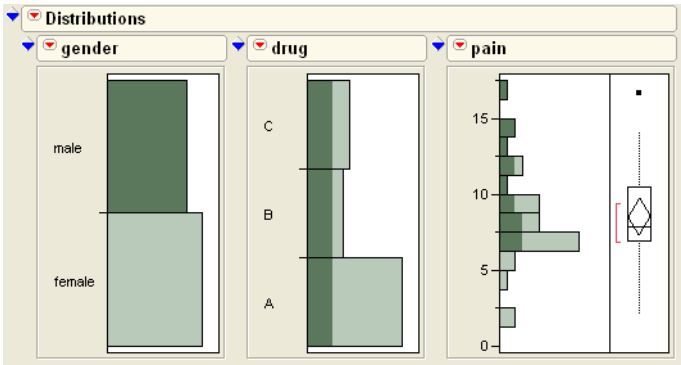
b. The grosses range from $45.4 million to $600.8 million with an average gross of $135.5 million.



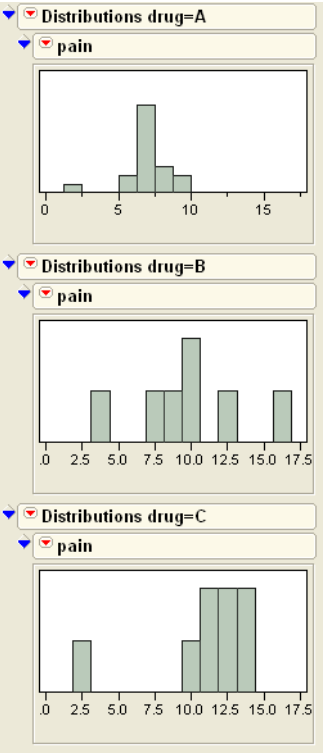
d. To create the subset, use **Rows > Row Selection > Select Where** and complete the dialog to select where Type equals Drama. Then, use **Tables > Subset** to create the data table.



2. a.The following picture has the males highlighted. There are far more females for drug A than males.

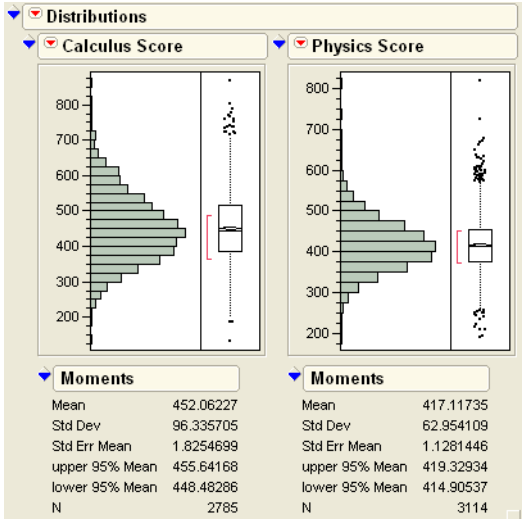


b. To produce this report, select **Analyze > Distribution**, assign pain to **Y, Columns** and drug to **By**. The means do not appear to be the same.



3.

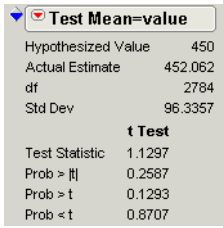
a.



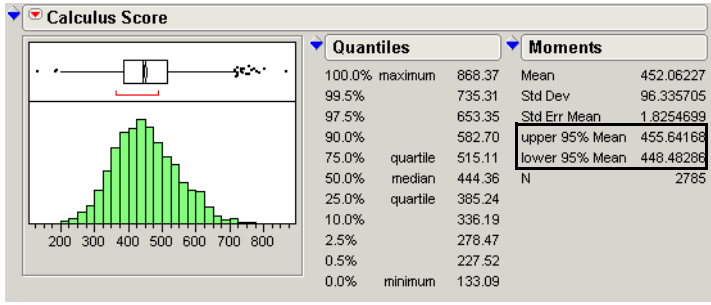
b. To produce the relevant report, select Calculus Score as **Y, Columns** and Region as **By**. The means for the four regions are 467.54, 445.1, 464.9, and 441.27 respectively.

c. The mean Physics scores for each of the four regions are 424.1, 404.8, 427.9, and 417.4 respectively.

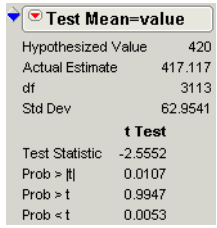
d. After requesting a distribution of the scores, use the **Test Mean** command from the platform menu to test that the mean is not 450. The following report appears, showing that there is not evidence that the mean is different from 450.



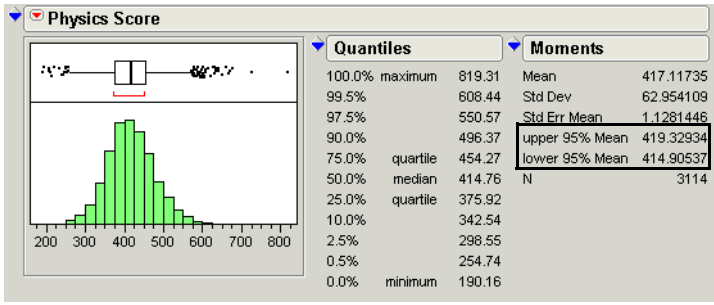
e. The confidence interval is shown in the **Moments** section of the report.



f. After requesting a distribution of the scores, use the **Test Mean** command from the platform menu. The resulting report shows that the mean appears to be less than 420.



g. The confidence interval is shown in the **Moments** section of the report.



4.

a. 1.44 g. "100% Natural Bran Oats & Honey", "Banana Nut Crunch", and "Cracklin' Oat Bran" appear to have unusually high amounts of fat.

b. "All Bran with Extra Fiber" and "Fiber One"

c. Cold Cereals: (8.15g, 10.78g); Hot Cereals (-2.46g, 5.13g)

5.

a. Auto and Robbery seem to be skewed, so they don't appear normal. The others have a bell-shaped appearance.

b. Nevada and New York

6.

a. Average height is 73.4 inches; average weight is 215.7 pounds.

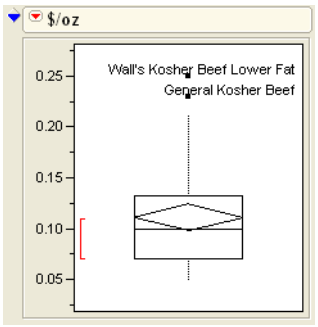
b. Smallest average weight is for wr (wide receivers); largest average weight is for dl (defensive linemen).

c. dl (defensive linemen) have the largest average neck measurements. lb (linebackers) can bench press the most weight.

