“The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.”

Marcel Proust

Introductory Guide

JMP, A Business Unit of SAS
SAS Campus Drive
Cary, NC 27513
## Contents

**JMP Introductory Guide**

1. **Introducing JMP**
   - What You Need to Know .......................... 3
   - Learning About JMP ............................. 3
   - Conventions Used in this Book ................. 4
   - Step 1: Start JMP ............................... 5
   - Step 2: Open a JMP Data Table .................. 5
   - Step 3: Learn About the Data Table .......... 6
     - Data Table Cursor Forms ....................... 7
     - Selecting Rows and Columns .................. 9
     - Specifying the Values’ Type .................. 9
   - Step 4: Select an Analysis ...................... 10
     - Casting Columns Into Roles .................. 10
   - Step 5: View the Output Report ............... 11
     - Graphs and Charts ............................ 11
     - Statistical Tables and Text .................. 12
   - Step 6: Save the JMP Session ................. 13
     - A Practice Tutorial .......................... 14
       - Open a Data Table ......................... 14
       - Select an Analysis ......................... 15
       - Spin the Cowboy Hat ....................... 16
       - Experiment with the Display Options ...... 17
       - Exit JMP .................................... 18

2. **Creating a JMP Data Table**
   - Starting a JMP Session ......................... 21
   - Creating Rows and Columns in a JMP Data Table 22
     - Add Columns ................................ 22
     - Set Column Characteristics ................ 23
     - Add Rows .................................... 24
   - Entering Data .................................. 24
     - Plotting Data ................................ 25
     - Document the Report ......................... 27
   - Chapter Summary ................................ 28
3 Summarizing Data

Look Before You Leap ................................................................. 31
Grouping Data .............................................................................. 32
Creating Statistics for Groups ......................................................... 33
Charting Statistics from Grouped Data ............................................. 34
Charting Statistics for Two Groups ................................................ 36
Finding a Subgroup with Multiple Characteristics .............................. 37
Comparative Scatterplots ............................................................... 38
What Has Been Discovered? .......................................................... 39
Finding the Best Points .................................................................. 40
Chapter Summary .......................................................................... 41

4 Looking at Distributions

Look Before You Leap ................................................................. 45
Graphical Display of Distributions ................................................... 45
Histograms of Nominal and Ordinal Variables ................................. 46
Histograms of Continuous Values .................................................... 47
Mosaic Bar Charts for Ordinal and Nominal Variables ....................... 48
Outlier Box Plots for Continuous Variables ...................................... 48
Quantile Box Plots for Continuous Variables .................................... 49
Report Tables .................................................................................. 50
Reports for Continuous Variables ................................................... 50
Frequency Table for Ordinal or Nominal Variables ........................... 51
Creating a Subset ........................................................................... 51
Add a Computed Column ............................................................... 51
Create a New View .......................................................................... 54
Chapter Summary .......................................................................... 55

5 Comparing Group Means

Look Before You Leap ................................................................. 59
Graphical Display of Grouped Data .................................................. 59
Choose Variable Roles .................................................................... 60
Show Points .................................................................................... 61
Fit Means Option ............................................................................ 61
Fit Quantiles ................................................................................... 63
Comparison Circles ......................................................................... 64
Quantify Results ............................................................................. 65
Analysis of Variance ....................................................................... 66
Mean Estimates and Statistical Comparisons .................................... 66
Chapter Summary .......................................................................... 67
Credits

Credits and Acknowledgments

Origin
JMP was developed by SAS Institute Inc., Cary, N.C. JMP is not a part of the SAS System, though portions of JMP were adapted from routines in the SAS System, particularly for linear algebra and probability calculations. Version 1 of JMP went into production in October, 1989.

Credits
JMP was conceived and started by John Sall. Design and development were done by John Sall, Chung-Wei Ng, Michael Hecht, Richard Potter, Brian Corcoran, Annie Dudley Zangi, Bradley Jones, Craig Hales, Chris Gotwalt, Paul Nelson, and Wenjie Bao. In the SAS Institute Technical Support division, Ryan Gilmore, Wendy Murphrey, Toby Trott, Peter Ruzsa, Rosemary Lucas, and Susan Horton provide technical support and conducted test site administration. Statistical technical support is provided by Craig Devault, Duane Hayes, Elizabeth Edwards, and Kathleen Kiernan. Nicole Jones, Jianfeng Ding, Jim Borek, Kyoko Tidball, and Hui Di provide ongoing quality assurance. Additional testing and technical support is done by Noriki Inoue, Kyoko Takenaka, and Masakazu Okada from SAS Japan. Bob Hickey is the release engineer.

The JMP manuals were written by Ann Lehman, Lee Creighton, John Sall, Bradley Jones, Erin Vang, and Meredith Blackwelder, with contributions from Annie Dudley Zangi and Brian Corcoran. Editing, creative services, and production was done by SAS Publications. Melanie Drake implemented the help system.


Acknowledgments
We owe special gratitude to the people that encouraged us to start JMP, to the alpha and beta testers of JMP, and to the reviewers of the documentation. In particular we thank Michael Benson, Howard Yetter (d), Andy Mauromostakos, Xan Gregg, Al Best, Stan Young, Robert Muenchen, Lenore Herzenberg, Ramon Leon, Tom Lange, Homer Hedges, Skip Weed, Michael Emptage, Pat Spagan, Paul Wenz, Mike Bowen, Lori Gates, Georgia Morgan, David Tanaka, Zoe Jewell, Sky Alibhai, David Coleman, Linda Blazek, Michael Friendly, Joe Hockman, Frank Shen, J.H. Goodman, David Iklé, Lou Valente, Barry Hembree, Dan O Berman, Jeff Sweeney, Lynn Vanatta, and Kris Ghosh. Also, we thank Dick Devaux, Gray McQuarrie, Robert Stine, George Fractions, Avigdor Cahane, Jose Ramirez, Gudmundur Axelsson, Al Fulmer, Cary Tuckfield, Ron Thisted, Nancy McDavid, Veronica Czitrom, and Tom Johnson.

We also thank the following individuals for expert advice in their statistical specialties: R. Hocking and P. Spector for advice on effective hypotheses; Robert Mee for screening design generators; Jason Hsu for advice on multiple comparisons methods (not all of which we were able to incorporate in JMP); Ralph
Credits and Acknowledgments

O’Brien for advice on homogeneity of variance tests; Ralph O’Brien and S. Paul Wright for advice on statistical power; Keith Muller for advice in multivariate methods, Harry M artz, Wayne Nelson, Ramon Leon, Dave Trindade, Paul Tobias for advice on reliability plots; Lijian Yang and J.S. Marron for bivariate smoothing design; George Milliken and Yuri Bulavski for development of mixed models; Will Potts and Cathy Maahs-Fladung for data mining; Clay Thompson for advice on contour plotting algorithms; and Tom Little, Blan Godfrey, Tim Clapp, and Joe Ficalora for advice in the area of Six Sigma.

For sample data, thanks to Patrice Strahle for Pareto examples, the Texas air control board for the pollution data, and David Coleman for the pollen (eureka) data.

Translations

Erin Vang coordinated localization. Noriki Inoue, Kyoko Takenaka, and Masakazu Okada of SAS Japan were indispensable throughout the project. Special thanks to Professor Toshiro Haga (retired, Science University of Tokyo) and Professor Hirohiko Asano (Tokyo Metropolitan University). Finally, thanks to all the members of our outstanding translation teams.

Past Support

Many people were important in the evolution of JMP. Special thanks to Jeffrey Perkinson, David DeLong, Mary Cole, Kristin Nauta, Aaron Walker, Ike Walker, Eric Gjertsen, Dave Tilley, Ruth Lee, Annette Sanders, Tim Christensen, Jeff Polzin, Eric Wasserman, Charles Soper, Yusuke O no, and Junji Kishimoto. Thanks to SAS Institute quality assurance by Jeanne Martin, Fouad Younan, and Frank Lassiter. Additional testing for Versions 3 and 4 was done by Li Yang, Brenda Sun, Katrina Hauser, and Andrea Ritter.

Also thanks to Jenny Kendall, John Hansen, Eddie Routten, David Schlotzhauer, and James Mulherin. Thanks to Steve Shack, Greg Weier, and Maura Stokes for testing JMP Version 1.

Thanks for support from Charles Shipp, Harold Gugel (d), Jim Winters, Matthew Lay, Tim Rey, Rubin Gabriel, Brian Ruff, William Lisowski, David Morganstein, Tom Esposito, Susan West, Chris Fehily, Dan Chilko, Jim Shook, Ken Bodner, Rick Blahunka, Dana C. Aultman, and William Fehliner.

Technology License Notices

JMP for the Power Macintosh was compiled and built using the CodeWarrior C compiler from Metroworks Inc.
JMP uses an extraordinary graphical interface to display and analyze data. JMP is software for interactive statistical graphics and includes:

- a data table window for editing, entering, and manipulating data
- a broad range of graphical and statistical methods for data analysis
- an extensive design of experiments module
- options to highlight and display subsets of data
- a formula editor for each table column to compute values as needed
- a facility for grouping data and computing summary statistics
- special plots, charts, and communication capability for quality improvement techniques
- tools for printing and for moving analyses results between applications
- a scripting language for saving and creating frequently used routines

This introductory chapter gives basic information about using JMP.
Contents

What You Need to Know ......................................................... 3
Learning About JMP .......................................................... 3
Conventions Used in this Book .............................................. 4
Step 1: Start JMP ............................................................... 5
Step 2: Open a JMP Data Table ............................................. 5
Step 3: Learn About the Data Table ........................................ 6
  Data Table Cursor Forms .................................................. 7
  Selecting Rows and Columns .............................................. 9
  Specifying the Values’ Type ............................................... 9
Step 4: Select an Analysis ..................................................... 10
  Casting Columns Into Roles .............................................. 10
Step 5: View the Output Report ............................................ 11
  Graphs and Charts .......................................................... 11
  Statistical Tables and Text ............................................... 12
Step 6: Save the JMP Session ............................................... 13
A Practice Tutorial ............................................................. 14
  Open a Data Table .......................................................... 14
  Select an Analysis .......................................................... 15
  Spin the Cowboy Hat ...................................................... 16
  Experiment with the Display Options ................................. 17
  Exit JMP ................................................................. 18
What You Need to Know

Before you begin using JMP, you should be familiar with:

- Standard operations and terminology such as click, double-click, Ctrl-click and Alt-click (Command-click and Option-click on the Macintosh), Shift-click, drag, select, copy, and paste.

- How to use menu bars and scroll bars, how to move and resize windows, and how to manipulate files in the desktop. If you are using your computer for the first time, consult the reference guides that came with it for more information.

- Minimal statistics. Even though JMP has many advanced features, you only need a minimal background of formal statistical training. All analysis include graphical displays with options that help you review and interpret the results. Each analysis also includes access to help windows that offer general help and some statistical details.

Learning About JMP

There are three ways to learn about JMP:

Use JMP’s online help system

- Select Help from the main menu. On Windows and Linux, Help contains Contents, Search, and Index commands, which access the JMP documentation. On the Macintosh, it displays all the JMP documentation and also contains searching capabilities.

- Click the Index tab on the JMP Starter to access a scrolling alphabetical reference that tells you how to generate specific analyses using JMP and lets you access further help for that topic.

- Choose Help from windows and from drop-down menus in JMP reports.

- After you generate a report, select the help tool (?) from the Tools menu and click the report. Context-sensitive help appears and shows information about the items in the report window.

Use the JMP Introductory Guide’s tutorials

This book, the JMP Introductory Guide, is a collection of tutorials designed to help you learn JMP strategies. When you installed JMP, a folder named Sample Data was also installed near the application. Each tutorial uses a file from the Sample Data folder. By following along with these step-by-step examples, you can quickly become familiar with JMP menus, options, and report windows.

Use this book in combination with other included books

The book you are reading now is the JMP Introductory Guide. See the following manuals for further documentation of JMP:

- The JMP User’s Guide has complete documentation of all JMP menus, an explanation of data table manipulation, and a description of the formula editor. There are chapters that show how to do common tasks such as manipulating files, transforming data table columns, and cutting and pasting JMP data, statistical text reports, and graphical displays.

- The JMP Statistics and Graphics Guide gives documentation of the Analyze and Graph menus. It documents analyses, discusses statistical methods, and describes all report windows and options.
Introducing JMP—Conventions Used in this Book

- The JMP Design of Experiments covers the DOE menu, the experimental design facility in JMP.
- The JMP Scripting Guide is a reference guide to the JMP scripting language (JSL) that lets you automate action sequences.

Conventions Used in this Book

Conventions used in this manual were devised to help relate written material to information that appears on-screen:

- The .jmp extension follows file names on the PC. When you installed JMP, a folder named Sample Data was also installed near the application. On the Macintosh, JMP sample data files have the same name, but show without an extension. Reference to names of JMP files, data tables, variable names, and items in reports appear in Helvetica to help distinguish them from surrounding text.
- Special information, warnings, and limitations are noted in sentences beginning with the bold word Note.
- Reference to menu names (File menu) or menu items (Save command) appear in Helvetica bold font.
- The notation to select a command from a menu is sometimes written as File > New, meaning “select the New command from the File menu.”
- Words or phrases that are important or have definitions specific to JMP are in italics the first time they appear.
Step 1: Start JMP

Start a JMP session by double-clicking the JMP application icon. Your initial view of JMP will be a menu bar, a tool bar, and the JMP Starter window (Figure 1.1).

Figure 1.1 First View of JMP (Windows)

Step 2: Open a JMP Data Table

There are several ways to open a data table:

- Selecting File > New (or clicking the New Data Table button on the JMP Starter window) creates and displays a data table with an empty data grid. First, add rows and columns, then type in or paste in new data. For details, see “Creating New Data Tables,” p. 8 of JMP User’s Guide.

- Selecting File > Open (or clicking the Open Data Table button on the JMP Starter window) presents a file selection window (Figure 1.2) with a list of existing tables. Select a file and click Open. For details, see “Opening Existing JMP Files,” p. 9 of JMP User’s Guide.
Step 3: Learn About the Data Table

Opening or creating a data table creates a data grid and table information panels, like the ones shown in Figure 1.3. The counts of table rows and columns appear in the corresponding panels to the left of the data grid. In the data grid, a row number identifies each row, and each column has a column name. Rows and columns are sometimes called observations and variables in an analysis.

The JMP data table window is a flexible way to prepare data. Using it, you can accomplish a variety of table management tasks, such as:

- Editing the value in any cell
- Changing a column's width by dragging the column line
- Hiding columns temporarily, or deleting columns permanently
- Adding rows, or rearranging the order of rows
- Adding columns, or rearranging the order of columns
- Selecting a subset of rows for analysis and saving that subset for further use
- Sorting or combining tables

For details, see “Using Data Tables,” p. 45 of JMP User’s Guide.

Data Table Cursor Forms

As you move the cursor around the data table, it changes forms. Its shape gives you information about performable actions. The following sections describe the different cursor forms.

Arrow Cursor

The cursor displays as 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 icon (▼) or diamond-shaped disclosure button (▼ ▲ on Windows and ▼ ▲ on the Macintosh) or when it is in the upper-left corner of the data grid—the area where rows and columns are deselected (as shown here).

I-Beam Cursor

The cursor is an I-beam when it is over text in the data grid or highlighted column names in the data grid or column panels. To edit text in the data grid:
1. Click the cell you wish to edit. The cell highlights.
2. Click again next to any character to mark an insertion point.
3. The I-beam deposits a vertical blinking bar.
4. Use the keyboard to make changes.

To edit a column name:
1. Click the column name to highlight the column.
2. Press the Enter key to change the I-beam cursor to an insertion point.
3. Use the keyboard to make changes.
Introducing JMP—Step 3: Learn About the Data Table

**Large Cross Cursor**

The cursor becomes a large cross when moved into a column or row selection area. When moved over a column name, you can edit the name. To do so:

1. Double-click the column name. The name becomes highlighted.
2. Begin typing.

The cross cursor can also be used to select rows and columns. To select a column, click the area above the column name.

See the next section, "Selecting Rows and Columns" on page 9, for a detailed explanation of selecting rows and columns.

**Double Arrow Cursor**

The cursor changes to a double-arrow cursor when positioned on a column boundary or on a panel splitter. Dragging the double-arrow cursor changes the column width or the panel size.