7.

a. No, there are more beef hot dogs considered.

b.



c. Beef: (146.2, 167.4) Meat:(145.7, 171.7) Poultry: (107.1, 130.4). Poultry has the lowest average, at 118.8.

8.

a. The data appear normally distributed. The Normal Quantile plot shows no reason to think the data is not normal.

b. 72.5 words per minute

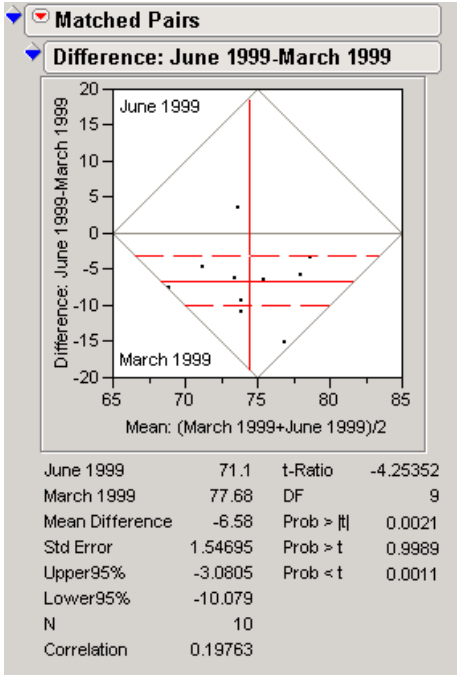
c. Regal: (76.1, 85.4) Speedytype: (76.1, 85.5) Word-o-matic: (54.8, 78.2)

Chapter 8, "The Difference between Two Means"

1.

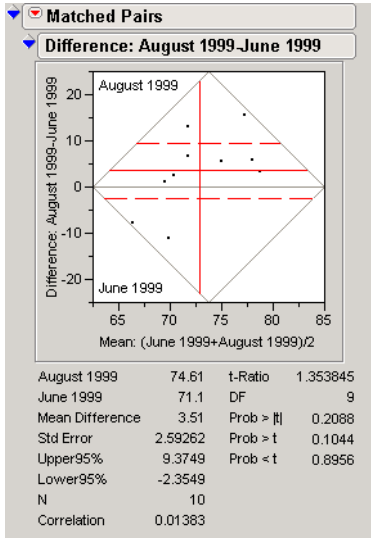
a. A matched pairs approach is more appropriate, since these are repeated measures over time.

b. The Matched Pairs platform yields the following report, showing a significant difference between the two months.



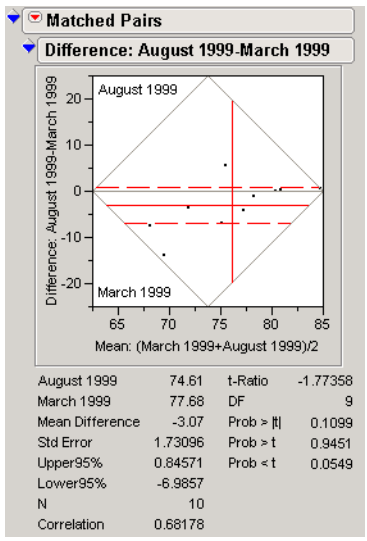
June 1999	71.1	t-Ratio	-4.25352
March 1999	77.68	DF	9
Mean Difference	-6.58	Prob > |t|	0.0021
Std Error	1.54695	Prob > t	0.9989
Upper95%	-3.0805	Prob < t	0.0011
Lower95%	-10.079		
N	10		
Correlation	0.19763		

c. There is no evidence for a significant difference between August and June.



August 1999	74.61	t-Ratio	1.353845
June 1999	71.1	DF	9
Mean Difference	3.51	Prob > |t|	0.2088
Std Error	2.59262	Prob > t	0.1044
Upper95%	9.3749	Prob < t	0.8956
Lower95%	-2.3549		
N	10		
Correlation	0.01383		

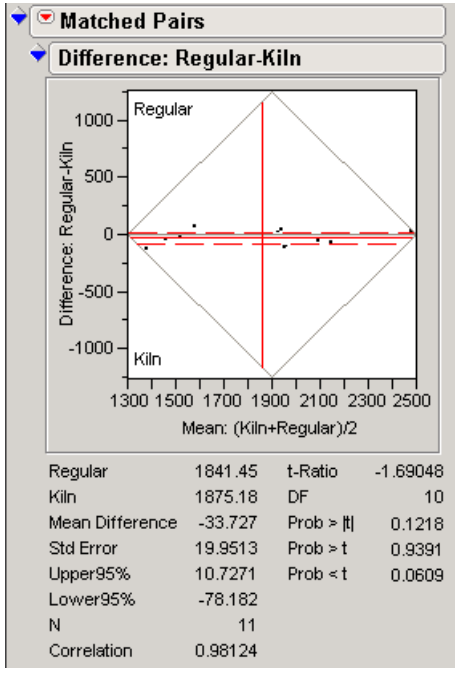
Similarly, there is no evidence for a difference between August and March.



August 1999	74.61	t-Ratio	-1.77358
March 1999	77.68	DF	9
Mean Difference	-3.07	Prob > |t|	0.1099
Std Error	1.73096	Prob > t	0.9451
Upper95%	0.84571	Prob < t	0.0549
Lower95%	-6.9857		
N	10		
Correlation	0.68178		

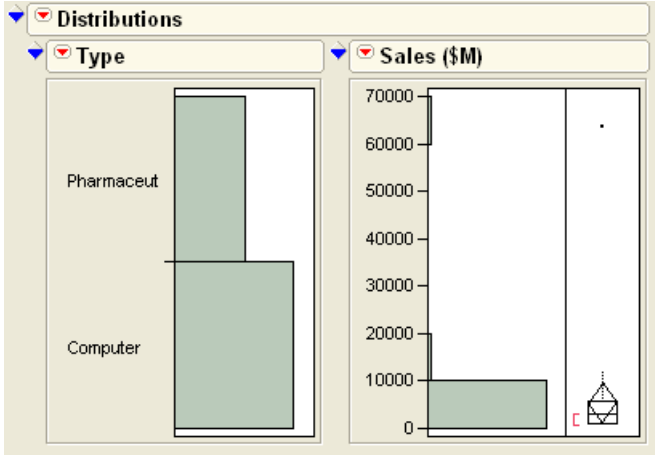
2.

b. There does not appear to be strong evidence between the two. However, the result is marginal and deserves further investigation.



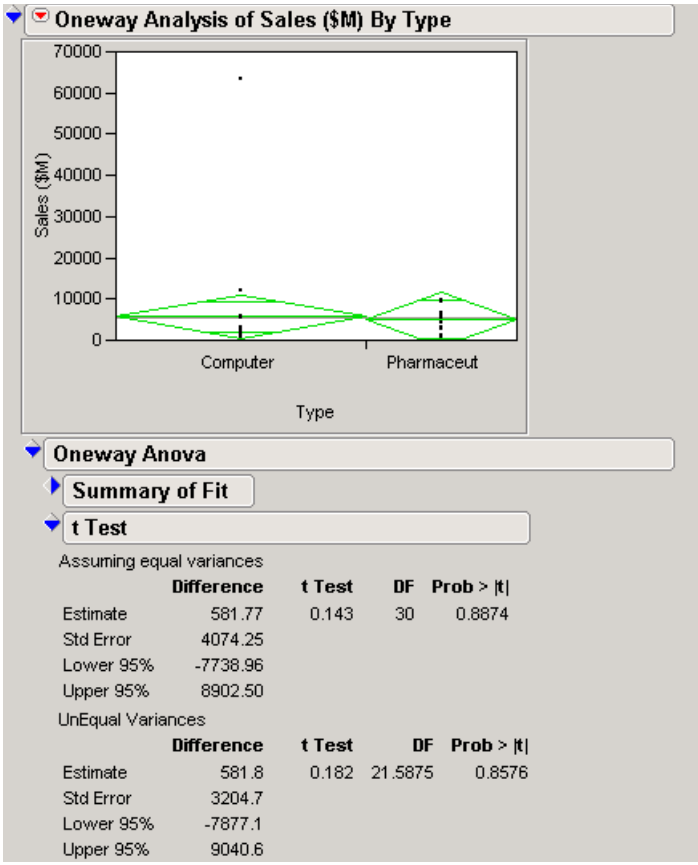
3.

a. The histograms are shown here. Note the outlier in the sales column and the skewed nature of the distribution.



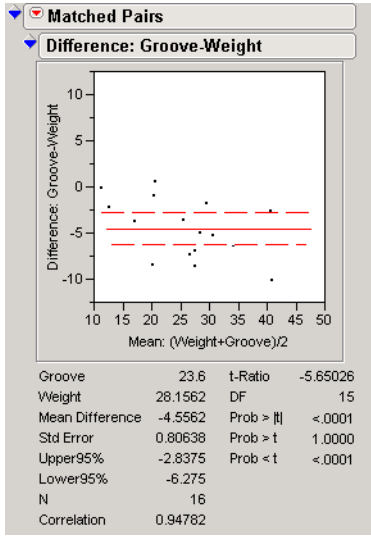
b. Grouped means are appropriate in this situation.

c. Using Fit Y by X with Sales as Y and Type as X allows for the **Means/Anova/t test** command to be used. It produces the report shown here, which does not show evidence for a difference.

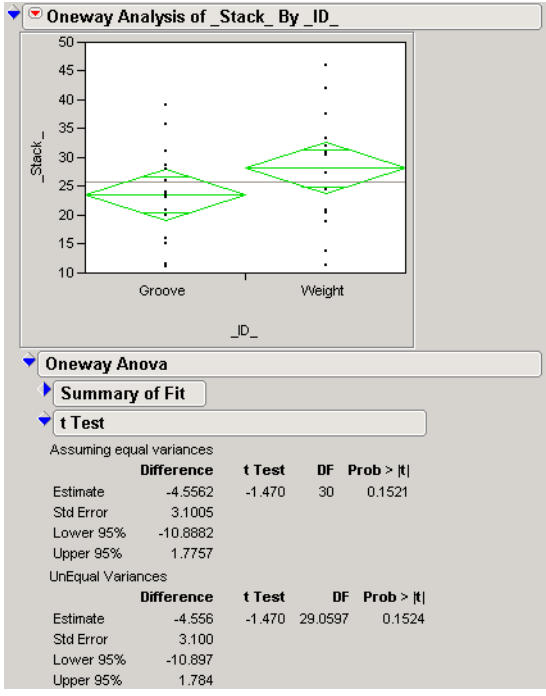


4.

a. The following report shows that there is a significant difference between the two measurements.



b. After stacking, the Fit Y By X platform can be used to reveal the following report. This test does not detect a difference.

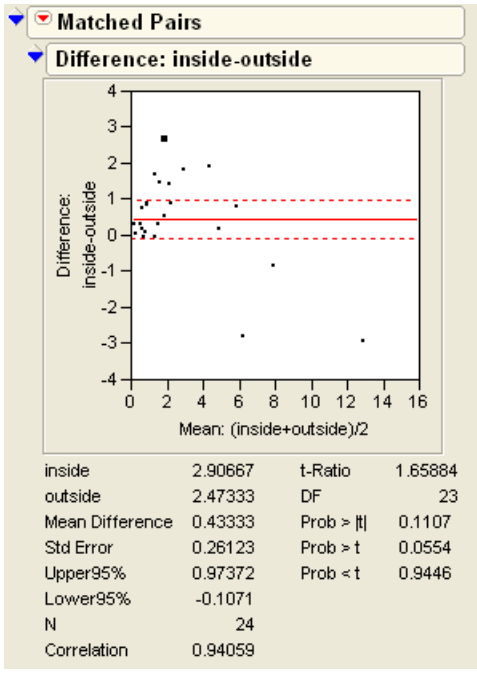


c. The matched pairs approach is appropriate in this case.

d. The scientist would not have detected a difference (in this case, a strong difference) with the wrong analysis.

5.

b. The following report comes from the Matched Pairs platform.



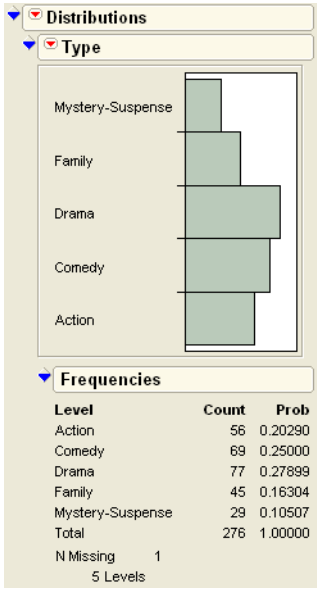
inside	2.90667	t-Ratio	1.65884
outside	2.47333	DF	23
Mean Difference	0.43333	Prob > |t|	0.1107
Std Error	0.26123	Prob > t	0.0554
Upper95%	0.97372	Prob < t	0.9446
Lower95%	-0.1071		
N	24		
Correlation	0.94059		

It has a p-value of 0.11, a nonsignificant (but barely so) value. More investigation is a good idea.