**Hand Cursor**

The cursor changes to a hand when you move the mouse over a red triangle icon (🌸) or diamond-shaped disclosure button (烔烔 on Windows and ⚪️ on the Macintosh). Click the red triangle to reveal the menu and select a menu icon. Click the diamond icon to open or close a panel.
Selecting Rows and Columns

Select rows and columns in a JMP data grid by highlighting them, as explained in Figure 1.4 and shown in Figure 1.5. For additional details, see “Selecting Rows and Columns,” p. 68 of JMP User’s Guide.

Figure 1.4 Ways to Select Rows and Columns

<table>
<thead>
<tr>
<th>Action</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highlight a row</td>
<td>Click the space that contains the row number.</td>
</tr>
<tr>
<td>Highlight a column</td>
<td>Click the background area above the column name. Or, click the column name</td>
</tr>
<tr>
<td>in the columns panels to the left of the data grid.</td>
<td></td>
</tr>
<tr>
<td>Extend a selection of rows or columns</td>
<td>Shift-click the first and last rows or columns of the desired range.</td>
</tr>
<tr>
<td>Make a discontiguous selection</td>
<td>Ctrl-click (Command-click on the Macintosh) the desired selections.</td>
</tr>
</tbody>
</table>

Figure 1.5 Select Rows and Columns

Specifying the Values’ Type

The small icon to the left of the column name in the columns panel is a clickable menu icon. Use it to declare the modeling type of the values in the column. JMP uses three modeling type to determine how to analyze the column's values.
Introducing JMP—Step 4: Select an Analysis

- **Continuous** ( ) Values are numeric measurements.
- **Ordinal** ( ) Values are ordered categories, which can have either numeric or character values.
- **Nominal** ( ) Values are numeric or character classifications.

Modeling types are changeable depending on how you want to look at your data. For example, a variable like age should be specified continuous to find the mean (average) age, but nominal or ordinal to find frequency counts for each age value.

The default modeling type is nominal for character values and continuous for numeric values. To assign a different modeling type to your variables:

1. Click the icon next to the variable name.
2. Select the appropriate modeling type.

For details, see “Specifying Data Types and Modeling Types,” p. 54 of JMP User’s Guide.

---

Step 4: Select an Analysis

There are a variety of analyses available through the Analyze and Graph menus in the main menu. An alternate way to access these analyses is through toolbar buttons and selections in the JMP Starter window.

Selecting an analysis in the Analyze or Graph menus produces graphs, charts, plots, and/or tables. For example, to see a histogram of columns in the data table you have open, select Analyze > Distribution. Then, complete the window and click OK.

Casting Columns Into Roles

After you select an analysis from the main menu, a window appears that asks you to cast columns into roles. For example, if you select Analyze > Fit Y by X from the main menu, you will see the window in Figure 1.6.

Figure 1.6 Fit Y by X Window

The JMP analysis methods are like stages or platforms for variables to dramatize their values. Each analysis requires information about which variables play what roles in an analysis.
The most typical variable roles are:

- **Y, Response** Identifies a column as a response or dependent variable whose distribution is to be studied.
- **X, Factor** Identifies a column as an independent, classification, or explanatory variable whose values divide the rows into sample groups.
- **Weight** Identifies a numeric column whose values supply weights for each response.
- **Freq** Identifies a numeric column whose values assign a frequency to each row for the analysis.

### Step 5: View the Output Report

After you have cast columns into their roles, JMP provides output reports that include graphics and text. For more detail than is presented below, see “Working With Output Reports,” p. 157 of *JMP User’s Guide*.

### Graphs and Charts

JMP reports are usually filled with graphs, charts, plots, and other graphical displays that show your results. For example, if you select **Analyze > Distribution** and assign several columns the **Y, Response** role in the Distribution window that appears, you create a report containing a graphical display of each column assigned the **Y, Response** role.

For the example shown in **Figure 1.7**, the **Distribution** command produces graphical displays which include:

- Histograms of both the brand and speed columns
- An outlier box plot of the continuous variable speed

**Figure 1.7** Distribution Histograms and Outlier Box Plot
Display Options

To enhance the default graphical displays that show your results, JMP provides options that you can add to them. These options are found by clicking the red triangle icon beside a report name. For example, the red triangle icon next to the histogram name lists available report options (Figure 1.8). For practice, try selecting different combinations of these options and watch the effect they have on the displays and reports.

Figure 1.8 Options for Nominal or Ordinal Variable in a Distribution Analysis

Statistical Tables and Text

In addition to the graphs, charts, plots, and other graphical displays in a report, JMP can also include text tables in a report. The kinds of tables given depend on whether a variable is continuous or categorical (ordinal or nominal).

When you installed JMP, a folder named Sample Data was also installed near the application. In that folder is a file named Typing Data.jmp. Using the Distribution command to produce a distribution report outputs text tables for the speed and brand variables (Figure 1.9). In this example:

- For a continuous variable, JMP displays a Quantiles table and a Moments table.
- For nominal and ordinal variables, JMP displays a Frequency table showing the total sample frequency, category frequencies, and associated probabilities.
Figure 1.9 Statistical Reports

![Statistical Report Table]

JMP also gives you the ability to change the appearance of these tables. For details, see “Formatting Report Tables,” p. 170 of *JMP User's Guide*.

### Step 6: Save the JMP Session

To save a report just as it appears in the report window, use the *Journal* command (select **Edit > Journal**) to duplicate it in a separate window titled *Journal*. Then append other reports to it or manipulate it in the journal window. Then, select **File > Save As**. For more detail than is presented below, see “Saving Reports,” p. 34 of *JMP User's Guide*. 
A Practice Tutorial

Before you begin the tutorials in the following chapters of this book, complete this brief practice tutorial that is a short guided tour through a JMP analysis. Follow the steps to see a three-dimensional spinning plot.

Open a Data Table

- Open the file called Cowboy Hat.jmp to begin a JMP session. When you installed JMP, a folder named Sample Data was also installed near the application. In that folder is a file named Cowboy Hat.jmp.

The data table shown in Figure 1.10 appears.
Figure 1.10 Cowboy Hat Data Table

This data table has three numeric columns and two row state columns. Columns x and y are x- and y-coordinates, and z is created using the function

\[ z = \sin \sqrt{x^2 + y^2} \]

Select an Analysis

To plot the three columns of information from the Cowboy Hat data table:

1. Choose the Spinning Plot command from the Graph menu.

When the mouse button is released, the window shown in Figure 1.11 prompts for selection of columns for the Spinning Plot analysis.

1. Select the x, y, and z columns from the column selector list on the left side of the window, and click Y, Columns.

These column names now appear in the list on the right side of the window. Columns are also selectable by dragging them to the Y, Columns box.

1. Click OK.

The spinning plot appears. Initially, the data points look like a two-dimensional plot because the z dimension is projected onto the x-y plane.
Spin the Cowboy Hat

- Select the hand tool from the Tools menu or from the cursor tool bar.
- Position the hand tool on the cowboy hat spinning plot, press the mouse button, and move the hand about.
  
  The cowboy hat moves in three dimensions.

  The plot can also spin by itself.
  - Press the Shift key, and give the plot a push with the hand tool.
  - To stop the spinning plot, click again in the spinning plot frame.
  
  The plot's spin can also be controlled using the spin control panel, which contains 11 small icons. These icons control the spinning plot.
The six directional arrows are buttons that control the spin of the three-dimensional image. The plot spins in the direction indicated by the arrow. The spin continues as long as the mouse button is held.

- Shift-click a directional arrow to make the plot spin continuously.
- Click anywhere in the display to stop the spin.

The other five icons control spin rate, plot scale, and various display options:

- Clicking either the + or - angle icon increases or decreases the rate of the next spin, respectively.
- The two four-way arrows adjust the scale of the plot as indicated by their arrows.
- The red triangle icon accesses the display options described in the section “Experiment with the Display Options,” p. 17.
- The home icon resets the spin axes to their original position.
- Experiment with these buttons to become familiar with their function.

**Experiment with the Display Options**

The display options, shown here, tailor the spinning plot. By default, the background is black and the axes are visible. Change these conditions by clicking the red triangle icon and selecting an option. Each option switches on and off alternately when it is selected.
Complete the experiment with the spinning cowboy hat by trying out a few of the display options.

Exit JMP

Close the Spinning Plot window and the Cowboy Hat data table with File > Close or by clicking the close box in the corner of the active window.

To close JMP:

Select File > Exit (Windows/Linux), File > Quit (Mac OS 9), or JMP > Quit JMP (Mac OS X).
This lesson evaluates a new drug developed to lower blood pressure. Data were recorded over a six-month period for the following treatment groups:

- 300 mg dose
- 450 mg dose
- placebo
- control

Following are the mean monthly blood pressure for each group, recorded in a journal. This lesson shows how to enter data values into the data table and to create a single neat and informative line chart that shows the study results.

**Objectives**

- Create rows and columns in a data table, one at a time and in groups.
- Enter data into JMP.
- Create a chart using the Chart command.
- Rescale axes in a plot.
- Animate a plot.

### Blood Pressure Study

<table>
<thead>
<tr>
<th>Month</th>
<th>Control</th>
<th>Placebo</th>
<th>300mg</th>
<th>450mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>March</td>
<td>165</td>
<td>163</td>
<td>166</td>
<td>168</td>
</tr>
<tr>
<td>April</td>
<td>162</td>
<td>159</td>
<td>165</td>
<td>163</td>
</tr>
<tr>
<td>May</td>
<td>164</td>
<td>158</td>
<td>161</td>
<td>153</td>
</tr>
<tr>
<td>June</td>
<td>162</td>
<td>161</td>
<td></td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>166</td>
<td>158</td>
<td></td>
<td></td>
</tr>
<tr>
<td>August</td>
<td>163</td>
<td>158</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Starting a JMP Session

- Double-click the JMP application icon to begin a JMP session.
- Use the New command in the File menu to create an empty data table like the one shown here.

Note: The new table includes one column and no rows. Add rows and columns to the table as needed. The panels to the left of the data grid display information about its rows and columns and the table as a whole.

The data values for this project are blood pressure statistics collected over six months and recorded in a notebook page as shown in Figure 2.1. This raw data can be quickly keyed into a JMP data table.

Figure 2.1 Notebook of Raw Study Data

<table>
<thead>
<tr>
<th>Blood Pressure Study</th>
<th>Month</th>
<th>Control</th>
<th>Placebo</th>
<th>300mg</th>
<th>450mg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>March</td>
<td>165</td>
<td>163</td>
<td>166</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>April</td>
<td>162</td>
<td>159</td>
<td>165</td>
<td>163</td>
</tr>
<tr>
<td></td>
<td>May</td>
<td>164</td>
<td>158</td>
<td>161</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td>162</td>
<td>161</td>
<td>158</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>July</td>
<td>166</td>
<td>158</td>
<td>160</td>
<td>148</td>
</tr>
<tr>
<td></td>
<td>August</td>
<td>163</td>
<td>158</td>
<td>157</td>
<td>150</td>
</tr>
</tbody>
</table>
Creating Rows and Columns in a JMP Data Table

JMP data tables have rows and columns, sometimes called observations and variables in statistical terms. The raw data in Figure 2.1 are arranged as five columns (treatment groups) and six rows (months March through August). The first line in the notebook names each column of values. These names can be used as column names in the new JMP table.

Add Columns

One way to begin is to first create the number of rows and columns that are needed.

- Click the red triangle icon in the columns panel to the left of the data grid and select Add Multiple Columns.

Add Multiple Columns displays the window shown here, which prompts for the number of columns to add, where to add them, and which type of columns to add.
Ask for five new columns. The default column names are Column 1, Column 2, and so on, but these can be changed at any time in the columns panel or at the top of the column in the data table. So, the next step is to type in meaningful names. Use the names from the data journal—Month, Control, Placebo, 300 mg, and 450 mg.

Edit a column name by clicking it in the data grid or in the columns panel.

Begin typing once the name is highlighted.

Click anywhere in the text of the name to reposition the edit cursor for entering text. Alternatively, drag the bar to select a portion of the text for replacement.

Set Column Characteristics

Columns can have different characteristics. By default, they contain numeric data. However, the values for month in this example are character values. To change the Month column from numeric to character:
Creating a JMP Data Table—Entering Data

- Highlight the column (not just the column name) by clicking either in the area at the top of the column or the area beside its name in the columns panel.
- Select the Column Info command in the Cols main menu (Cols > Column Info) to display the window in Figure 2.2.
- Change Month to a character variable, as shown in Figure 2.2, by clicking the box beside Data Type.

The Column Info window is also used to change other column characteristics and to access the JMP formula editor for computing column values.

Figure 2.2 Change Data Type

Add Rows

Adding rows is also easy.

- Choose Rows > Add Rows.
- Ask for six new rows.

Alternatively, double-click anywhere in the body of the table to automatically fill it with new rows up through the position of the cursor.

- Select File > Save As to name the table BP Study.jmp and save it.

The data table is now ready to hold data values. To summarize the table evolution so far, you:

- Began with a new untitled table.
- Added enough rows and columns to accommodate the raw data.
- Tailored the characteristics of the table by giving the table and columns descriptive names.
- Changed the data type of the Month column to accept character values.

Entering Data

Entering data into the data table is as simple as typing values into their appropriate table cells.

- Type the values from the study journal (Figure 2.1) into the BP Study.jmp table as shown here.

When entering data into the data table:
Edit the cell value by moving the cursor into a data cell and double-clicking. The cursor becomes a flashing vertical bar.

Correct a mistake by dragging the text entry bar across the incorrect entry and typing the correction over it.

Press the Tab key to move the highlight one cell to the right. Press Shift-Tab to move the highlight one cell to the left. Moving the highlight 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.

Press the Return/Enter key on the keyboard to move the highlight down one cell. Press Enter on the numeric keypad to move the highlight to the right. Press Shift-Return/Enter to move the highlight in the opposite direction.

Plotting Data

The Analyze and Graph menu commands must know which columns to work with and what to do with them. This section shows how to plot the months across the horizontal (x) axis and the columns of blood pressure statistics for each treatment group overlaid on the vertical (y) axis.

Select Graph > Chart.

The window in Figure 2.3 appears.

Assign x and y roles and choose the type of chart. This example specification is for a bar chart, with data (as opposed to statistics) as chart points.

Assure that the default choice Vertical is selected from the chart type drop-down list.

Select the four continuous variables in the Select Columns list.

Click the Statistics button and select Data from the drop-down list.
Enter Month as the **X, Level** column.

Click **OK**.

**Figure 2.3** Assign Roles and Choose Information For Chart

JMP displays an overlaid bar chart of the data.

**Rescale the Plot Axis**

By default, y-axis scaling begins at zero and the overlay chart looks like the one shown here. But, to present easy-to-read information, the y-axis needs to be rescaled and the chart needs labels. To do this,

1. Double-click the y-axis area, which accesses the Axis Specification window (**Figure 2.4**).

This window gives the ability to:

- Set the minimum and maximum of the axis scale.
- Specify the tick mark increment.
- Request minor tick marks.
- Request grid lines at major or minor tick marks.
• Format numeric axes.
• Use either a linear or log-based scale.

**Figure 2.4** Axis Specification Window

In this example, the plotted values range from about 145 to 175.

- Enter these figures into the Axis Specification window for **Minimum** and **Maximum**.
- Also, the increment for the tick marks needs to change from 50 to 1.

- Enter a 1 in the **Increment** box.

Peek ahead at **Figure 2.5** to see the dramatic result of axis rescaling.

**Note:** The magnifier tool ( ), found in the **Tools** menu or the cursor tool bar, can also be used to change the scale of graphs. Drag the magnifier diagonally across the points of interest to see the chart automatically adjust. Alt-click the plot frame to reset the plot to its original scale.

To change the name of the axis from **Y** to **Blood Pressure**,

- Click the axis name in the report.
- Type the new name in the resulting edit box.

**Document the Report**

The chart also needs a title and other documentation to make it easy to interpret. The annotate tool ( ) places text on the report. Refer to **Figure 2.5** see where to place the following steps.

- Select the annotate tool.
- Click and drag in the report to create a text box.
- Release the button and enter the text for the title **Comparison of Treatment Groups**.
- Click outside the annotation to quit editing the text.
- Repeat to enter the footnote **XYZ Blood Pressure Study 2001**.

**Note:** Double-click any report title bar to edit the text on the bar.
Chapter Summary

A study was done to evaluate the effect of a new drug on blood pressure. To complete this analysis, you:

- Used the **New Data Table** command in the **File** menu to create a new JMP table.
- Created the appropriate number of rows and columns for the data.
- Typed the data into the empty data grid.
- Used the **Chart** command in the **Graph** menu to request a bar chart of blood pressure measures over time.
- Tailored the chart with a specific axis scale and axis name, and added a plot title and footnote with the annotate tool.

**Figure 2.5** Line Chart with Modified Y-Axis, Titles, and Footnotes
The hot dog is a questionable item on a school cafeteria menu because of its reputation as an unhealthy food, possibly classified in the junk food category. Many students feel this is unpatriotic and are upset. This lesson examines the hot dog as a menu item, but not before looking into the multitude of brands available. The data shows information about cost, nutritional ingredients of concern, and taste preference for 54 hot dog brands. This information is sufficient to provide a summary of hot dog statistics and to identify the brands that are:

- most nutritious
- least costly
- best tasting

The taste, cost, and nutritional variables used in this chapter are an enhancement of data from Moore, D. S., and McCabe G. P., (1989), *Introduction to the Practice of Statistics*, and *Consumer Reports* (1986). The brand names were changed to fictional names, and the taste preference labels correspond to a taste preference scale.

**Objectives**

- Find and mark subgroups of data
- Produce scatterplots using the **Fit Y by X** command and use them as discovery tools
- Label individual points in plots
- Produce and plot summary statistics
Contents

Look Before You Leap ......................................................... 31
Grouping Data ................................................................. 32
  Creating Statistics for Groups ........................................... 33
  Charting Statistics from Grouped Data ............................... 34
  Charting Statistics for Two Groups ................................. 36
Finding a Subgroup with Multiple Characteristics ................. 37
  Comparative Scatterplots ............................................... 38
  What Has Been Discovered? .............................................. 39
  Finding the Best Points ................................................... 40
Chapter Summary ............................................................. 41
Look Before You Leap

When you installed JMP, a folder named Sample Data was also installed near the application. In that folder is a file named Hot Dogs.jmp. Open the Hot Dogs.jmp file to see the data shown in Figure 3.1.

**Figure 3.1** Hot Dogs Data Table

<table>
<thead>
<tr>
<th>Product Name</th>
<th>Type</th>
<th>Taste</th>
<th>$/oz</th>
<th>$/lb Protein</th>
<th>Calories</th>
<th>Sodium</th>
<th>Protein/Fat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy Hill Sparer</td>
<td>Beef</td>
<td>Bland</td>
<td>0.11</td>
<td>14.23</td>
<td>186</td>
<td>495</td>
<td>1</td>
</tr>
<tr>
<td>Georgies Skinless Beef</td>
<td>Beef</td>
<td>Bland</td>
<td>0.17</td>
<td>21.70</td>
<td>181</td>
<td>477</td>
<td>2</td>
</tr>
<tr>
<td>Special Market’s Premium Beef</td>
<td>Beef</td>
<td>Bland</td>
<td>0.11</td>
<td>14.49</td>
<td>178</td>
<td>425</td>
<td>1</td>
</tr>
<tr>
<td>Silver’s Beef</td>
<td>Beef</td>
<td>Medium</td>
<td>0.15</td>
<td>20.48</td>
<td>140</td>
<td>322</td>
<td>1</td>
</tr>
<tr>
<td>Hungry Hugh’s Jumbo Beef</td>
<td>Beef</td>
<td>Medium</td>
<td>0.10</td>
<td>14.47</td>
<td>134</td>
<td>402</td>
<td>1</td>
</tr>
<tr>
<td>Great Dinner Beef</td>
<td>Beef</td>
<td>Medium</td>
<td>0.11</td>
<td>15.45</td>
<td>130</td>
<td>507</td>
<td>1</td>
</tr>
<tr>
<td>RJL Kosher Beef</td>
<td>Beef</td>
<td>Medium</td>
<td>0.21</td>
<td>25.25</td>
<td>123</td>
<td>370</td>
<td>2</td>
</tr>
<tr>
<td>Wonder Kosher Skinless Beef</td>
<td>Beef</td>
<td>Medium</td>
<td>0.20</td>
<td>24.02</td>
<td>138</td>
<td>322</td>
<td>2</td>
</tr>
<tr>
<td>Happy Kids Jumbo Beef</td>
<td>Beef</td>
<td>Medium</td>
<td>0.14</td>
<td>19.00</td>
<td>175</td>
<td>479</td>
<td>1</td>
</tr>
<tr>
<td>Midwest Beef</td>
<td>Beef</td>
<td>Medium</td>
<td>0.14</td>
<td>19.00</td>
<td>140</td>
<td>375</td>
<td>1</td>
</tr>
</tbody>
</table>

The Hot dogs.jmp table has the following information:

- The columns called Type, Calories, Sodium, and Protein/Fat (an index ratio of protein to fat) give information about nutrition. The Type column has values Meat, Poultry, and Beef.
- Cost information is in columns $/oz (dollars per ounce of hot dog) and $/lb Protein (dollars per pound of hot dog protein).
- Three categories of taste are coded Bland, Medium, and Scrumptious in the Taste column.

To get a feel for these data, use the **Distribution** command.

1. Select **Analyze > Distribution** to see the window in **Figure 3.2**.
2. Select all the variables except Product Name and click the **Y, Columns** button.
3. Click **OK**.

**Figure 3.2** Distribution Command

Examine the resulting report to see the distributions and levels of each variable.
Of course, health is a primary concern of a school cafeteria. It is interesting to see if type of hot dog plays a role in healthfulness. In particular:

- Which type of hot dog has the fewest calories?
- Is the amount of sodium different in the three types of hot dog?
- Which hot dogs have the highest protein content?
- Which hot dogs taste good and are healthy?

To address these issues, the data need to be grouped into hot dog type and taste preference categories with summary statistics computed for each group. The Summary command in the Tables menu is the JMP facility for grouping data and computing summary statistics.

The Summary command creates a JMP window that contains a summary table. This table summarizes columns from the active data table, called its source table. The Hot Dogs.jmp table is the source table in this example. A summary table has a single row for each level (value) of a specified variable.

Select Tables > Summary.
Select Type and click the Group button to see the window as shown in Figure 3.3.
Click OK.

Figure 3.3 Summary Window

The Hot Dogs By (Type) summary table (Figure 3.4) appears in a new window. The Type column lists hot dog type and the NRows column gives the frequency of each type in the source table.

A summary table is not independent of its source table. It has these characteristics:
When rows are highlighted in the summary table, their corresponding rows highlight in the source table.

The summary table is not saved when closed. Use File > Save As to specify a name and location for the table. JMP creates a standard data table from the summary table.

**Figure 3.4** Summary Table for Type of Hot Dog

<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th>N Rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beef</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Meat</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>Poultry</td>
<td>3</td>
<td>17</td>
</tr>
</tbody>
</table>

**Note:** To create an independent copy of the summary table, use the Tables > Subset command to copy the whole table.

**Creating Statistics for Groups**

Next, expand the summary table with columns of statistics. Summary tables have an additional command in the columns panel called Add Statistics Column that accesses the Summary window so statistical summary columns can be added to the table at any time.

To follow along with this example, do these steps:

- Click the red triangle icon and select the Add Statistics Column command from the columns panel on the left side of the screen.
- Select Calories, Sodium, and Protein/Fat in the column selector list of the window, as shown in Figure 3.5.
- Click the Statistics button and select Mean. You should now see new column names in the Statistics list.
- Click OK.
Figure 3.5 Summary Window

The new columns of statistics are displayed in the Hot Dogs By (Type) table (top table of Figure 3.6).

Repeat the previous steps to create a second summary table of Hot Dogs by Taste to look at health factors and hot dog tastiness.

The Hot Dogs By (Taste) summary table shows average calories, sodium, and protein-to-fat ratio for each taste category (bottom table of Figure 3.6).

Figure 3.6 Summary Statistics for Hot Dog Groups

Charting Statistics from Grouped Data

The summary tables in Figure 3.6 show the summary statistics in tabular form, but bar charts are better for visual comparison. The Chart command on the Graph menu can also summarize data and then create charts of the summarized data.

Make sure the Hot Dogs.jmp table is active.
Select **Graph > Chart**.

- Assign variable roles as shown in **Figure 3.7**.

Charts like those below should appear.

**Figure 3.7** Charting Data
It appears that poultry hot dogs have fewer calories on average than the other two hot dog types. Also note that the poultry hot dogs have slightly more sodium. The most visible difference is that the protein-to-fat ratio appears much higher in poultry hot dogs.

Repeat the above steps using taste to produce another set of bar charts.

It may not be surprising to see from the bar charts that hot dogs rated as bland tasting have (on average) more calories, more sodium, and a lower protein-to-fat ratio. It can be seen from the data table that scrumptious tasting hot dogs have the lowest average calories and sodium content. However, medium tasting hot dogs have the highest protein-to-fat ratio, and they compare well with respect to the other nutritional factors.

**Charting Statistics for Two Groups**

Next, it is useful to know the frequency of the three taste responses for each type of hot dog.

Make sure the Hot Dogs jmp table is active and use **Tables > Summary** again and select both Type and Taste as grouping variables.

This produces the table shown here. There is one row for each taste response within each type of hot dog. The Nrows column lists the frequency in the source table of each type-taste combination.
3 Summarizing Data—Finding a Subgroup with Multiple Characteristics

Select the Graph > Chart command, with both grouping variables as X Level and the Nrows column with the y (statistics: data) role. This produces the chart shown here. In this example, there are side-by-side charts that show the frequency for each taste within each type of hot dog.

**Note:** Graph > Chart can also be used to directly chart data grouped by two variables; the data don’t have to be grouped first by Tables > Summary.

To label each bar with the frequency it represents:

- Select the Rows > Row Selection > Select All Rows command.
- Select the Rows > Label command.

The chart above shows that the poultry hot dogs excelled in nutrition factors and that most people find them medium-tasting. However, because the sodium content appears slightly high in some poultry brands, more investigation is needed.

---

**Finding a Subgroup with Multiple Characteristics**

- Select Rows > Clear Row States to remove the labels from the source table rows (and the bar charts) and deselect any selected rows.

This prepares the summary table, Hot Dogs by (Type, Taste), for assignment of special markers to identify each type of hot dog. Continue the search for the ideal hot dog.

In the Hot Dogs by (Type, Taste) summary table:

- Shift-click or click and drag over the medium and scrumptious beef rows (2 and 3) to select them.
- Use the Markers command in the Rows menu to assign them the Z marker.
Deselect those rows.

Shift-click or drag the medium and scrumptious meat rows (5 and 6) to select them, assign them the Y marker, and deselect them.

Shift-click or drag the medium and scrumptious poultry rows (8 and 9), assign them the X marker, and deselect them.

The type-taste summary table now looks like the one shown here, and the corresponding rows in the Hot Dogs.jmp table are marked likewise.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Beef</td>
<td>bland</td>
</tr>
<tr>
<td>2</td>
<td>Beef</td>
<td>Medium</td>
</tr>
<tr>
<td>3</td>
<td>Beef</td>
<td>scrumptious</td>
</tr>
<tr>
<td>4</td>
<td>Meat</td>
<td>bland</td>
</tr>
<tr>
<td>5</td>
<td>Meat</td>
<td>Medium</td>
</tr>
<tr>
<td>6</td>
<td>Meat</td>
<td>scrumptious</td>
</tr>
<tr>
<td>7</td>
<td>Poultry</td>
<td>bland</td>
</tr>
<tr>
<td>8</td>
<td>Poultry</td>
<td>Medium</td>
</tr>
<tr>
<td>9</td>
<td>Poultry</td>
<td>scrumptious</td>
</tr>
</tbody>
</table>

**Comparative Scatterplots**

Now, examine the relevant variables with scatterplots to identify specific points (brands). The **Fit Y by X** command in the **Analyze** menu produces scatterplots when both the x and y are continuous numeric variables.

The following scatterplots graphically show the relationship of cost and the nutritional factors together.

- Deselect the rows in the **Analyze** menu.

Select **Analyze > Fit Y by X**.

Make your selections in the window, giving $/lb Protein to the y role and both $/oz and Protein/Fat to the x role.

Click **OK**.

This produces $/lb Protein by $/oz and a $/lb Protein by Protein/Fat scatterplots.

Choose **Type** as the grouping variable from the list of variables in the Grouping window.

Repeat this action for the $/lb Protein by Protein/Fat scatterplot.

Choose **Density Ellipses > .90** to make a density ellipse visible.

Repeat to complete a similar **Fit Y by X** analysis with Calories as y and Sodium as x.

These commands produce the $/lb Protein by $/oz, the $/lb Protein by Protein/Fat, and the Calories by Sodium scatterplots shown in **Figure 3.9** and **Figure 3.10**.

The 0.90 ellipses in the scatterplots show the shape of the bivariate response for each type of hot dog. The special markers identify the taste and type of each point.
Summarizing Data—Finding a Subgroup with Multiple Characteristics

Figure 3.9 Scatterplots Comparing Cost, Taste, and Nutritional Factors

To further identify and highlight points of interest:

- Select the brush tool ( ) from the Tools menu.
- Press the Alt key (Alt-shift on Linux and Option on Macintosh) and drag the brush in the lower left quadrant of the Calories by Sodium scatterplot, as shown in Figure 3.10.

These points represent brands with both low sodium and low calories. The highlighted points of these healthiest brands also highlight in the other scatterplots.

Figure 3.10 Select Low Sodium and Low Calorie Brands

What Has Been Discovered?

The costs of meat and beef brands range from low to high. However, it is not surprising to see the tight low-cost cluster of poultry brands (X-marked) at the lower left of the $/lb Protein by $/oz scatterplot. The highlighted points include poultry brands, one meat brand, and one beef brand. The selected beef point (Z-marked) is in the upper-right corner of the plot, which places it in the most expensive category. The single meat point (Y-marked) is more costly than the poultry brands but less than the beef brands.

A bigger surprise appears in the $/lb Protein by Protein/Fat scatterplot. As the protein-to-fat ratio increases, the cost per pound of protein stays about the same. Further, the poultry brands not only cost the least but also contain the most protein. Most of the selected points are in the three highest protein categories.
The density ellipses on the Calories by Sodium scatterplot show clearly that the poultry brands have about the same range of sodium content as the meat and beef brands, but many poultry brands have fewer calories.

**Finding the Best Points**

Now there is sufficient information to identify several hot dog brands as possible cafeteria menu items.

1. Click the Hot Dogs.jmp table to make it active.
2. Note that the Product Name column is designated as a Label column.

   ![Hot Dogs Table](image)

   The poultry (X-marked) brands are acceptably economical, and some of them have high protein content. Few meat or beef brands compared well.

1. Select the Arrow tool from the Tools menu.
2. Click inside the Calories by Sodium scatterplot to deselect all points.
3. Shift-click to highlight the two poultry brands with the least calories and lowest sodium content.
4. Again Shift-click to highlight the lone meat point (Y-marked) that has the least sodium of all brands, is low in calories, has a moderate protein count, and is average in price.
5. Select the Label/Unlabel command in the Rows menu to display the brand names of highlighted points (Thin Jack Veal, Calorie-less Turkey, and Estate Chicken), as shown in Figure 3.11.
Figure 3.11  Labeling Ideal Points

As a final step, use Analyze > Fit Y by X to look again at the two scatterplots that compare costs.
- Select Analyze > Fit Y by X.
- Assign $/lb Protein as Y.
- Assign both $/oz and Protein/Fat as X.
- Click OK.

The plot to the left in Figure 3.12 shows that the Estate Chicken brand is the most economical of the three labeled brands (showing $/oz as continuous). The plot to the right indicates that the Calorie-less Turkey brand is in the group with the highest proportion of protein (showing Protein/Fat as nominal).

Figure 3.12  Winning Hot Dog Brands.

Chapter Summary

This lesson examined different hot dog brands for a cafeteria menu. A JMP table has data for 54 brands of hot dog showing type of hot dog, taste preference, nutritional factors, and cost factors.

To find the ideal hot dog, we:
- Created a summary table that group the data by hot dog type and by taste preference within each hot dog type.
Summarizing Data—Chapter Summary

- Used Graph > Chart to chart summary statistics and identify the subset of hot dog brands that are both the most nutritious and the best tasting.
- Assigned different markers to each type of hot dog.
- Used Analyze > Fit Y by X to see scatterplots that compare cost factors and nutritional factors.
- Selected the points representing the lowest cost, most nutritious, and used the Label/Unlabel command in the Rows menu to identify the “Calorie-less Turkey” brand as a possible cafeteria hot dog.

Looking at Distributions

The students in a local school are participating in a health study. This lesson summarizes basic information about the students for the school system’s health care specialists. The data collected include age, sex, weight, and height.