Chapter 9, "Comparing Many Means: One-Way Analysis of Variance"

1.

a. From the distribution platform:

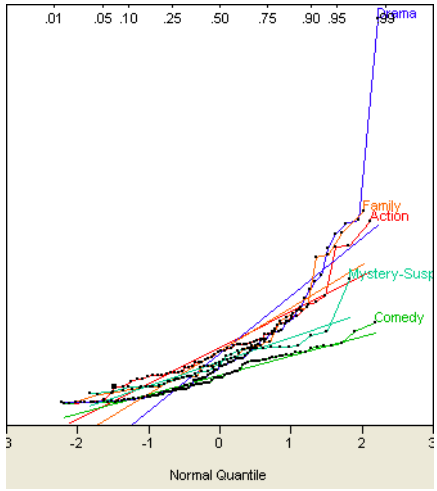


There are not an equal number of movies of each type.

b. With $p = 0.0020$. there is evidence for a difference between at least two movie types.

c. Action and Drama are not different from all other movie types.

d. The normal quantile plot looks like the following.



Since the lines appear to have very different slopes, a Welch ANOVA is not a bad idea.

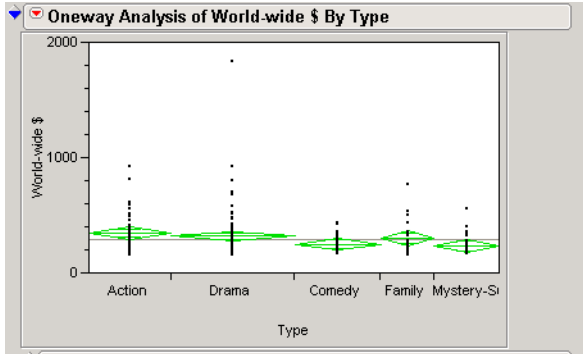
Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	1.7421	4	271	0.1410
Brown-Forsythe	3.8784	4	271	0.0044*
Levene	5.6545	4	271	0.0002*
Bartlett	21.1239	4	.	<.0001*

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
9.2975	4	109.85	<.0001*

The output form the UnEqual Variances command shows four tests that the variances are unequal. Three support the conclusion that they are not equal. The Welch ANOVA shows a similar conclusion to the parametric ANOVA: there is a difference among the movie types.

b. The analysis of variance does show differences among the types.



Oneway Anova

Summary of Fit

Rsquare	0.056497
Adj Rsquare	0.038525
Root Mean Square Error	170.6263
Mean of Response	293.1665
Observations (or Sum Wgts)	215

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Type	4	366093.6	91523.4	3.1437	0.0154
Error	210	6113799.8	29113.3		
C. Total	214	6479893.4			

Means for Oneway Anova

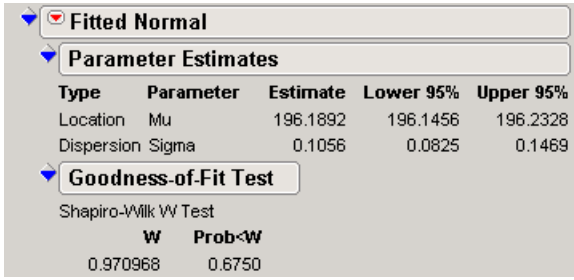
Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Action	43	344.509	26.020	293.21	395.80
Drama	65	318.748	21.164	277.03	360.47
Comedy	45	248.747	25.435	198.61	298.89
Family	28	298.636	32.245	235.07	362.20
Mystery-Suspense	34	233.615	29.262	175.93	291.30

Std Error uses a pooled estimate of error variance

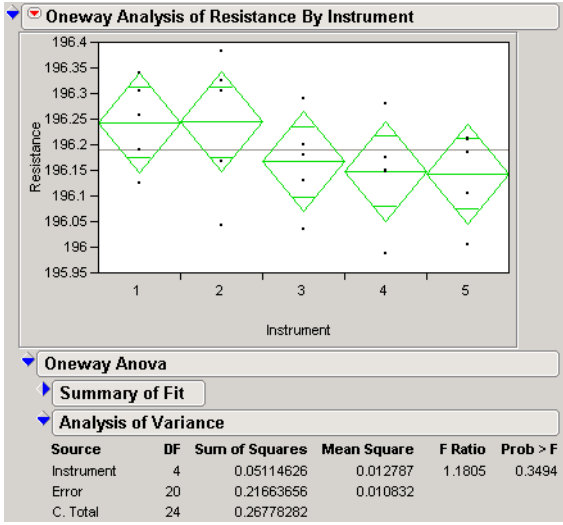
c. They do not appear to be different from the other groups.

2.

a. After examining the histograms, **Fit Distribution > Normal** overlays a normal curve. From the Fitted Normal outline bar, Goodness of Fit displays the report shown here. The data appear to be normal.



b. Fit Y by X is used to generate the ANOVA shown here. There is no evidence that the instruments differ.



3.

a. 299,852.4 km/sec

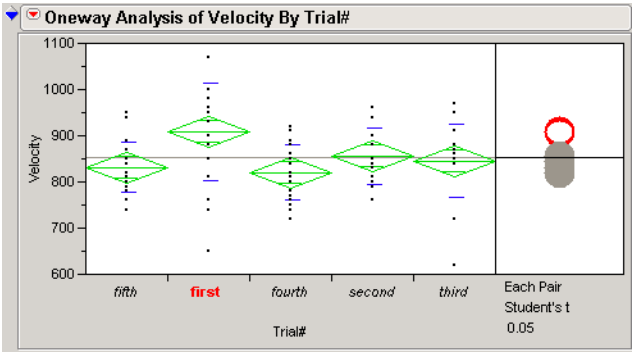
b.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Trial#	4	94514.00	23628.5	4.2878	0.0031
Error	95	523510.00	5510.6		
C. Total	99	618024.00			

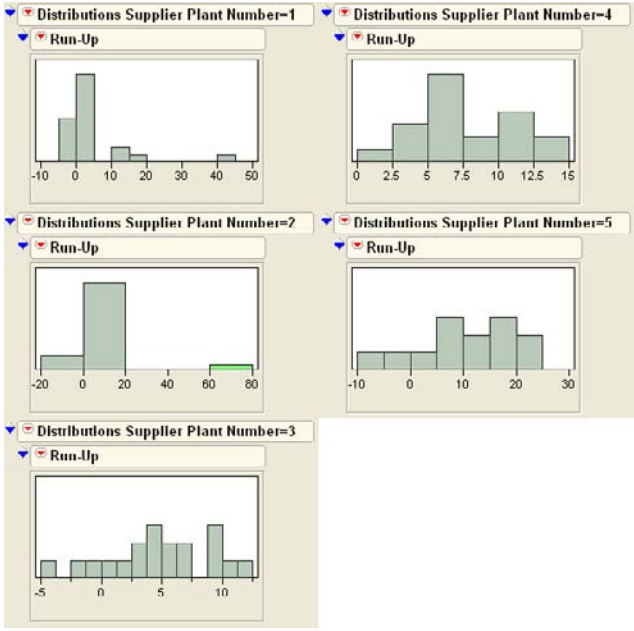
There is evidence that the trials differ.

c.



d. Excluding the first group and re-computing the mean results in a mean of 299838.25.