Data summaries are needed to document the sample of participating students and identify any students with unusual characteristics who may need special attention. Therefore, this lesson produces reports with graphs and short, straightforward explanations.

Objectives

- Use the distribution analysis to explore several variables at once.
- Produce reports of moments, quantiles, frequencies, and proportions.
- Use the formula editor to compute a column’s value.
- Use the Tables > Subset command to create a subset of a data table.
4

Contents

Look Before You Leap ...................................................... 45
Graphical Display of Distributions ........................................ 45

- Histograms of Nominal and Ordinal Variables ....................... 46
- Histograms of Continuous Values ...................................... 47
- Mosaic Bar Charts for Ordinal and Nominal Variables ............. 48
- Outlier Box Plots for Continuous Variables ......................... 48
- Quantile Box Plots for Continuous Variables ....................... 49

Report Tables .................................................................. 50

- Reports for Continuous Variables .................................. 50
- Frequency Table for Ordinal or Nominal Variables ............... 51

Creating a Subset ................................................................ 51

- Add a Computed Column .............................................. 51
- Create a New View ..................................................... 54

Chapter Summary ............................................................ 55
Look Before You Leap

The first step in this analysis is to become familiar with the data in the Students.jmp file. Looking at the information in the JMP data table helps us decide which summary charts and tables to use in the health report.

Open the Students.jmp data table to see the data table shown in Figure 4.1.

**Figure 4.1** Students.jmp Data Table

The file contains the height, weight, age, sex, and an identification number for each student participating in the health study. The data table is in order by age, and sex is ordered within each age group. Even though there are only five columns of information, these variables address the following questions:

- How many boys and how many girls are there?
- How old are they?
- What is the average height and weight of the students?
- Are there any students drastically younger or older than the average age?
- Are there any students whose height or weight might signal the need for special attention?

**Graphical Display of Distributions**

To summarize the data:

Select the **Distribution** command from the **Analyze** menu.
In the window that appears, select the age and sex columns as **Y, Columns**.

- Click **OK**.

**Histograms of Nominal and Ordinal Variables**

After selecting **Analyze > Distribution** and completing the window, you will see a window that displays histograms for analysis variables. The histogram for ordinal or nominal variables like age and sex has a bar for each level (value) of the variable. Display options (described later) show additional plots and graphs such as the mosaic plot to the right of the histogram in **Figure 4.2** and **Figure 4.3**.

**Figure 4.2** Histogram and Mosaic Plot of the age Variable
Figure 4.3 Histogram and Mosaic Plot of the sex Variable

Histograms of Continuous Values

Continue by looking at the distribution of height and weight values in the sample of students.

- Click the Students data table to make it the active window.
- Again choose the Distribution command from the Analyze menu.
- Designate the height and weight columns as Y, Column variables.
- Click OK.

Histograms are displayed for height and weight, as shown in Figure 4.4.

Figure 4.4 Histograms of height and weight

Both height and weight appear to have approximately normal (bell-shaped) distributions, but notice the extremely high weight value. It will be examined more closely later.

It is important to present data in the best possible form. Sometimes it pays to experiment with the shape of a histogram by changing the number of bars or their arrangement on the axis.

To adjust the histogram bars:

- Select the hand from the graph cursor tool bar.
- Position the hand on the bars and press the mouse button to grab the plot.
Looking at Distributions—Graphical Display of Distributions

Move the hand to the left to increase the bar width and combines intervals (see “Graphs and Charts,” p. 11).

The number of bars decreases as the bar size increases.

Move the hand to the right to decrease the bar width, showing more bars.

Move the hand up or down to change the boundaries of the bins.

The height of each bar adjusts according to the new number of observations within each bin.

Exchange the hand for the arrow cursor after adjusting the bars.

Mosaic Bar Charts for Ordinal and Nominal Variables

The figure to the right is a mosaic plot for the age variable. A mosaic plot accompanies each frequency histogram for ordinal or nominal variables. The mosaic plot is a way to visualize the proportion of each ordinal or nominal level within the sample. The mosaic plot has a section for each level of the variable, where the size of the section is proportional to the corresponding group's size. Think of a mosaic plot as a bar chart with its bars stacked end to end.

Mosaic plots can be displayed or hidden by clicking the red triangle icon and selecting Mosaic Plot in the list of display options described in “Quantile Box Plots for Continuous Variables,” p. 49.

Outlier Box Plots for Continuous Variables

The outlier box plot (see Figure 4.5) is a schematic that shows the sample distribution and allows identification of points with extreme values, sometimes called outliers.

The ends of the box are the 25th and 75th quantiles, also called the quartiles. The difference between the quartiles is the interquartile range. The line across the middle of the box identifies the median sample value.

The lines extending from each end of the box are sometimes called whiskers. The whiskers extend from the ends of the box to the outermost data points that fall within the distance computed as quartile ± 1.5*(interquartile range). Points beyond the whiskers indicate extreme values that are possible outliers. To label a point, click the point to highlight it, and then select Rows > Label/Unlabel.

The bracket along the edge of the box identifies the shortest half, which is the most dense 50% of the observations.
Quantile Box Plots for Continuous Variables

In histograms, you can hide and show plot options by clicking the triangle icon and selecting from the menu that appears. For example, by clicking the triangle and selecting Quantile Box Plot, JMP gives the plot shown in Figure 4.6.

A quantile box plot shows the location of preselected percentiles, sometimes called quantiles, on the response axis. The median shows as a line in the body of the box. The ends of the box locate the 25th and 75th quantiles. The number of other quantile lines depends on the available space. The accompanying text report lists the data values for each of the standard quantiles. The box also contains a means diamond. The two diamond points within the box identify the 95% confidence interval of the mean. The line that passes through the two diamond points spanning the box identifies the sample mean.

Looking at the quantile box plot and means rectangle together helps see if data are distributed normally. If data are distributed normally (bell shaped), then the 50th quantile and the mean are the same and other quantiles show symmetrically above and below them.

Try different combinations of display options and watch the effect they have on the graphical displays and text reports.
Report Tables

Tables of statistical summaries are displayed with graphs. The tables that JMP produces depend on whether a variable is continuous, ordinal, or nominal. Click the red triangle icon and select **Display Options** to reveal available tables.

- For continuous variables, the **Display Options** submenu has commands for viewing or hiding quantiles and moments tables.
- For nominal and ordinal variables, the **Display Options** submenu has the **Frequencies** command for viewing or hiding a frequency table.

The diamond-shaped disclosure button (↑ ☰ on Windows and Linux and ▲ ☰ on the Macintosh) at the top-left of each report opens and closes it.

### Reports for Continuous Variables

The figure to the right is the report for the continuous variable height. The disclosure buttons open the report tables showing Quantiles and Moments.

- The Quantiles table displays the maximum value, minimum value, and other values for selected quantiles.
- The Moments table displays the mean, standard deviation, and other summary statistics.

**Note:** The **Test Mean** option from the menu on the variable’s title compares the variable’s mean to any specified constant and gives a table that shows the resulting t-test. There is also a test for the standard deviation.
Frequency Table for Ordinal or Nominal Variables

The report for nominal and ordinal variables has a different table from those produced for continuous variables. The report window for the nominal or ordinal (categorical) variables sex and age has frequency tables that show these items:

- **Level** lists each value of the response variable.
- **Count** lists the number of rows found for each level of a response variable.
- **Prob** lists the probability of occurrence for each level of a response variable. The probability is computed as the count divided by the total frequency of the variable, shown at the bottom of the table.

The following two statistics are not displayed by default. Right-click (Ctrl-click on the Macintosh) the table and select the **Columns** menu to reveal them.

- **StdErr Prob** lists the standard error of the probabilities.
- **Cum Prob** contains the cumulative sum of the column of probabilities.

Creating a Subset

Be on the alert for any unusual subjects, such as students who have extreme height or weight values. A good indicator of extreme values is the ratio of weight to height.

Add a Computed Column

To examine the ratio of weight to height, create a new column called ratio, computed as weight divided by height. To do this:

- Click the Students data table to make it the active window.
- Click the red triangle icon in the columns panel and select **New Column**.
To create the new column of weight-to-height ratios, complete the New Column window as in Figure 4.7.

1. Type the new name, ratio, in the Column Name area.
2. The default data type is Numeric and is correct as is.
3. The modeling type is Continuous and is correct as is.
4. Click the drop-down menu beside Format and set the format for ratio in the data grid to Fixed Dec with two decimal places.
5. Click the New Property button and select Formula, as shown in Figure 4.7, to compute values for the new column.

Figure 4.7 New Column Window

Construct the formula that calculates values for the ratio column as follows:

1. Highlight the empty term in the formula and select weight from the list of column names in the upper-left corner of the formula editor.
2. Press the divide (÷) key on the formula editor keypad.
3. With the empty denominator term highlighted, select height from the list of column names.
When the formula is complete, click **Apply** or **OK** on the formula editor, or just close its window. The new column called *ratio* is now in the *Students* data table as shown here. Its values are the computed weight-to-height ratio for each student.

![Formula Editor](image)

Now look at the distribution of the *ratio* variable.

- Select **Analyze > Distribution** and assign the new column (*ratio*) to the **Y, Columns** role.
- Click **OK**.

One way to identify subjects that have extreme values is to highlight histogram bars for the highest and lowest values. To highlight more than one bar, press the Shift key and click the desired bars.

- Select the bars for the two lowest and one highest bar (Figure 4.8).

**Note**: In the histogram shown in Figure 4.8, the plot frame was resized. To resize any plot, move the cursor to the lower right of the plot frame until it appears as a diagonal double-arrow, and then drag to change the plot size.
Looking at Distributions—Creating a Subset

**Figure 4.8** Histogram of Ratio with Bars Highlighted at Extreme Values

The highlighted bars in the histogram represent a ratio either greater than or equal to 3.5 or less than 1.5. The corresponding points automatically highlight in the data table and in all other reports generated from the Students data table.

Create a New View

Looking in the Students data table allows examination of the selected rows, but scrolling through a large data grid can be tedious. For the final report to the health researchers, include a separate list containing only the highlighted students—those with extreme values. To do this, use **Tables** menu commands to create new data tables or modify existing tables.

- Select **Tables > Subset** or click the red triangle icon in the tables panel and select **Subset**.
- Click **OK** to accept the default choices presented in the window.

This creates a new data table that has only the selected rows and columns from the active data table. The new data table, shown in **Figure 4.9**, contains only the students that have extreme weight-to-height ratios. By default, the table is named Subset of Students. Change the name by clicking the existing name (Subset of Students) in the panel located on the top left side of the window. The
table can be saved as a JMP file, exported for use in another application, or printed from the JMP session.

**Figure 4.9** Data Table Containing a Selected Subset

<table>
<thead>
<tr>
<th>Subset of Students</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>3</td>
<td>11</td>
<td>F</td>
<td>57</td>
<td>63</td>
</tr>
<tr>
<td>Distribution Age and Sex</td>
<td>3</td>
<td>11</td>
<td>F</td>
<td>64</td>
<td>60</td>
</tr>
<tr>
<td>Distribution Height and Weight</td>
<td>3</td>
<td>11</td>
<td>F</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>Source</td>
<td>4</td>
<td>11</td>
<td>F</td>
<td>62</td>
<td>65</td>
</tr>
</tbody>
</table>

**Chapter Summary**

In this chapter, the demographic and vital data of students participating in a health study were summarized. The profile was completed using the `Distribution` command and the data management features of the JMP data table.

The `Distribution` command displayed histograms and box plots or stacked bar (mosaic) bar charts for each variable assigned the role of response variable (y). Using display and text report options to look more closely at the data, the following actions were completed:

- Adjusted the number of bars and the scale of the histograms
- Produced supporting statistical reports showing moments and quantiles of numeric variables and frequencies and proportions of nominal and ordinal variables
- Created a new column in the data table computed as a function of existing columns
- Highlighted histogram bars to identify a subset of rows in the data table
- Created a new data table from a subset of highlighted rows

Graphs and text reports can be printed directly from JMP. Graphs and reports can be copied to a JMP journal or into other applications to complete a report for the school system health care specialists. See the chapter “Univariate Analysis” in the JMP Statistics and Graphics Guide for more information about distributions.
Comparing Group Means

In keeping with a recent corporate policy to modernize operations, all the typewriters in the typing pool are to be replaced with modern word processors. The typists are eager for this change and willingly participated in a study to help decide what kind of equipment to buy. The company selected three different brands of machine to test. These machines were randomly assigned to three groups of typists with comparable typing skills. The typists completed typing tests and recorded their words-per-minute scores.

This lesson finds out if the typing scores are significantly better on any one brand of machine than on the others.

Objectives

• Use the Fit Y by X command to produce plots and analyses appropriate for a one-way analysis of variance.
• Use JMP’s interactive capabilities to examine differences among groups.
• Produce text reports to display differences among groups.
Look Before You Leap

The first step is to become familiar with the data. The typing test scores are in a JMP file so that they can be reviewed and the kind of analysis determined.

When you installed JMP, a folder named Sample Data was also installed near the application. In that folder is a file named Typing Data.jmp. Open the file Typing Data.jmp.

The Typing Data table appears in the form of a data grid, as shown here.

![Typing Data Table](image)

The data table has columns named brand and speed. The modeling type for each column shows to the left of each column name in the columns panel. The character variable brand has nominal (nom) values and the numeric variable speed has continuous (cont) values.

There are 17 rows that represent typing scores for 17 typists. However, the number of participants in the groups differs because some of the scheduled participants did not show up for the study. Perhaps other statistics for the groups differ also. In particular:

- Is the mean (average) typing speed the same for each brand?
- Do any one of the three brands of word processor stand out from the others?
- Does it make a difference as to which brand the typists use?

Graphical Display of Grouped Data

Comparing the mean typing scores of each word processor brand involves analyzing two variables, so use the Fit Y by X command from the Analyze menu.
Comparing Group Means—Graphical Display of Grouped Data

**Note:** Selecting **Fit Y by X** allows you to perform:

- Categorical analysis when both \( x \) and \( y \) have nominal or ordinal values
- Analysis of variance when \( x \) is nominal or ordinal and \( y \) has continuous values, as in the example shown here
- Logistic regression when \( x \) is continuous and \( y \) has nominal or ordinal values
- Regression analysis when both \( x \) and \( y \) have continuous values.

### Choose Variable Roles

To discover if typing speed is related to (dependent on) a brand of word processor, follow these steps:

- Choose **Analyze > Fit Y by X**.
- Select brand as **X, Factor** and speed as **Y, Response** (see Figure 5.1).
- Click **OK**.

The plot shown in Figure 5.2 appears.

**Figure 5.1** The Fit Y by X Window

Selecting **Fit Y by X** and completing the window produces a statistical analysis appropriate for the variable roles (\( x \) and \( y \)) and the modeling type (continuous and nominal or ordinal) of each variable.

- **Y, Response** identifies a response (dependent) variable.
- **X, Factor** identifies a classification (independent) variable.

The next step is to choose an analysis that investigates if there is a statistical difference between the group mean values.
Show Points

Each of the typing test scores is plotted for each brand of word processor. Note that the distance between tick marks on the brand axis is proportional to the sample size of each group. The mean typing score for the total sample is shown as a horizontal line across the plot.

Figure 5.2 One-way Analysis for scores by brand

It is easy to see at a glance that most participants who used the SPEEDYTYPE machines typed faster than the others.

Fit Means Option

Now look at more graphical information about the distribution of typing scores by using commands accessed by clicking the red triangle icon on the title bar at the upper-left of the plot. Note that the Display Options submenu initially shows only the Points, Grand Mean, and X Axis proportional options in effect. Most of the options are toggles, which means the option turns alternately on and off each time it is selected.

Click the red triangle icon on the title bar and select Means/Anova.

This produces the appropriate analysis of variance reports. It automatically activates the Mean Diamonds option from the Display Options submenu, which draws a 95% means diamond for each group, as shown in Figure 5.3.
Comparing Group Means—Graphical Display of Grouped Data

Figure 5.3 Example of the Means Diamond Option

The means diamond has a line drawn at the mean (average) value of words-per-minute for each brand of word processor.

The upper and lower points of the means diamond span a 95% confidence interval computed from the sample values for each machine.

The width of each diamond spans the distance on the horizontal axis proportional to the group size.

Overlap lines within each diamond are drawn at ±(\sqrt{2}/2)C1. For groups with equal sample sizes, the marks that appear not to overlap indicate that two group means could be significant at the 95% confidence interval.

Figure 5.4 Means Diamond with X Axis Proportional Option Turned On (Left) and Off (Right)

The mean scores of the REGAL and WORD-O-MATIC word processors appear to be nearly the same, but note that the SPEEDYTYPE mean is much higher (Figure 5.3).
**Fit Quantiles**

The next logical step is to check the distribution of points within each group. This gives a better idea of the spread of the values and shows the distance of extreme values from the center of the data. The **Quantiles** option accessed by clicking the red triangle icon is useful for making this comparison.

- Click the red triangle icon and select **Quantiles**.

This displays the report in **Figure 5.5**, which lists the standard percentiles for each word processor. The median (50th percentile) is the typing speed that divides the sample in half. This means that 50% of the typists had speeds greater than the median, and the other half had lower speeds.

**Figure 5.5 Quantiles Table**

<table>
<thead>
<tr>
<th>Level</th>
<th>Minimum</th>
<th>10th</th>
<th>25th</th>
<th>Median</th>
<th>75th</th>
<th>90th</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGAL</td>
<td>56</td>
<td>65</td>
<td>68</td>
<td>70.5</td>
<td>72</td>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td>SPEEDY TYPE</td>
<td>77</td>
<td>77</td>
<td>78</td>
<td>84</td>
<td>87</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>WORD-O-MATIC</td>
<td>81</td>
<td>61</td>
<td>61.25</td>
<td>64.25</td>
<td>77</td>
<td>77</td>
<td></td>
</tr>
</tbody>
</table>

The **Quantiles** command automatically selects the **Quantile Boxes** display option, which overlays a quantile box plot on each group of typing scores, as shown in **Figure 5.6**.

**Figure 5.6 Fit Quantiles Option**

**Figure 5.7** illustrates the quantile box plot. The median, or 50th quantile, shows as a line in the body of the box. The top and bottom of the box represent the 75th and 25th quantiles, also called the upper and lower quantiles. The box encompasses the interquantile range of the sample data. The 10th and 90th quantiles show as lines above and below each box.

Looking at the quantile box plot and the means diamond together helps show if data are distributed normally within a group. If data are normally distributed (bell shaped), the 50th percentile and the mean are the same and the other quantiles are arranged symmetrically above and below the median.
Comparing Group Means—Graphical Display of Grouped Data

**Figure 5.7** Quantiles Box Plot

The quantile box plots (Figure 5.6) show a difference in variation of scores across the three groups. The scores in the REGAL group cluster tightly around the mean score but the WORD-O-MATIC scores show much more variation. However, even with this variation among the groups, the SPEEDYTYPE brand still appears to promote the best performance.

**Comparison Circles**

To complete the typing data inspection:

- Click the red triangle icon and choose **Compare Means > All Pairs, Tukey HSD**.

This option produces statistical reports (discussed later) and automatically selects **Comparison Circles** in the Display Options submenu. It also draws a set of comparison circles to the right of the plot that provides a graphical test of whether the mean typing scores are statistically different. Comparison circles for the three word-processor groups are shown in **Figure 5.8**.

The center of each circle is aligned with the mean of the group it represents. For the Student's t-test, the diameter of each circle spans the 95% confidence interval for each group. Whenever two circles intersect, the confidence intervals of the two means overlap, suggesting that the means may not be significantly different. Whenever two circles do not intersect, the group means they represent are significantly different.

- Click the SPEEDYTYPE comparison circle.

This graphically illustrates that the SPEEDYTYPE machine is statistically better than the other machines. The comparison circles highlight to show the statistical magnitude of the difference between typing scores. Circles for groups that are statistically the same have the same color.
The comparison circle for the SPEEDYTYPE brand does not intersect with either of the other two. The REGAL and WORD-O-MATIC brands are statistically slower than SPEEDYTYPE but do not appear different from each other. A later section, “Mean Estimates and Statistical Comparisons,” p. 66, discusses the multiple comparison tests the comparison circles represent.

Quantify Results

Now, examine the report beneath the plot that consists of several tables. The Summary of Fit table, shown here, summarizes the typing data distribution with these statistics:

- **Rsquare (R²)** quantifies the proportion of total variation in the typing scores resulting from different word processors rather than from different people.
- **Rsquare Adj** adjusts R² to make it more comparable over models with different numbers of parameters.
- **Root Mean Square Error (RMSE)** is a measure of the variation in the typing scores that can be attributed to different people rather than to different machines.
- **Mean of Response** is the mean (average) of all the typing scores.
- **Observations** is the total number of scores recorded.
Comparing Group Means—Quantify Results

Analysis of Variance

The second table given by the Means/Anova command is a standard analysis of variance table. If there are only two group levels, the report also includes a t-test table.

Note that the value of the F-probability (Prob>F) for the Analysis of Variance is 0.0004. This implies that differences as great as seen in this typing trial are expected only four times in 10,000 similar trials if the word processors did not really promote different typing performances.

The Analysis of Variance table has the following information:

- **Source** lists the sources of variation: brand, Error, and C. Total.
- **DF** is the degrees of freedom associated with the three sources of variation.
- **Sum of Squares** (SS for short) identifies the sources of variation in the typing scores.
- **C. Total** is the corrected total SS. It divides (partitions) into the SS attributable to brand and the SS for Error. The brand SS is the variation in the typing scores explained by the analysis of variance model, that hypothesizes the word processors are different. The Error SS is the remaining or unexplained variation.
- **Mean Square** is the sum of squares divided by its associated degrees of freedom.
- **F Ratio** is the model mean square divided by the error mean square.
- **Prob>F** is the probability of obtaining a greater F-value if the mean typing scores for the word processors differed only because different people were typing on them rather than because the word processors promoted different scores in any way.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>brand</td>
<td>2</td>
<td>630.83256</td>
<td>264.463</td>
<td>14.6634</td>
<td>0.0004</td>
</tr>
<tr>
<td>Error</td>
<td>14</td>
<td>255.00000</td>
<td>19.230</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>19</td>
<td>784.23256</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean Estimates and Statistical Comparisons

To see the list of means for each group, look at the Means for Oneway Anova table. This table summarizes the scores for each brand and reveals what level of performance to expect.

The Means for Oneway Anova table shows the following information:

- **Level** lists the name of each group.
- **Number** is the number of scores in each group.
- **Mean** is the mean of each group.
- **Std Error** is the standard error of each group mean.
- **Lower 95%** is the lower 95% confidence interval for the group means.
- **Upper 95%** is the upper 95% confidence interval for the group means.
The `Compare Means` command gives several multiple comparison options to statistically compare pairs of groups. This example uses the `All Pairs, Tukeys HSD` option, which performs a statistical means comparison for the three pairs of means using the Tukey-Kramers HSD (honestly significant difference) test (Tukey 1953, Kramer 1956). This means comparison method compares the actual difference between group means with the difference that would be significantly different. The difference needed for statistical significance is called the LSD (least significant difference).

The graphical results show as the comparison circles previously seen in Figure 5.8. The circles' centers represent the actual difference in the group means. The corresponding report is the Means Comparisons table, which shows the actual absolute difference between each mean and the LSD. The top half of the report gives information based on a Student's t comparison of each pair. The bottom half shows the results of the Tukey-Kramer multiple comparison tests. Pairs with a positive value are significantly different. The Means Comparison table confirms the visual results in Figure 5.8.

### Means Comparisons

<table>
<thead>
<tr>
<th>Level</th>
<th>Mean</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEEDTYPE</td>
<td>90.800</td>
<td>87.909</td>
<td>93.691</td>
</tr>
<tr>
<td>REGAL</td>
<td>70.300</td>
<td>67.409</td>
<td>73.191</td>
</tr>
<tr>
<td>WORD-O-MATIC</td>
<td>65.400</td>
<td>62.509</td>
<td>68.291</td>
</tr>
</tbody>
</table>

Positive values show pairs of means that are significantly different.

Chapter Summary

In this chapter, the difference in mean typing scores for three brands of word processor was summarized using the `Fit Y by X` command in the `Analyze` menu. This command was also used to:

- Plot the typing scores for the three brands of word processor.
- Overlay a means diamond on each group of typing scores to compare the means of each group.
5 Comparing Group Means—Chapter Summary

- Overlay a quantile box plot on each group of typing scores to compare the shape of the distribution of scores in each group.
- Produce comparison circles to visualize the difference in mean typing scores.
- Compute and display a one-way analysis of variance table, which confirmed that at least one pair of means is statistically different.
- Display a table of the group means and standard errors.
- Display a table showing the multiple comparison statistical test results for group means.

Using the selection tool from the **Tools** menu, the graphs or tables can be copied and prepared in a report for presentation. The analysis concludes that, in this typing trial, the SPEEDYTYPE word processor produced significantly higher scores than either of the other two brands.

See the chapter "One-way Layout" of the *JMP Statistics and Graphics Guide* for a complete discussion of one-way analysis of variance.
Analyzing Categorical Data

Survey data are frequently categorical data rather than measurement data. Analysis of categorical data begins by simply counting the number of responses in categories and subcategories. Counting is easy, but interpreting the relationship between categories based on counts is more complex. It requires computing probabilities and evaluating the likelihood of these probabilities compared to expectations.

For example, an American automobile manufacturer—feeling the pinch of competition from foreign auto sales—needs a market analysis before proceeding with a multimillion-dollar advertising campaign. A random sample of people is surveyed. The auto manufacturer wants to know each participant’s age, sex, marital status, and auto information. The auto information consists of the manufacturing country, the car’s size, and the car’s type, and whether it is a family, work, or sporty car. This information may provide the advertising experts with direction for the upcoming advertising campaign.

Who buys what?

Objectives

- Use the Fit Y by X command to compare two variables consisting of categorical data.
- Use the formula editor to re-code a categorical variable as a numeric variable.
- Produce and examine graphs and statistics appropriate for the comparison of proportions such as Chi-squared tests and mosaic plots.
Contents

Look Before You Leap ......................................................... 71
Open a Data Table .......................................................... 71
Address the Research Question ........................................ 71
Modify the Data Table ....................................................... 72
Contingency Table Reports .................................................. 75
Cast Variables Into Roles .................................................. 75
Contingency Table Mosaic Plots .......................................... 76
Chapter Summary .............................................................. 79
Look Before You Leap

The first step is to become familiar with the data. Begin by reviewing the data to determine the best way to proceed with the market analysis.

Open a Data Table

When you installed JMP, a folder named Sample Data was also installed. In that folder is a file named Car Poll.jmp. Open Car Poll.jmp.

The rows and columns of the Car Poll data table display as shown here.

The Car Poll data were collected from a random sample of people in a specific geographic area. The columns panel shows that Age is a numeric variable and is assigned the continuous ( ) modeling type. The other five columns are character variables with nominal ( ) modeling types.

Address the Research Question

The basic research question asks,

“Is the response probability for country of manufacture, size of car, or type of car a function of the age, sex, or marital status of the owner?”

Look at the data table to see what specific relationships lend insight into this question. The relationships between the following automobile characteristics and demographics are of interest:

- manufacturing country by age
- manufacturing country by sex
- manufacturing country by marital status
- size of car by age
- size of car by sex
- size of car by marital status
- type of car by age
- type of car by sex
- type of car by marital status
Modify the Data Table

Better summary information can sometimes be obtained from age groups rather than specific ages. In fact, dividing people into two age groups is often the basis for a valuable broad analysis. So, the median age—the age that divides the sample into two equal age groups—should be found.

The distribution of a variable and its corresponding quantiles often shows a good way to form sample groups. Use the distribution of the age column to find a reasonable value of age that divides the sample into two groups.

Select Analyze > Distribution as shown here.

When the Distribution window appears, select age as the analysis column.

A histogram is displayed with an accompanying outlier box plot, Quantiles table, and Moments table. The Quantiles table, shown here, identifies 30 as the median age.

The next step is to create a new column whose values identify whether a subject's age is greater than 30, or is less than or equal to 30.

Click the red triangle icon on the columns panel and select New Column, or select Cols > New Column.

This displays the New Column window, shown in Figure 6.1, which is used to define column characteristics.

Data Type, Data Source, and Modeling Type options define the new column's characteristics. Enter characteristics for the new column as follows:

- Type the new name (call it age group) in the Column Name text box.
- Because the new column has grouping values instead of measurements, select Character from the box beside Data Type.

The Nominal selection beside the Modeling Type box automatically shows when the data type is character.
Click the **New Property** button and select **Formula**.
You are presented with the formula editor window shown in **Figure 6.2**.

---

**Figure 6.1** New Column Window

![Figure 6.1](image1)

**Figure 6.2** Formula Editor Window

![Figure 6.2](image2)

Suppose 0 represents ages greater than the median (30) and 1 represents the ages less than or equal to the median. To create a formula that divides the sample into two groups, follow these steps:

1. Click **Conditional** in the function selector list and select the **If** function.
Highlight the expression term, denoted \( expr \).

Choose \( a \leq b \) from the Comparison functions.

Highlight the left side of the comparison clause and click \( age \) in the column selection list.

Double-click the right side of the comparison clause to obtain a text entry box.

Enter 30 for the numeric comparison.

Double-click the term denoted then clause.

Enter "1" (in double quotes because this column is a character variable).

Double-click the term denoted else clause.

Enter "0" (with double quotes).

The complete equation should look like the one shown here.
Click Apply, OK, or the formula editor's close box to fill the new column with calculated values. **Note:** Instead of using the buttons in the formula editor, you can double-click the outermost nesting box to create a single text entry box and enter \( \text{if(age<=30, "1", "0")} \). Then, press Enter (or Return) or click outside the text box, and the formula appears in formatted form.

---

### Contingency Table Reports

The nominal age grouping variable shows the relationship of age to the other nominal variables using contingency tables. To look at combinations of two variables:

- Choose **Analyze > Fit Y by X**.

JMP does the statistical analysis appropriate for a variable's modeling types and role assignments.

### Cast Variables Into Roles

Assign analysis roles to variables by choosing an analysis from the **Analyze** menu and making selections in the window that appears. In this investigation, the country, size, and type columns are dependent response (y) variables; sex, marital status, and age group are independent (x) variables. This example shows how to complete the **Fit Y by X** window (**Figure 6.3**).

- Select the three y variables (country, size, and type).
- Click the **Y, Response** button.
- Assign the x variables (sex, marital status, and age group) by selecting them and clicking the **X, Factor** button.
- Click **OK** when finished.
Contingency Table Mosaic Plots

If both x and y have either nominal or ordinal values, JMP displays a mosaic plot with accompanying text reports for each combination of columns assigned x and y modeling roles.

A mosaic chart has side-by-side divided bars for each level of its x variable. The bars are divided into segments proportional to each discrete level (value) of the y variable. The mosaic chart in Figure 6.4 shows the relationship of marital status to the manufacturing country. The width of each bar is proportional to the sample size. When the lines dividing the bars align horizontally, the response proportions are the same. When the lines are far apart, the response rates of the samples might be statistically different.

Figure 6.4 Mosaic Plot Axes

Sex and country do not appear to have any relationship at all. The proportion of automobiles from the three manufacturing countries is about the same for each sex.
The country by age group mosaic plot shows that the proportion of American car owners 30 years or over is only slightly greater than the proportion of American car owners under age 30.

The most significant relationship is seen between marital status and country. The mosaic plot, shown previously in *Figure 6.4*, and its supporting Tests table (*Figure 6.5*), suggest that married people are more likely than single people to own American cars.

There are statistical text tables accompanying each mosaic plot. To open and close these reports,

- Click the disclosure button (on Windows/Linux and on the Macintosh) next to any report title bar.

- Double-click any title bar to edit its text.

The Likelihood Ratio and Pearson Chi-squared tests evaluate the relationship between an automobile’s country of manufacture and the marital status of owner. If no relationship exists between country and marital status, a smaller Chi-squared value than the one computed in this survey would occur only seven times in 100 similar surveys.

*Figure 6.5* Table of Statistical Tests for Marital Status By Country

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>LOGLike</th>
<th>P-Square (U)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>2.57004</td>
<td>0.0000</td>
</tr>
<tr>
<td>Error</td>
<td>299</td>
<td>295.88121</td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>301</td>
<td>290.45125</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>303</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>ChiSquare</th>
<th>Prob&gt;ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>6.140</td>
<td>0.0765</td>
</tr>
<tr>
<td>Pearson</td>
<td>5.301</td>
<td>0.0700</td>
</tr>
</tbody>
</table>

These statistical results reveal that American auto manufacturers might want to direct advertising plans toward married couples.

- Scroll the report to see the relationship between size of car and each x variable (sex, marital status, and age).