4. a.



b.

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
1	22	687.000	31.2273	-3.251
2	22	1120.50	51.2955	0.635
3	19	830.500	43.7105	0.754
4	19	1081.00	56.8947	1.568
5	13	833.000	64.0769	2.250

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
15.3185	4	0.0041*

Median Test (Number of Points Above Median)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
1	22	4.000	0.181818	-3.331
2	22	12.000	0.545455	0.540
3	19	8.000	0.421053	0.714
4	19	13.000	0.684211	1.837
5	13	10.000	0.769231	2.119

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
15.7366	4	0.0034*

Van der Waerden Test (Normal Quantiles)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
1	22	-10.591	-0.48139	-2.672
2	22	2.206	0.10029	0.557
3	19	3.154	0.16599	0.839
4	19	4.985	0.26292	1.329
5	13	6.540	0.50327	2.026

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
11.2432	4	0.0240*

All three give the same (significant) result.

c. An ANOVA shows a non-significant result.

5.

a. There is a difference among the regions for Calculus scores.

b. There is a difference among the regions for Physics scores.

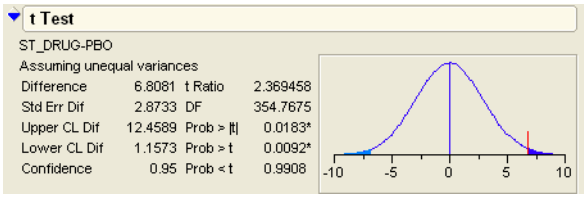
c. Groups 1 and 3 appear to be quite different from groups 2 and 4.

6. a. The ANOVA shows that there is a difference among groups, p=0.0004.

b. Speedytype is clearly superior to the other two brands.

7. There is no statistical evidence (but a complete answer would include a p-value!)

8. a.



We use the unequal variance **t test** command to see that there is a significant difference between the two groups.

b. There is no evidence to support a difference between genders.

c. There is weak support that there is a difference among races. The unequal variances suggest a Welch ANOVA, which shows a difference. A cautious investigator would investigate further.

e. The weights are significantly different between the two groups. However, we can account for that in the analysis, as you will see in a future chapter.

9. a. Yes.

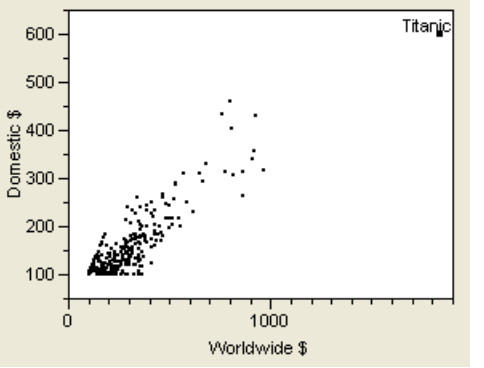
b. Yes.

c. mean(A)=6.98; mean(B)=9.83; mean(C)=10.89

d. There is a noticeable difference between the males and females. For the females, there was no difference in the three drugs. For the males, there was a strong difference.

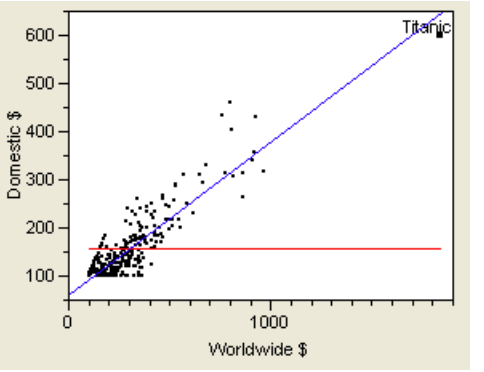
Chapter 10, "Fitting Curves through Points: Regression"

1. a.



One movie is definitely separated from the rest of the movies.

c.



The linear model does explain the data better than the simple mean model. (Why?)

d. The model with excluded outliers is probably a better summary of a typical movie's performance.

f. For comedy, the equation of the regression line is

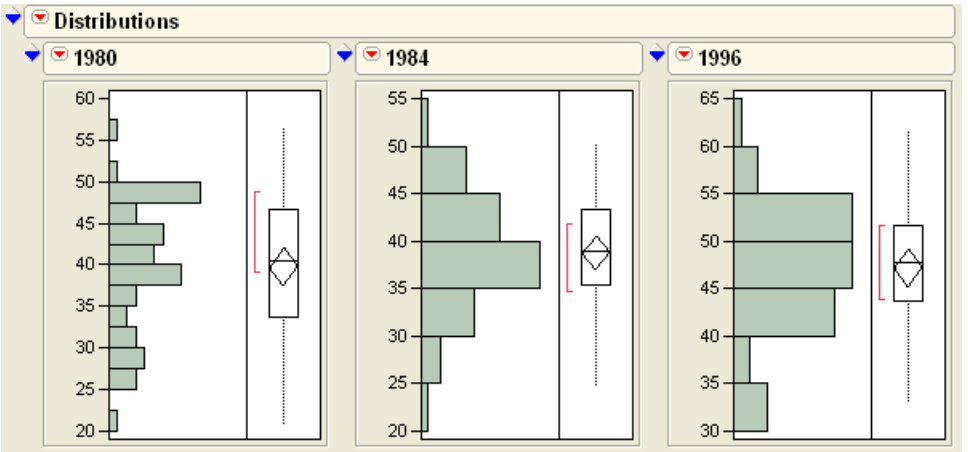
$$\text{Domestic \$} = 76.906592 + 0.2865119\ \text{Worldwide \$}$$

For Drama, the equation of the regression line is

$$\text{Domestic \$} = 60.516832 + 0.327075\ \text{Worldwide \$}$$

There seems to be a difference in the slopes for these two types. It may make sense to use a separate linear prediction for each type of movie or (as you will see in a later chapter) use movie type as a factor in the model.

2. a.



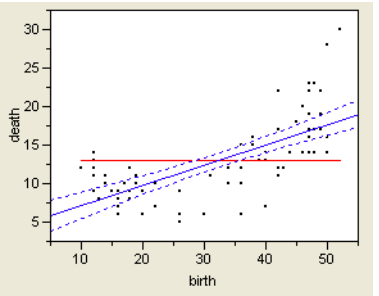
The mean for 1996 seems about 10 points higher than for the other two elections. Perhaps 1996 (Clinton vs. Dole) was a more contested election than the others (Reagan vs. Carter, Reagan vs. Mondale).

b. 1996 vs 1980: 0.593; 1984 vs. 1980: 0.704. Correlations are stronger for years that are closer together.

c. Probably not. The next presidential election is far removed from 1996, 1984, and 1980.