The three mosaic plots indicate no relationship between car size and gender, marital status, or age group. This is seen numerically by looking at the Contingency Tables and the Tests tables beneath each of the mosaic plots (see *Figure 6.6*).

Note that by default, Col% and Row% also appear in the Contingency Tables. Right-click (Ctrl-click on the Macintosh) the table to access the Columns menu to turn columns on and off.

The Chi-squared values support the hypothesis that the purchase of large, medium, and small cars is not significantly different across the sex, marital status, and age group factor levels. The Chi-squared probabilities range from 0.06 to 0.30, so you should expect smaller Chi-squared values to occur six to 30 times in 100 similar surveys.

It probably makes no difference what size cars appear in advertisements.
The market survey categorizes cars based on both size and type, where a car's type is work, sporty, or family.

Scroll to see the plots that show the relationship between type of car and the three x variables.

The mosaic plots in Figure 6.7 and Figure 6.8 show that the type of car varies for levels of marital status and age group. As perhaps expected, many of the cars owned by married people are family automobiles, while the largest proportion of cars owned by single people are sporty cars.

Figure 6.7 Reports for Type of Car and Marital Status

So, American automobile manufacturers may choose to focus advertisements toward married couples buying family-type automobiles.

It follows logically that a relationship between age group and type of car also exists because older people are more likely to be married. Figure 6.8 shows graphically that the proportion of people over 30 years old who own family cars is much greater than those under 30. The small Chi-squared values support
the significant difference in proportions. The Chi-squared values of 0.0005 mean that proportions as varied as these are expected to occur only five times in 1,000 similar surveys.

**Figure 6.8** Reports for Type of Car and Age Group

![Contingency Analysis of type By age group](image)

### Chapter Summary

This chapter looked at relationships between categorical variables obtained from a survey. The survey recorded age, sex, marital status, and information about the type of automobile owned by a random sample of people in the same geographical area. The auto information included manufacturing country, size, and type of car. Car types were classified as work, sporty, and family. The question “Is the size of car, type of car, or manufacturing country related to the age, gender, or marital status of the owner?” was investigated.

The **Fit Y by X** command produced nine mosaic charts with supporting statistical summaries that show:

- No relationship between either sex or age and manufacturing country.
- A significant relationship between marital status and manufacturing country with married people more likely to own American cars than single people.
- No relationship between sex, age, or marital status and size of car.
- No relationship between sex and type of car.
- Significant relationships between marital status and type of car. As might be expected, married people over 30 years old were more likely to own family type cars than younger, single people.
The chapter “Contingency Tables Analysis” in the JMP Statistics and Graphics Guide discusses analyzing categorical data in more detail.

For more information about using the formula editor, see the chapter “Using the Formula Editor” in the JMP User's Guide.
Regression and Curve Fitting

This lesson demonstrates the interactive regression capabilities of JMP.

The data is from Eppright et al (1972) as reported in Eubank (1988, p. 272). The study subjects are young males. The variables in the data table are age (in months) and the ratio of weight to height. A third variable classifies the subjects into two groups based on age. The goal is to describe and model the growth pattern of subjects for the age range given in the data table.

Objectives

- Use the Fit Y by X command to fit least-squares lines to continuous data.
- Fit polynomial curves and cubic splines to the data set and explore their goodness of fit.
- Journal and save analysis results.
- Use the Group By command to fit different lines to certain groups of data.
Look Before You Leap

The first step is to become familiar with the data. Begin by reviewing the data to determine the best way to proceed with the regression.

Open a JMP File

When you installed JMP, a folder named Sample Data was also installed near the application. In that folder is a file named Growth.jmp. Open Growth.jmp.

A partial listing of the Growth.jmp data table is shown here. There are 2 columns and 72 rows. The ratio column contains the average weight-to-height ratio for each age group in the study. The age groups range from 0.5 to 71.5 months.

The modeling type for each column is shown to the left of the variable name in the columns panel. Both columns have a continuous modeling type ( ), as needed for a regression analysis in JMP.

The purpose of the analysis is to determine if the ratio values are related to (or dependent on) the age values.

Select an Analysis

To fit regression curves:

Select the Analyze > Fit Y by X.

Note: The Fit Y by X analysis does four kinds of analyses, depending on the modeling type of the variable:

- Regression analysis when both x and y have continuous values, as in this example.
- Categorical analysis when both x and y have nominal or ordinal values.
- Analysis of variance when x is nominal and y has continuous values.
- Logistic regression when x is continuous and y has nominal or ordinal values.

Choose Variable Roles

The Fit Y by X command first displays the Fit Y by X window. Y identifies a response or dependent variable and x identifies a classification or independent variable. To choose variable roles:

Highlight ratio and click Y, Response.

Highlight age and click X, Factor, as shown in Figure 7.1.

Click OK.
Now investigate if the ratio of weight to height is a function of age.

**Fitting Models to Continuous Data**

The scatterplot shown in Figure 7.2 is the result of the Fit Y by X analysis. It is easy to see that the growth pattern is not random. A straight line regression is a good baseline fit to compare with other regression curves.

**Figure 7.2** Scatterplot of ratio by age
When clicked, the red triangle icon on the scatterplot title bar reveals a variety of fitting commands and additional display options. Options include a Show Points toggle, fitting commands, and other features. The Show Points command alternately hides or displays the points in the plot. Fitting options can be as simple as fitting a straight line or involved as drawing density ellipses. Fitting options can be used repeatedly to overlay different fits on the same scatterplot.

Begin with a simple line and try different techniques after inspecting the initial straight line regression fit.

The Fit Mean Command

- Click the red triangle icon and select Fit Mean.

This is the baseline fit that hypothesizes that there is no relationship between $x$ and $y$. All other fits compare to this fit. Since the Fit Mean table is closed by default:

- Click the disclosure button ( £ on Windows/Linux and ^v on the Macintosh).

This displays a report that shows:

- The sample mean (arithmetic average) of the response variable.
- The standard deviation of the response variable.
- The standard deviation of the response mean.
- The error sum of squares for the simple mean model.

The Fit Line Command

To fit a simple regression line through the data points,

- Click the red triangle icon and select Fit Line.

The regression line minimizes the sum of squared distances from each point to the line of fit. Because of this property, it is sometimes referred to as the line of best fit.
Each time a fit is selected from the red triangle icon, the regression equation and another red triangle icon for that fit show beneath the scatterplot, as shown here.

Click the red triangle icon to reveal commands that show confidence curves and give the ability to save predicted and residual values as new data table columns. The **Save Predicteds** command saves the prediction equation for the fit with the new column of predicted values. The fit can be removed from the scatterplot at any time with the **Remove Fit** command.

### The Summary of Fit Table

Clicking the red triangle icon and selecting **Fit Line** produces a Summary of Fit table, which summarizes the linear fit.

- **Rsquare (R^2)** quantifies the proportion of total variation in the growth ratios accounted for by fitting the regression line.
- **Rsquare Adj** adjusts R^2 to make it more comparable over models with different numbers of parameters.
- **Root Mean Square Error (RMSE)** is a measure of the variation in the ratio values that is attributable to different people rather than to different ages.
• **Mean of Response** is the arithmetic mean (average) of the ratio values.
• **Observations** is the total number of nonmissing values.

**Note:** Also notice that the first line of the report is the regression equation, which is editable.

### The Analysis of Variance Table

In addition to producing a Line of Fit table, clicking the red triangle icon and selecting **Fit Line** produces an Analysis of Variance table.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>0.6665172</td>
<td>0.666517</td>
<td>334.4435</td>
</tr>
<tr>
<td>Error</td>
<td>70</td>
<td>0.1667645</td>
<td>0.00002338</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>71</td>
<td>1.0032735</td>
<td>x0.0001</td>
<td></td>
</tr>
</tbody>
</table>

The elements of the table give an indication of how well the straight line fits the data points:

- **Source** identifies the sources of variation in the growth ratio values (**Model**, **Error**, and **C. Total**).
- **DF** records the associated degrees of freedom for each source of variation.
- **Sum of Squares** (SS for short) quantifies the variation associated with each variation source. The **C. Total SS** is the corrected total SS computed from all the ratio values. It divides (partitions) into the SS for **Model** and SS for **Error**. The **Model SS** is the amount of the total variation in the ratio scores explained by fitting a straight line to the data. The **Error SS** is the remaining or unexplained variation.
- **Mean Square** lists the Sum of Squares divided by its associated degrees of freedom (DF) for Model and Error.
- **F Ratio** is the regression (Model) mean square divided by the Error mean square.
- **Prob > F** is the probability of a greater F-value occurring if the ratio values differed only because of different subjects rather than because the subjects are different ages.

In this example, the significance of the F-value is 0.0001, which strongly indicates that the linear fit to the weight/height growth pattern is significantly better than the horizontal line that fits the sample mean to the data.

### The Parameter Estimates Table

In addition to producing a Line of Fit table and an Analysis of Variance table, clicking the red triangle icon and selecting **Fit Line** produces a Parameter Estimates table.

| Term | Estimate | Std Error | T Ratio | Prob > |t| |
|------|----------|-----------|---------|---------|---|
| Intercept | 0.666517 | 0.012587 | 53.007  | <0.0001 |
| age | 0.00002338 | 0.0000392 | 10.61   | <0.0001 |

- **Term** lists the parameter terms in the regression model.
- **Estimate** lists estimates of the coefficients in the regression line equation.
- **Std Error** lists estimates of the standard error of the parameters.
7 Regression and Curve Fitting—Fitting Models to Continuous Data

- **t Ratio** is the parameter estimate divided by its standard error.
- **Prob > |t|** is the probability of a greater absolute t-value occurring by chance alone if the parameter has no effect in the model.

The significant F-ratio in the Analysis of Variance table tells the student that the regression line fits significantly better than the horizontal line at the mean (the simple mean model). However, while the regression line looks like a good fit for age groups above seven months, it does not describe the data well for ages younger than seven months.

### The Exclude Command

Because the low-age points are the trouble spots for the linear fit, remove them from the analysis and try fitting the model to the remaining values.

To highlight these outliers and exclude them from the analysis:

- Select the lasso tool from **Tools** menu or toolbar.
- Drag the lasso around the points to be excluded.
- Select **Rows > Exclude/Unexclude** to exclude the selected points.
- Select **Rows > Markers** to assign the X marker to the excluded points.
- Click the red triangle icon in the title bar and choose the **Fit Line** command again to see the results of excluding the low-age points.

The scatterplot shown here has both regression lines. The low-age points still show on the plot but are not included in the second regression line's computation.

### Journaling JMP Results

After completing this part of the exploratory regression analysis:
Choose **Edit > Journal**.

The first time the **Journal** command is selected during a **JMP** session, a journal window opens and fills with the graphs and tables from the report window.

The open journal file contains all reports from the active report window. Plots can be resized, opened, or closed, as can outlines. This allows for printing of certain parts of the report.

However, the journal file is a like a word processing document and is not linked to the **JMP** session. The journal file can be saved, but when reopened, a journal file does not re-create its associated **JMP** session.

Choose **Save As** from the **File** menu to save the journal.

The window similar to the one shown here prompts for a file name, and appends .jrn to the file name to identify the file type.

Leaving a journal file open causes each subsequent use of the **Journal** command to append results in the active window at the end of the journal contents.

Name the journal **Regression Results**.

On Windows, change the **Save as type** to **RTF Files (*.RTF)** and click **Save**.

On the **Macintosh**, change the **Format** type to **RTF document** and click **Save**.

On **Linux**, change the **Save as type** option to **.rtf (Rich Text File)** and click **Finished**.

Navigate to the file’s directory on your system and open the file.

The file should open in your default word processor as shown here. Note that the graphics are saved as graphics, and the reports are saved as text tables.
The Fit Polynomial Command

Now, examine a polynomial fit for comparison to the linear fit. A linear regression is simply a polynomial of degree 1.

- Click the Bivariate Fit of ratio By age report window to assure it is active.
- Highlight all the excluded rows to include all rows in the analysis.
- Choose Rows > Exclude/Include include these points in the next analysis.
- Keeping the rows highlighted, select Rows > Markers and return the marker shape to the default.
- Click the red triangle icon in the title bar, select Fit Polynomial > 2, quadratic, which allows the fit to have curvature.
- Remove the line of fit that excluded the lower age groups by clicking the second (modified) regression line's red triangle icon and selecting Remove Fit.
- Remove the Fit Mean results so that only the polynomial fit and the line fit to all the data points remain.
Click the red triangle icon and select **Fit Polynomial > 3, cubic** to overlay a polynomial curve of degree 3 on the scatterplot.

Again select the **Edit > Journal** command to append these results to the existing journal.

**Figure 7.3** Comparison of Linear Fit and Polynomial Fits of Degree 2 and 3

The tables in **Figure 7.3** show the $R^2$ value from the Summary of Fit tables for the linear fit, the second degree polynomial fit, and the third degree fit. As polynomial terms are added to the model, the regression curve appears to fit the data better.

### The Fit Spline Command

Even the polynomial fit of degree 3 does not quite reach the outlying points of the very young subjects. A free-form function that acts as if it smooths the data, such as a smoothing spline, may be better.

Use the **Remove Fit** command on both polynomial fits, so that only the first linear regression line shows on the scatterplot.

Click the red triangle icon on the title bar and select **Fit Spline** three times, with lambda values of 10, 1,000, and 100,000.

Lambda is a tuning factor that determines the flexibility of the spline. The **Fit Spline** command submenu (shown to the left in **Figure 7.4**) lists lambda values. The three new fits are overlaid on the scatterplot.
Regression and Curve Fitting—Fitting Models to Continuous Data

Figure 7.4 Comparison of Spline Fits

By inspecting the plot, see that the lambda = 10 curve is too flexible and therefore local error has too great an effect on it. The lambda = 100,000 curve is too stiff. It is so straight that it does not reach down to model the lower ages closely. However, the lambda of 1,000 curve fits well. Its shape is not influenced by local errors, and it appears to fit the data smoothly.

If a report of these results is needed, journal these results.

Select Edit > Journal.

The Journal command appends the scatterplot with spline fits and text reports to the open journal file. After journaling the final analyses, the following draft notes about the spline-fitting technique can be added at the bottom of the journal window.

Select the annotate tool from the tool bar.

Click and drag a large box at the bottom of the report.

Add the following text to the box.

"This fitting technique applies a cubic polynomial to the interval between points; the polynomial is joined such that the curve meets at the same point with the same slope to form a continuous and smooth curve. A small enough lambda could make such a curve go through every point, which would model the error, not the mean. A moderate lambda value forces the curve to be smoother, i.e., less curved. This is accomplished by adding a curvature penalty to the optimization that minimizes the sum of squares error."

By comparing various regression fits, notice that both the polynomial fits and the spline fit with moderate flexibility best describe the data. These models show that infants grow most rapidly during the first months of life and that growth rate decreases significantly at approximately 12 months.
Fitting By Groups

“The Exclude Command,” p. 88 in this chapter, shows how to overlay a linear fit for the whole sample with a linear fit for children over the age of one year. Carry this idea one step further with overlay fits to compare children under the age of one year with children over one year.

1. In the Growth.jmp data table, create a new column called group to act as a grouping variable.
2. Right-click (Command-click on the Macintosh) and select Formula from the menu that appears.
3. Enter the formula shown in Figure 7.5.

This assigns the value Babies to each child less than 12.5 months old, and Toddlers to children who are 12.5 months or older.

**Figure 7.5** Computed Age Grouping Variable

- Click the Bivariate report to make it the active window.
- Clear the Smoothing Spline fits still showing, such as those seen in Figure 7.4, using each fit’s Remove Fit command.
- Click the red triangle icon and select Group By to display the window shown here.
Select group, the newly created grouping variable, and click OK.
Choose the Fit Line command.

With a grouping variable (group) in effect, the overlaid regression lines shown in Figure 7.6 appear automatically. The points that correspond to each regression give a dramatic visualization of the steep growth rate for babies during the first year of life compared to the more moderate growth rate of toddlers and small children age one to five years.

Figure 7.6 Regression Lines for Levels of a Grouping Variable

Chapter Summary
To analyze some bivariate data, the Fit Y by X command was used to examine a variety of regression model fits. The task was to model and describe the growth pattern of subjects over a range of ages. You measured growth using the ratio of weight to height and accomplished this task by:
- Fitting mean to use as a baseline comparison to other regression models and evaluate the fit using statistical text reports.
- Fitting a straight line as a first guess for a model.
- Excluding outliers and again fitting a straight line to compare the $R^2$ values given by the Summary of Fit tables for both lines.
- Fitting second and third degree polynomials to see if they model the growth pattern more realistically.
- Fitting smoothing splines with lambda values of 10, 1,000, and 100,000 and comparing them with each other and with the linear fit.
- Clicking the red triangle icon and selecting the grouping facility (Group By) to compare growth rates of babies under the age of one year with toddlers from age one to five years.
- Using the Journal command to append each of these regression reports and graphs to a journal file.
A Factorial Analysis

This lesson examines two treatments of popcorn. The plain, everyday type has been around for years, but researchers claim to have discovered a special treatment of corn kernels. This new process supposedly increases the popcorn yield as measured by popcorn volume from a given measure of kernels.

Is this true? If so, how much is the increase? Are these increases the same for all groups of conditions? The special treatment raises the cost of the popcorn, so the increase in yield must be significant enough to warrant the higher costs.

The popcorn data used in this chapter and for examples in the JMP User's Guide and the JMP Statistics and Graphics Guide are artificial, but the experiment was inspired by experimental data reported in Box, Hunter, and Hunter (1978).

Objectives

- Learn techniques to analyze a designed factorial experiment using the Fit Model command.
- Evaluate and interpret effects using JMP's interactive graphical techniques.
- Examine supporting text reports.
- Evaluate the significance of interaction effects using interaction plots.
- Save a model's predicted values for each observation.
8

Contents

Look Before You Leap ................................................................. 97
Open a Data Table ................................................................. 97
What Questions Can Be Asked? .................................................. 97
The Fit Model Window ............................................................ 98
Graphical Display: Leverage Plots ............................................. 99
Quantify Results: Statistical Reports ....................................... 101
Analysis of Variance ............................................................... 101
Summary Reports For The Whole Model .................................. 103
Summary Reports for Effects .................................................... 103
Chapter Summary ................................................................. 106
Look Before You Leap

The popcorn yield data are the result of a designed experiment. The same amounts of different kinds of corn were methodically popped under different conditions. First, look at the data to review the results of the popcorn experiment.

Open a Data Table

When you installed JMP, a folder named Sample Data was also installed near the application. In that folder is a file named Popcorn.jmp. Open Popcorn.jmp.

The Popcorn data table displays in spreadsheet form as shown here.

For the experiment, the corn was popped under controlled conditions. Plain popcorn and specially-treated gourmet popcorn were each popped in large or small amounts of oil and in large or small batches. Two trials were done for both kinds of corn under all popping conditions.

This experimental design is called a factorial design. The experiment has three factors, usually called main effects, which are:

- Kind of popcorn (plain or gourmet)
- Amount of cooking oil (little or lots)
- Cooking batch size (large or small)

What Questions Can Be Asked?

The appropriate statistical analysis for a factorial design addresses the following questions about the main effects:

- Is there an overall difference in yield between plain and gourmet popcorn?
- Is there an overall difference in yield between cooking in lots of oil instead of a small amount of oil?
- What is the difference in yield between cooking several small batches instead of one large batch?

Analysis of a factorial experiment also provides information about the interaction between the main effects as addressed by the following questions:

- Does the amount of cooking oil have the same effect on both types of popcorn? In other words, is there an interaction effect between popcorn type and amount of cooking oil used?
- Is there an interaction effect between batch size and type of popcorn?
- Is there an interaction effect between batch size and amount of oil used?
- Are there interaction effects among the three main effects?
The Fit Model Window

- Select Analyze > Fit Model.

The Fit Model command lets you specify and analyze complex models like the factorial design in this experiment.

The Fit Model command displays the Fit Model window shown in Figure 8.1. This window is used to define the type of model, the model response variable, and model effects.

To specify the factorial model:
- Select yield from the Select Columns list.
- Click the Y button.
- Select popcorn, oil amt, and batch from the Select Columns list.
- Click the Macros button and select Full Factorial. This adds all main effects and interactions (crossed effects) to the Construct Model Effects list (Figure 8.1).
- Further tailor the model by adding effects or removing unwanted effects with the Add and Remove buttons. In this case, remove the three-way interaction term by clicking popcorn*oil amt*batch and selecting Remove.

Figure 8.1 Fit Model Window

- Select Effect Leverage from the box beside Emphasis at the top right of the Fit Model window.
- Click Run Model to estimate the model parameters and view the results.
Graphical Display: Leverage Plots

The **Fit Model** command graphically displays the whole model and each model effect as the leverage plots shown in **Figure 8.2** through **Figure 8.5**. It is possible to tell at a glance whether the factorial model explains the popcorn data and which factors are most influential.

The whole model plot to the left in **Figure 8.2** shows actual yield by predicted yield values with a regression line and 95% confidence curves. The regression line and the 95% confidence curves cross the sample mean (the horizontal line), which show that the whole factorial model (all effects together) explains a significant proportion of the variation in popcorn yield.

There is also a significant difference in yield between the two types of popcorn, as shown in the right-hand leverage plot for the popcorn main effect. The small p-values beneath the plots quantify the significant model fit and popcorn effect.

**Figure 8.2** Leverage Plots of Actual by Predicted and of Popcorn Effect

In **Figure 8.3**, the confidence curves for oil amt and the popcorn*oil amt interaction do not cross the horizontal mean line (rather, they encompass the mean line). This shows that neither of these factors significantly affected popcorn yield.

**Figure 8.3** Leverage Plots for the Oil Amt and Its Interaction with Popcorn
The leverage plots in Figure 8.4 show that the batch size effect (batch) and the interaction between popcorn type and batch size (popcorn*batch) are significant effects. This means that the size of the batch makes a difference in the popcorn yield. Furthermore, the significant interaction means that batch size affects each type of popcorn differently.

Figure 8.4  Leverage Plots for Batch and Its Interaction with Popcorn

The two leverage plots shown in Figure 8.5 show that there is no significant interaction between amount of oil and batch size.

Figure 8.5  Leverage Plots for Other Interaction Effects

For more information about interpretation of leverage plots, see the chapters “Understanding JMP Analyses” and “Standard Least Squares: Introduction” and the appendix “Statistical Details” of the JMP Statistics and Graphics Guide.
Quantify Results: Statistical Reports

Because oil amt and its interactions with other effects are not significant, fit the popcorn data again without these effects. The new model should have the significant factors (type of popcorn, batch size), and their interaction term. This approach condenses the statistical reports that show estimates of yield under the different conditions of interest. Use the same Fit Model command as before.

- Click the Fit Model window to make it the active window.
- If the window is closed, click the red triangle icon on the report and select Script > Redo Analysis to open a new Fit Model window.

Use the following method to specify the two-factor model:

- From the full factorial model, select unwanted effects listed in the Construct Model Effects box and click Remove.
- Click Run Model.

Analysis of Variance

The whole model leverage plot in Figure 8.6 shows that the two-factor model describes the popcorn experiment well. Examine the tables that accompany the whole model leverage plot.

The Analysis of Variance table that accompanies the whole model leverage plot quantifies the analysis results. It lists the partitioning of the total variation of the sample into components. The ratio of the Mean Square components forms an F-statistic that evaluates the effectiveness of the model fit. If the probability associated with the F-ratio is small, then the analysis of variance model fits better statistically than the simple model that contains only the overall response mean.
The Analysis of Variance table shows these quantities:

- **Source** identifies the sources of variation in the popcorn yield values (Model, Error, and C. Total).
- **DF** records the degrees of freedom for each source of variation.
- **Sum of Squares** (SS for short) quantifies the variation in yield. C. Total is the corrected total SS. It is divided (partitioned) into the SS for Model and SS for Error. The SS for Model is the variation in the yield explained by the analysis of variance model, which hypothesizes that the model factors have a significant effect. The SS for Error is the remaining or unexplained variation.
- **Mean Square** is a sum of squares divided by its associated degrees of freedom (DF).
- **F Ratio** is the model mean square divided by the error mean square.
- **Prob > F** is the probability of a greater F-value occurring if the variation in popcorn yield resulted from chance alone rather than from the model effects.

In this example, the F-value is 0.0001. This implies that the difference found in the popcorn yield produced by this experiment is expected only 1 time in 10,000 similar trials if the model factors do not affect the popcorn yield.
Summary Reports For The Whole Model

Other tables in the Fit Model report provide statistical summaries. The Summary of Fit table shows the numeric summaries of the response for the factorial model:

- **Rsquare (R²)** of 0.809 tells the scientist that the two-factor model explains nearly 81% of the variation in the data.
- **Rsquare Adj** adjusts R² to make it more comparable over models with different numbers of parameters.
- **Root Mean Square Error** (sometimes called the RMSE) is a measure of the variation in the yield scores that can be attributed to random error rather than differences in the model’s factors.
- **Mean of Response** is the mean (average) of the yield scores.
- **Observations** is the total number of recorded scores.

The F-test probabilities in the Effect Test table tell the scientist that all model effects explain a significant proportion of the total variation. There is also a table that gives the parameter estimates for the model.

Summary Reports for Effects

Now look at the summary tables for each effect in the model. The tables for the main effects are shown here. The Least Squares Means table lists the least squares means and standard errors for each level of the model factors, without considering the interaction between them. In this balanced example, the least squares means are simply the sample means of each factor level.

The nature of the interaction is important in the interpretation of the popcorn experiment. To examine the significant popcorn*batch interaction,

- Click the red triangle icon from the popcorn*batch title bar and select LSMeans Plot, or click the red triangle icon from the Response yield title bar and select Factor Profiling > Interaction Plots.

This command plots the least squares means for each combination of effect levels, as shown in Figure 8.7.
The Least Squares Means table for the popcorn*batch effect tells the whole story. Batch size makes no difference for the plain brand popcorn, but popping in small batches increases the yield in the new gourmet brand.

Because the factorial model with two-factors is a good prediction model, save the prediction formula.

Click the red triangle icon in the Response yield title bar and select Save Columns > Prediction Formula, as shown in Figure 8.8.
This command creates a new column in the Popcorn data table called Pred Formula yield that contains the predicted values for each experimental condition.

The prediction formula, shown at the bottom of Figure 8.8, becomes part of the column information. To see this formula:

- Highlight the new column name (Pred Formula yield).
- Select Formula from theCols menu.

The prediction formula can be copied to the clipboard using standard cut and paste techniques.

Results show that popcorn should be packaged:

- in small packages so that the yield will be good.
- in family size packages with smaller packets inside.
- in family size packages with popping instructions that clearly state the best batch size for good results.
Chapter Summary

In this chapter, a designed experiment evaluated the difference in yield between two types of popcorn. A three-factor factorial experimental design was the basis for popcorn popping trials. The results were analyzed by using the Fit Model command. The following results were found:

- The leverage plots for the factorial analysis of three factors showed one main effect and its associated interactions to be insignificant.
- A more compact two-factor analysis with interaction adequately described the variation in yield for the popcorn trials.
- The interaction between the two main effects was significant. The Least Squares Means table for the interaction showed how the two types of popcorn behaved under different popping conditions.

The new, more expensive gourmet popcorn had better yield than the plain everyday type only if popped in small batches.
Exploring Data
Advanced Example with Principal Components

Exploration is the search to find something new—the endeavor to make some discovery. For data analysis, exploratory study is often the most fruitful part of the analytical process because it is the most open to serendipity. Something noticed in a data set can be the seed of an important advance.

There are two important aspects of exploration:

• What is the pattern or shape of the data?
• Are there points unusually far away from the bulk of the data (outliers)?

When exploring data composed of many variables, the great challenge is dealing with this high dimensionality. There can be many variables that have interesting relationships, but it’s hard to visualize the relationship of more than a few variables at a time.

Objectives

• Use graphical techniques to search for outliers in one, two, three, and higher dimensions.
• Perform a principal components analysis and examine it graphically.
• Examine outliers graphically using Mahalanobis distance.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solubility Data</td>
<td>109</td>
</tr>
<tr>
<td>One-Dimensional Views</td>
<td>109</td>
</tr>
<tr>
<td>Two-Dimensional Views</td>
<td>110</td>
</tr>
<tr>
<td>Three-Dimensional Views</td>
<td>112</td>
</tr>
<tr>
<td>Principal Components and Biplots</td>
<td>112</td>
</tr>
<tr>
<td>Using Colors, Markers, and the Brush Tool</td>
<td>116</td>
</tr>
<tr>
<td>Multivariate Distance</td>
<td>117</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>118</td>
</tr>
</tbody>
</table>
Exploring Data—Solubility Data

This lesson examines compounds for those with unusual solubility patterns in various solvents. When you installed JMP, a folder named Sample Data was also installed near the application. In that folder is a file named Solubility.jmp. Data from an experiment by Koehler, Grigorus, and Dunn (1988) are in the Solubility.jmp file.

Open Solubility.jmp.

There are 72 compounds tested with six solvents, in columns called 1-Octanol, Ether, Chloroform, Benzene, Carbon Tetrachloride, and Hexane.

The Labels column in the table should serve as a label variable (Figure 9.1) so when you plot them, the compound names instead of row numbers identify points. Although this is already done for you in Solubility.jmp, you should know that to assign the label role to columns, you should select the columns, then select Cols > Label/Unlabel or click the red triangle in the columns panel and select Label/Unlabel from the resulting menu.

Figure 9.1  Solubility Data Table

There are six solvent variables, but there are no six-dimensional graphics. However, it is possible to look at six one-dimensional graphs, 15 two-dimensional graphs, and 20 three-dimensional spinning plots. Using principal components, a representation of higher dimensions can be displayed.

One-Dimensional Views

The Distribution command helps you summarize data one column at a time. It does not show any relationships between variables, but the shape of the individual distributions helps identify the one-dimensional outliers.

To begin exploring the solubility data:

Choose Analyze > Distribution.
Select the six solubility columns and click the Y, Columns button.

Click OK.

Their histograms, resized and trimmed of other output, are shown in Figure 9.2.

Click any histogram bar.

That bar, and all other representations of that data, are highlighted in all related windows. To see how outlying values are distributed in the other histograms:

Shift-click the outlying bars in each histogram.

This identifies the outlying rows in each single dimension.

Use the Rows > Markers palette to assign the X marker to these selected rows.

The markers show in the data table and in subsequent plots.

**Figure 9.2 One-Dimensional Views**

To create a new data table that contains only the outlying rows:

Use the Tables > Subset command as shown here.

Click OK to accept the default settings.

Scroll through the new subset table to see the compound names of the one-dimensional outliers.

**Two-Dimensional Views**

Return to Solubility.jmp.
Select Analyze > Multivariate Methods > Multivariate.

Highlight all the continuous columns in the table and click the Y, Columns button.

Click OK.

This displays a correlation matrix and a scatterplot matrix of all 30 two-dimensional scatterplots (Figure 9.3).

The one-dimensional outliers appear as Xs in each scatterplot. Note in the scatterplot matrix that many of the variables appear to be correlated, as evidenced by the diagonal flattening of the normal bivariate density ellipses. There appear to be two groups of variables that correlate among themselves but are not very correlated with variables in the other group.

Figure 9.3 Two-Dimensional View

The variables Ether and 1-Octanol appear to make up one group, and the other group consists of the remaining four variables. These two groups are outlined on the scatterplot matrix shown in Figure 9.3.

Scan these plots looking for outliers (points that fall outside the bivariate ellipses) of a two-dimensional nature and identify them with square markers using the following steps.

- Shift-click each outlier.

- Select Rows > Markers and select the square marker from the palette.

Now, both one- and two-dimensional outliers are identified.
Three-Dimensional Views

To see points in three dimensions:
- Select **Graph > Spinning Plot**, which opens the three-dimensional Spinning Plot window.
- Add all six continuous variables to the **Y, Columns** list.
- Click **OK**.
- After the plot appears, click and drag the X, Y, and Z axis tags to any combination of three variables. The goal is to look for points away from the point cloud for each combination of three variables. To aid in this search:
  - Rotate and examine each three-dimensional plot with the spin controls.
  - Select the hand tool from the tools palette.
  - Click and drag inside the plot to rotate the plot in any direction.
  - Select the arrow tool from the tools palette when the plot is at an angle to see outliers.

**Figure 9.4** shows two three-dimensional outlying points in the view of Ether by 1-Octanol by Benzene that hadn’t been apparent before. To label them:
- Shift-click these points.
- Select **Rows > Label/Unlabel**.

Their labels, METHYLACETATE and ACETONE, appear on the plot.