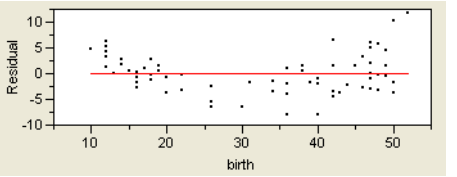
3. a. Upper Volta and Afghanistan are outliers for the death rate variable.

b.



The line is a better predictor than the mean, but a line is not an appropriate model (see part d of this question).

c.



There is a pattern to these residuals (it's U-shaped).

d. Because of the pattern of the residuals (and, actually, the pattern of the original data—clearly nonlinear), a line is not an appropriate model.

4. a. 1-Octonol

b. Right-click on the Parameter Estimates report and select **Columns> Lower 95%** and **Columns Upper 95%** to reveal the confidence interval we want.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	0.600659	0.054128	11.10	<.0001*	0.4927051	0.7086129
Hexane	0.8638495	0.034004	25.40	<.0001*	0.7960313	0.9316677

Chapter 11, "Categorical Distributions"

1. a. The Test Probabilities command from the Distribution report gives the following results:

Test Probabilities

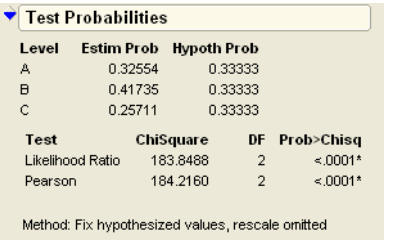
Level	Estim Prob	Hypoth Prob
Blue	0.21000	0.20000
Brown	0.22000	0.30000
Green	0.18000	0.20000
Red	0.23000	0.20000
Yellow	0.16000	0.10000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	6.0786	4	0.1934
Pearson	6.4333	4	0.1690

Method: Fix hypothesized values, rescale omitted

Based on these numbers, there's no reason to dispute the company's claim.

2. We entered the data into two columns of a data table, one representing the machines and the second representing the counts. Then, we selected **Analyze > Distribution** with Machine as **Y** and the counts as **Freq**. When the Distribution report appeared, we used **Test Probabilities** to test that they were all the same.



Test Probabilities

Level	Estim Prob	Hypoth Prob
A	0.32554	0.33333
B	0.41735	0.33333
C	0.25711	0.33333

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	183.8488	2	<.0001*
Pearson	184.2160	2	<.0001*

Method: Fix hypothesized values, rescale omitted

They clearly are different.

Chapter 12, "Categorical Models"

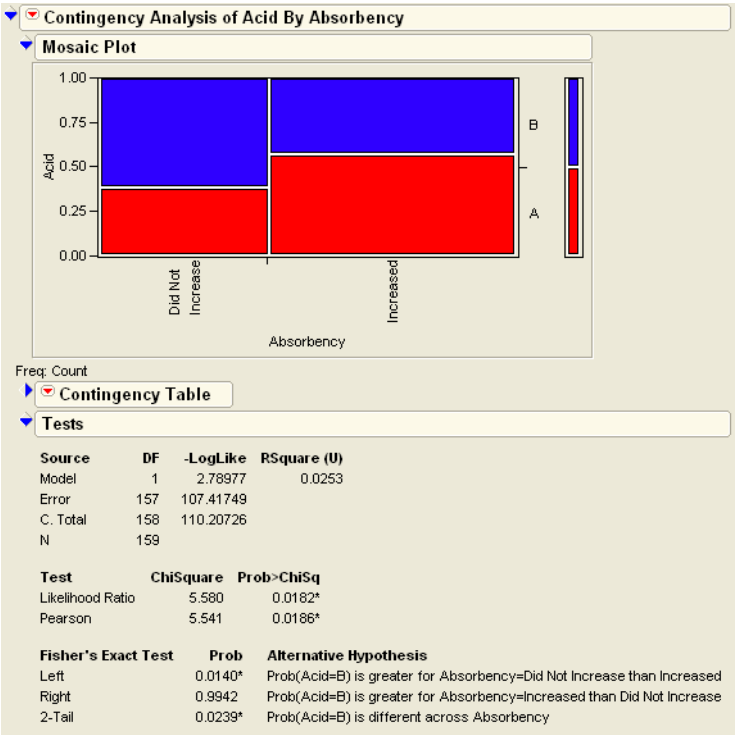
1. a. No

b. Yes

c. Looks: Yes; Money: No

d. Grades: No; Sports: No; Looks: No; Money: No

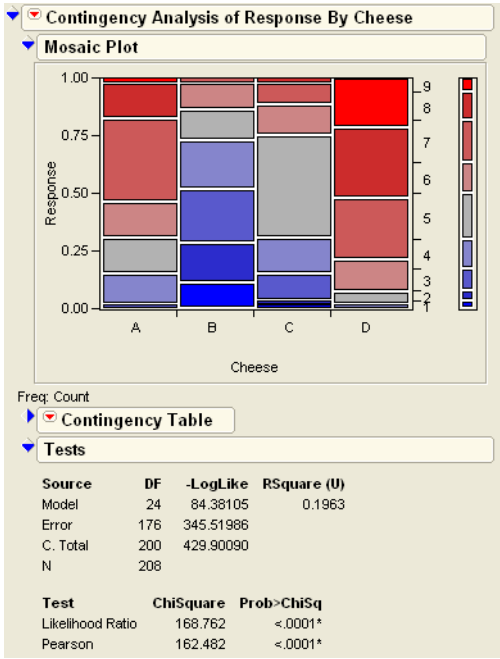
2.



There is evidence of a difference in absorbency between the two acids.

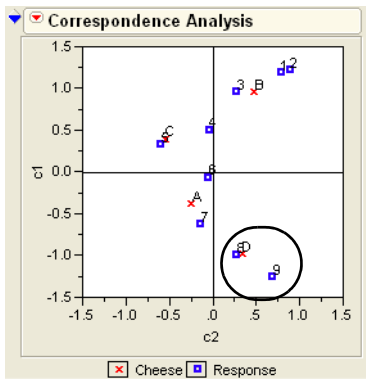
3. a, b:

The first two parts of this question can be answered with the following report, produced with Fit Y By X with Count as a **Freq** variable.



There is an obvious difference since the *p*-values are so low.

c.



Cheese D is associated with high tasted scores. A comes in second place.

4.

a. The Distribution platform tells us that there were 885 crew, 325 first class, 285 second class, and 706 third class.

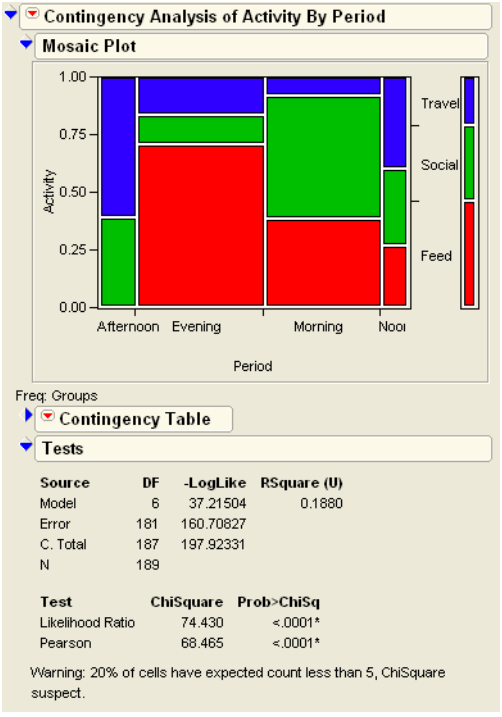
b. 109

c. 2201

d. Reject the null hypothesis of no difference with $p < 0.0001$.

e. Reject the null hypothesis of no difference with $p < 0.0001$.

5. a. Clear evidence of a difference in activity for different times of day



Chapter 13, "Multiple Regression"

1. a. 0.533

   b. In increases markedly, to 0.893.

2. b. The effect with the least significance is Shoulder, which has a p-value of 0.90.

   c. After several repetitions, Fore, Waist, Height, and Thigh remain in the model.

   d. We reach an identical model.

3. a. Dreaming and Non-Dreaming. Note that this model is degenerate since there is no variability in Dreaming that is not accounted for by the Non-Dreaming variable.

b. Dreaming and Exposure.

d. Forward stepwise gives a model with Dreaming, Gestation, and Danger.

e. Mixed stepwise gives the same model as part d.

4. a. There are many approaches. Using a forward stepwise model with 0.05 as the criteria to enter reveals that population is the only variable that enters into the model.

Chapter 14, "Fitting Linear Models"

1. a. The Summary of Fit and Analysis of Variance tables are the same in both platforms.

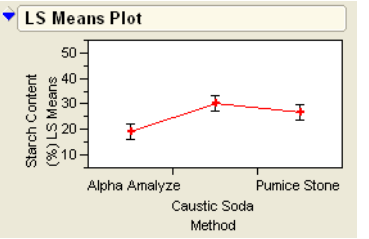
Summary of Fit

RSquare	0.216702
RSquare Adj	0.200211
Root Mean Square Error	8.636243
Mean of Response	25.51663
Observations (or Sum Wgts)	98

Analysis of Variance

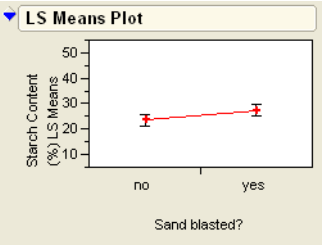
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1960.2363	980.118	13.1410
Error	95	7085.5450	74.585	Prob > F
C. Total	97	9045.7813		<.0001*

b. This plot (like the comparison circles) tell us that Caustic Soda removes the most starch.



c. All effects are significant.

d.

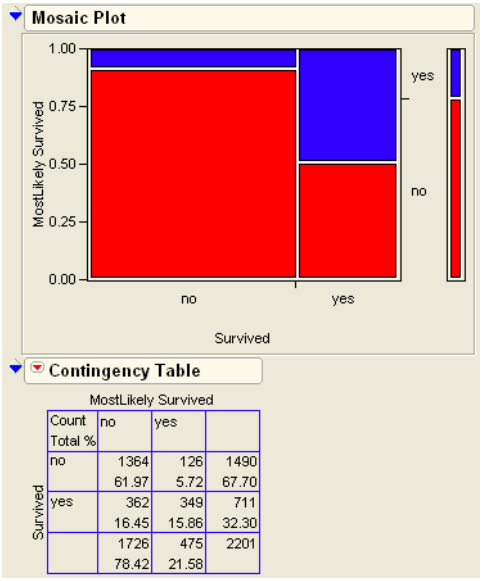


The Method LS means plot gives similar information as in part (b). The Sand Blasted plot says that sand blasting removes more starch than not.

e. It is ($F = 28.66$, $p < 0.001$).

f. No interaction effects are significant.

2. b.



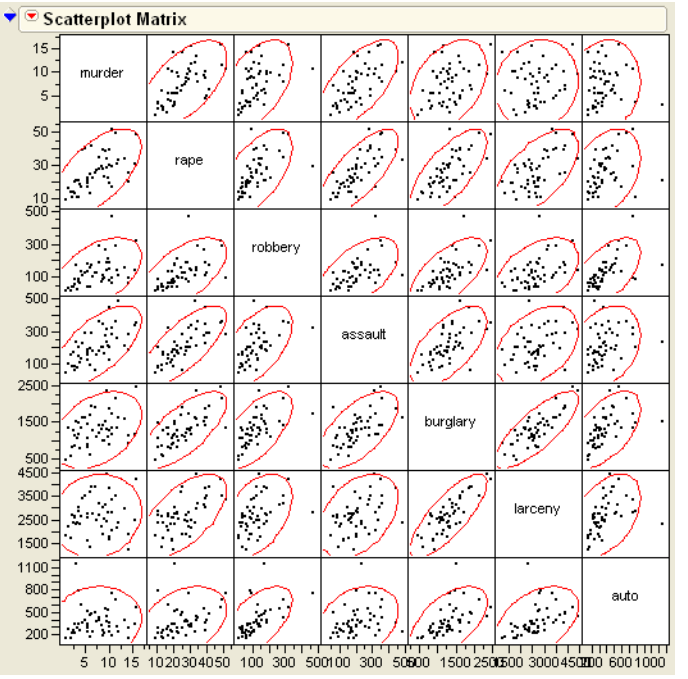
The model predicted correctly about 77% of the time.

c. The principle of Maximum Likelihood guarantees the most accurate predictions for the given data.

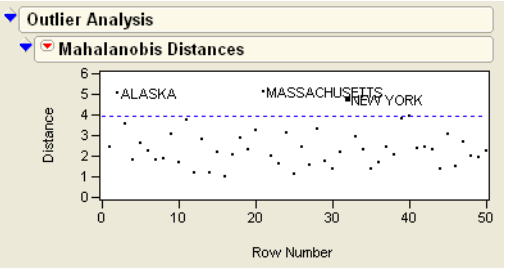
3. b. Long jump, high jump, 400m, 100m hurdles, 1500m

Chapter 15, "Bivariate and Multivariate Relationships"

1. a.



b.



c. Auto theft and robbery seem to appear in a different direction from the others.

d. 2 or 3

2. b. 2

c. Partial answer: Population and Employment are two variables that seem to occupy the same space.

Chapter 17, "Exploratory Modeling"

1.

a. Although the R2 value is not high, we have some weak conclusions that may warrant further study.

The first split tells us that gender plays an important role in predicting HDL cholesterol, and that men's HDL cholesterol tends to be lower than women's. For the men, the next most important factor was age—specifically, whether you are over or under 30. For the women, the next most important factor was weight—whether you weigh more or less than 128 lbs.

2.

a. Minimization of Time appears at Length=2.5, Freq=250, and Space = 0.721. This is from using the following model.

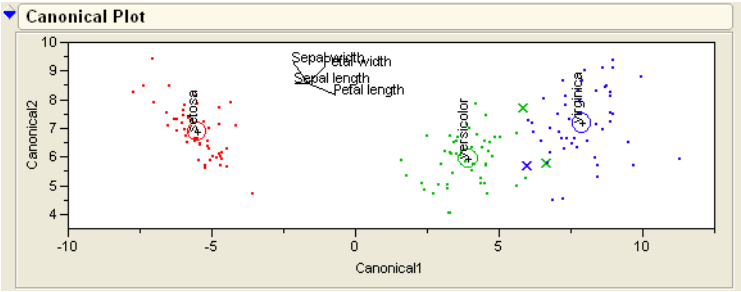
$$
\begin{bmatrix}
-3.2452922536306 \\
+3.45254884159199*\mathrm{Squish}\begin{bmatrix} 1.23349639217962 \\ +-4.0313426311121*Length \\ +-0.0079663004945*Freq \\ +5.55404075455771*Space \end{bmatrix} \\
+2.2135073851837*\mathrm{Squish}\begin{bmatrix} 19.6355924059005 \\ +1.24904647690406*Length \\ +-0.0469271850159*Freq \\ +-17.262879335114*Space \end{bmatrix} \\
+2.76379447848082*\mathrm{Squish}\begin{bmatrix} -8.2031099033942 \\ +-1.8298210665791*Length \\ +0.00754281425927*Freq \\ +13.2247711095988*Space \end{bmatrix}
\end{bmatrix} + 7.9375
$$

*4.6363118279061

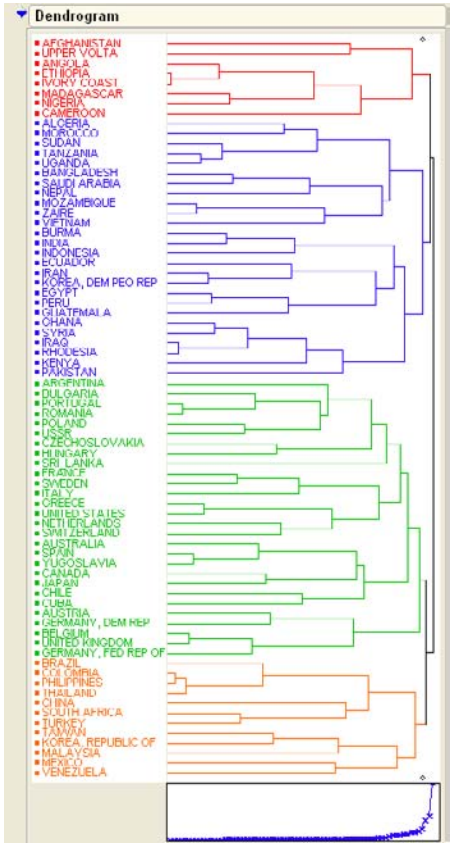
Chapter 18, "Discriminant and Cluster Analysis"

1.

With Species as X and the four measurements as Y, JMP produces a model that misclassifies only three flowers (these misclassifications are shown as Xs on the plot).



2. The Scree plot suggests around 4 clusters.



What looks similar among the clustered countries?

Chapter 20, "Time Series"

1. The AR(2) model is closer to white noise than the spiky AR(1). Both models could be improved.
2. AR(2)
3. AR(3,1)
4. ARI(2,1)
5. Suggestion: ARIMA(0,0,0)(1,1,1)

Technology License Notices

The ImageMan DLL is used with permission of Data Techniques, Inc.

SAS INSTITUTE INC.'S LICENSORS MAKE NO WARRANTIES, EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, REGARDING THE SOFTWARE. SAS INSTITUTE INC.'S LICENSORS DO NOT WARRANT, GUARANTEE OR MAKE ANY REPRESENTATIONS REGARDING THE USE OR THE RESULTS OF THE USE OF THE SOFTWARE IN TERMS OF ITS CORRECTNESS, ACCURACY, RELIABILITY, CURRENTNESS OR OTHERWISE. THE ENTIRE RISK AS TO THE RESULTS AND PERFORMANCE OF THE SOFTWARE IS ASSUMED BY YOU. THE EXCLUSION OF IMPLIED WARRANTIES IS NOT PERMITTED BY SOME STATES. THE ABOVE EXCLUSION MAY NOT APPLY TO YOU.

IN NO EVENT WILL SAS INSTITUTE INC.'S LICENSORS AND THEIR DIRECTORS, OFFICERS, EMPLOYEES OR AGENTS (COLLECTIVELY SAS INSTITUTE INC.'S LICENSOR) BE LIABLE TO YOU FOR ANY CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES (INCLUDING DAMAGES FOR LOSS OF BUSINESS PROFITS, BUSINESS INTERRUPTION, LOSS OF BUSINESS INFORMATION, AND THE LIKE) ARISING OUT OF THE USE OR INABILITY TO USE THE SOFTWARE EVEN IF SAS INSTITUTE INC.'S LICENSOR'S HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. BECAUSE SOME STATES DO NOT ALLOW THE EXCLUSION OR LIMITATION OF LIABILITY FOR CONSEQUENTIAL OR INCIDENTAL DAMAGES, THE ABOVE LIMITATIONS MAY NOT APPLY TO YOU. SAS INSTITUTE INC.'S LICENSOR'S LIABILITY TO YOU FOR ACTUAL DAMAGES FOR ANY CAUSE WHATSOEVER, AND REGARDLESS OF THE FORM OF THE ACTION (WHETHER IN CONTRACT, TORT (INCLUDING NEGLIGENCE), PRODUCT LIABILITY OR OTHERWISE WILL BE LIMITED TO $50.00.

Index

Q

R

S