**Figure 9.4** Spotting Outliers in a Three-Dimensional View

---

Principal Components and Biplots

Because many of the variables in the Solubility.jmp table are highly correlated, there is not a lot of scatter in six dimensions. The scatter is oriented in some directions but is flattened in other directions.
To illustrate this:
- Remove the labels from METHYLACETATE and ACETONE.
- Select Graph > Spinning Plot.
- Add only two highly correlated variables, in this case Ether and 1-Octanol.
- Click OK.
- Click the red triangle icon in the Spinning Plot title bar and select Principal Components.

The results are shown in Figure 9.5. Note that because the data are highly correlated, the scatter in the points runs in a narrow ellipse whose principal axis is oriented in the direction marked P1.

**Figure 9.5 Two Correlated Variables with Principal Components**

To see the greatest variation of the data in one dimension,
- Rotate the axis so that the first principal component, P1, is horizontal.

The technique of extracting these orientations that capture the highest variance is called principal components analysis. Principal components capture the most variation possible in the smallest number of dimensions.

Use principal components to explore all six dimensions with the following steps:
- Select Graph > Spinning Plot.
- Add all six continuous variables to the Y, Columns list.
- Click OK.
- Click the red triangle icon and select Principal Components.

The result is the Principal Components table in Figure 9.6. The cumulative percent row (CumPercent) shows that the first three principal components account for 97.8% of the six-dimensional variation.
Exploring Data—Solubility Data

**Figure 9.6** Principal Components Text Report

<table>
<thead>
<tr>
<th>Principal Components</th>
<th>Eigenvalue</th>
<th>0.7903</th>
<th>0.9402</th>
<th>0.1339</th>
<th>0.0611</th>
<th>0.0471</th>
<th>0.0217</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>79.7500</td>
<td>57.7501</td>
<td>2.2309</td>
<td>1.6102</td>
<td>0.7952</td>
<td>0.3020</td>
<td></td>
</tr>
<tr>
<td>Cum Percent</td>
<td>79.7500</td>
<td>95.9540</td>
<td>97.3848</td>
<td>98.3539</td>
<td>99.3930</td>
<td>100.0000</td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td>0.37441</td>
<td>0.65981</td>
<td>-0.19377</td>
<td>-0.66842</td>
<td>0.21608</td>
<td>0.18674</td>
<td></td>
</tr>
<tr>
<td>Ether</td>
<td>0.34834</td>
<td>0.63434</td>
<td>0.11873</td>
<td>0.62084</td>
<td>-0.20850</td>
<td>0.11495</td>
<td></td>
</tr>
<tr>
<td>Chloroform</td>
<td>0.49490</td>
<td>-0.20804</td>
<td>-0.64850</td>
<td>0.30590</td>
<td>0.43608</td>
<td>0.18763</td>
<td></td>
</tr>
<tr>
<td>Benzene</td>
<td>0.44911</td>
<td>-0.64715</td>
<td>-0.21994</td>
<td>-0.09495</td>
<td>-0.49649</td>
<td>0.08085</td>
<td></td>
</tr>
<tr>
<td>Carbon Tetrachloride</td>
<td>0.43152</td>
<td>-0.37936</td>
<td>0.16497</td>
<td>0.24735</td>
<td>-0.49649</td>
<td>0.04918</td>
<td></td>
</tr>
<tr>
<td>Hexane</td>
<td>0.42217</td>
<td>-0.37117</td>
<td>0.06828</td>
<td>0.18235</td>
<td>0.40526</td>
<td>0.23436</td>
<td></td>
</tr>
</tbody>
</table>

On the spinning plot, principal components appear as additional rays labeled P1, P2, and so on, as shown in **Figure 9.7**. These rays are the projections of the six-dimensional direction of the principal component in the three dimensions shown. P1 tends to be the longest ray because it shows the direction of greatest variance. P2 appears as the second longest ray and describes the next most prominent direction. The other rays are labeled successively as their variances decrease.

**Figure 9.7** Principal Component Rays in Three of the Variables

A spinning plot of the first three principal components produces the best three-dimensional representation of the six-dimensional space.

Drag the X, Y, and Z axes icons to the spaces labeled Prin Comp 1, Prin Comp 2, and Prin Comp 3, respectively.

In this space, CAFFEINE appears as an outlier (see **Figure 9.8**).

In the principal component space, the principal components are the axes and the variables show as rays. A plot showing both variable rays and points in an approximation of a high-dimensional space is called a biplot (Gabriel 1971). The configuration of the variable rays in **Figure 9.8**, shows how the variables relate. Note that eth and oct seem to dominate one direction, while the other four variables cluster to define the other directions. As more dimensions (variables) are condensed into principal components, the angles between variables become indicators of their correlation. In a factor analysis, these directions are further refined, mapping the variables into clusters called factors.
When looking for outliers, it is often revealing to examine the space of the last three principal components instead of the first three. The last principal components define the least popular directions of the scatter (directions with the least variation). If a point is unusual in a multivariate sense, then its prominence in the least popular direction suggests it is an outlier. To use this strategy,

- Drag the X, Y, and Z axes icons to the spaces labeled Prin Comp 4, Prin Comp 5, and Prin Comp 6, respectively.
- Rotate the plot to find an outlier.
- Click the outlier and select Rows > Label/Unlabel.

Your spinning plot should show SULFATHIAZOLE as the most unusual value. Most other points should be in a tight cluster near the center.

**Figure 9.9 Outlier in the Last Principal Components**

Another strategy is to plot each variable with the first two principal components and then with the last two. As an example,
Exploring Data—Using Colors, Markers, and the Brush Tool

Drag the X, Y, and Z axes icons to the spaces labeled Prin Comp 1, Prin Comp 2, and hex, respectively.
You will see that hexane is on the z-axis with the first principal components plotted on the x- and y-axes.

**Figure 9.10** Hexane and the First Two Principal Components

Drag the X, Y, and Z axes icons to the spaces labeled Prin Comp 5, Prin Comp 6, and CCl₄, respectively.
The variable Carbon Tetrachloride should be plotted with the last two principal components.

**Figure 9.11** Carbon Tetrachloride and Last Two Principal Components

**Using Colors, Markers, and the Brush Tool**

Though a given plot shows two or three dimensions geometrically, aspects of other dimensions can be added by plotting the points with different markers and colors. In addition, using the brush tool ( ) or the lasso tool ( ) shows where points in areas of one plot show up on other plots.
Multivariate Distance

The basic concept of distance in several dimensions relates to the correlation of the variables. For example, in a multivariate scatterplot cell for Benzene by Chloroform (Figure 9.3), HYDROQUINONE is located away from the point cloud. This compound is not particularly unusual in either the $x$ or $y$ direction alone, but it is a two-dimensional outlier because of its unusual distance from the strong linear relationship between the two variables. The ellipse is a 95% density contour for a bivariate normal distribution with the means, standard deviations, and correlation estimated from the data. The concept of distance that takes into account the multivariate normal density contours is called Mahalanobis distance.

Though only three dimensions can be visualized at a time, the Mahalanobis distance can be calculated for any number of dimensions. To produce a plot of the Mahalanobis distance:

- Select Outlier Analysis from the menu accessed by the red triangle at the top of the multivariate report.

Figure 9.12 shows the Mahalanobis distance by the row number for each data point. To label these points:

- Select the brush tool ( ) from the tools palette.

- While holding down the Shift key, drag the brush over the points labeled in Figure 9.12. These are the four points with the greatest Mahalanobis distances.

- Select Rows > Label/Unlabel.
Chapter Summary

In this example, commands from the Analyze and Graph menus were used for data exploration to locate and identify unusual points. The data were first examined in one dimension using the Distribution command and then in two dimensions using the Multivariate command to look for unusual points in histograms and scatterplots.

Next, the Spinning Plot command in the Graph menu was used to plot three columns at a time. The technique of principal components was used to summarize six dimensions and to plot principal component rays. The Principal Components table showed that the first three principal components accounted for more than 97% of the total variation. To locate multivariate outliers, each column was plotted with the first two principal components and then with the last two principal components.

Finally, the Outlier Analysis command in the Multivariate report produced the Mahalanobis outlier distance plot, which summarizes the points in six dimensions. The multivariate outliers were highlighted and labeled in this multi-dimensional space.

Multiple regression is the technique of fitting or predicting a response by a linear combination of several regressor variables. The fitting principle is like simple linear regression, but the space of the fit is in three or more dimensions, making it more difficult to visualize. With multiple regressors, there are more opportunities to model the data well, but the process is more complicated.

This chapter begins with an example of a two-regressor fit that includes three dimensional graphics for visualization. The example is then extended to include six regressors (but unfortunately no seven-dimensional graphics to go with it).

Objectives

- Illustrate the concept of a fitting plane using graphical techniques.
- Combine data tables using the **Concatenate** command.
- Explore a three-dimensional version of a leverage plot.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerobic Fitness Data</td>
<td>121</td>
</tr>
<tr>
<td>Fitting Plane</td>
<td>123</td>
</tr>
<tr>
<td>Compute a Grid for Plotting</td>
<td>124</td>
</tr>
<tr>
<td>Combine the Data Tables</td>
<td>125</td>
</tr>
<tr>
<td>Fit Planes to Test Effects</td>
<td>127</td>
</tr>
<tr>
<td>Leverage Plots: The Hypothesis-Eyed View</td>
<td>130</td>
</tr>
<tr>
<td>Whole Model Tests</td>
<td>132</td>
</tr>
<tr>
<td>More and More Regressors</td>
<td>133</td>
</tr>
<tr>
<td>Interpreting Leverage Plots</td>
<td>134</td>
</tr>
<tr>
<td>Collinearity</td>
<td>135</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>136</td>
</tr>
</tbody>
</table>
Aerobic Fitness Data

Aerobic fitness can be evaluated using a special test that measures the oxygen uptake of a person while running on a treadmill for a prescribed distance. However, it would be more economical to evaluate fitness with a formula that predicts oxygen uptake with simpler measurements.

To identify such an equation, runtime and fitness measurements were taken for 31 participants who ran 1.5 miles. The participants’ ages were also recorded.

When you installed JMP, a folder named Sample Data was also installed near the application. In that folder is a file named Fitness.jmp. Open Fitness.jmp.

The data are shown in Figure 10.1. For purposes of illustration, certain values of MaxPulse and RunPulse have been changed from data reported by Rawlings (1988, p. 124).

Figure 10.1 Partial Listing of the Fitness.jmp Data File

Investigate Age and Runtime as predictors of oxygen uptake using the Fit Model analysis.

Choose Analyze > Fit Model and you should see the window shown here.
To specify a multiple regression model with two effects,

- Highlight Oxy in the Select Columns column.
- Click the Y button.
- Highlight both Age and Runtime.
- Click the Add button to specify them as model effects.
- Click Run Model.

You should now see the tables shown in Figure 10.2. These statistical reports are appropriate for a response variable and factor variables that have continuous values.

Figure 10.2 Statistical Text Reports
Clicking the red triangle icon and selecting **Save Columns** displays a list of save commands. To save predicted values and the prediction equation for this model:

- Click the red triangle icon and select **Save Columns > Prediction Formula**.

This command creates a new column in the Fitness data table called Pred Formula Oxy. Its values are the calculated predicted values for the model. To see the column's formula:

- Double-click the Pred Formula Oxy column.

The Column Info window box opens and displays the formula

\[ 88.4356809 - 0.1509571 \times \text{Age} - 3.1987736 \times \text{Runtime} \]

This formula defines a plane of fit for Oxy as a function of Age and Runtime.

- Click **Cancel** to close the Column Info window and return to the data table window.

Do the following for use later in this example:

- Select no columns and all rows (**Rows > Row Selection > Select All Rows**) and assign them the x marker.

- Double-click to the right of the last column in the table to create a new column.

- Name the new column **Markers**.

- Double-click the new column to access the Column Info window.

- Change the data type of the column to **Row State**.

- Click **OK** to close the Column Info window.

- Click the R ( ) or Rs (on Macintosh and Linux) icon beside the column name located in the column panel (to the left of the data grid).

- Select **Copy from Row States** to copy the markers to the data column.

---

**Fitting Plane**

JMP can show relationships between Oxy, Runtime, and Age in three dimensions with a spinning plot.

- Select the **Spinning Plot** command in the **Graph** menu tool bar.

- Add Oxy, Age, and Runtime as **Y, Columns** in the window.

- Click **OK**.

The plot shown in **Figure 10.3** has the **White Background** and the **Box** display options in effect.
Impose a fitting plane showing a grid of values that satisfies the prediction formula for Oxy. The next sections proceed through the following steps:

- Creating a new table having a grid of Age and Runtime values to use as x- and y- axes, and their predicted Oxy scores for the z-axis.
- Concatenating this grid table to the Fitness data table.
- Using the Spinning Plot command as before (Figure 10.3).

**Compute a Grid for Plotting**

- On Windows and Linux, select **File > New > Data Table**. On the Macintosh, select **File > New**.
- In the empty data table that appears, create three numeric continuous columns called Age, Runtime, and Oxy. These columns should match the corresponding names in the Fitness table for later concatenation.
- For convenience, rename the data table by clicking the name Untitled in the left column of the table and typing in the new name: Grid.
- Right-click the Age column (Command-click on the Macintosh) and select Formula from the menu that appears.
- Click Numeric in the function list and select the Count function.
- Enter 30 in the from slot.
- Enter 60 in the to slot.
- Select the steps slot.
- Click Row in the functions list and select NRow.
- Highlight the entire NRow slot, then click the root button in the formula editor's keypad.

- Repeat similar steps to give the Runtime column a formula to compute values that range from 8 to 14. Therefore, supply the from, to, and steps arguments as
Note: Highlight the NRow root slot, then press the comma key on the computer keyboard to access the fourth (initially hidden) argument.

- Click the Fitness table to make it active.
- Right-click (Command-click on the Macintosh) the Pred Formula Oxy column and select Formula from the menu that appears.

When the formula editor opens, select Edit > Copy to copy the formula to the clipboard.

- Click the Grid data table to make it active.
- Right-click (Command-click on the Macintosh) the Oxy column and select Formula from the menu that appears.
- Select Edit > Paste to paste the formula into the formula editor.
- Click OK to close the formula editor. See “Using the Formula Editor,” in the JMP User’s Guide, for information about modifying formulas in the formula editor window.
- Select Rows > Add Rows command to add 49 rows (giving a 7-by-7 grid).

The table automatically fills with values. Next, create unique markers (different from those used in the Fitness data table).

- Create a new column named Markers in the Grid data table.
- Select Cols > Column Info or double-click the new column to access the Column Info window.
- Make the data type Row State.
- Click OK to close the Column Info window.
- Select all rows and select Rows > Markers. Choose the diamond marker.
- Click the R ( ) or Rs (on Macintosh and Linux) icon beside the column name located in the column panel (to the left of the data grid).
- Select Copy from Row State from this menu.

Now the spinning plot can show this fitted plane.

Combine the Data Tables

The Fitness and Grid tables must be combined to plot the predicted Oxy values for the Age by Runtime grid along with the observed Oxy values.

- Click the Fitness table to make it active.
- Select Tables > Concatenate.

The concatenation window appears.
Select Grid from the list of open tables showing in the Concatenate window.

Click the Add button to add the Grid table to the list of tables to concatenate.

Click the Concatenate button.

The Fitness and Grid tables append end to end and form a new untitled table.

Rename the data table by clicking the name Untitled in the left column of the table and typing in the new name: Fitplane.

Figure 10.4 shows a partial listing of the new table. The first 31 rows have the observed Oxy values. Rows 32 to 80 (the rows from the grid table just created) have predicted Oxy values. Rows that were not in the Grid table have missing values.

Figure 10.4 Partial Listing of the Table with Computed Grid Values

To use the row state information,

Click the R ( ) or Rs (on Macintosh and Linux) icon beside the column name located in the column panel (to the left of the data grid).

Select the Copy to Row State command.

Select Graph > Spinning Plot.

Specify Age, Runtime, and Oxy as the Y, Columns to be plotted.

Click OK.

The result is a graphic dramatization of the actual points and the plane of fit (Figure 10.5).

Spin the plot to see how the observed points scatter about the plane.
Figure 10.5  Observed Points using Age, Oxy, and Runtime with the Predicted Plane of Fit

Click the home button on the spin control panel.
When the axes are in the position, the plot is identical to one of Oxy by Age.
Rotate the plot about the y-axis using the spin control panel until it reaches the position shown on the right in Figure 10.6.
This shows a plot of Oxy by Runtime, which illustrates a stronger relationship.

Figure 10.6  Rotate to See Relationships

Fit Planes to Test Effects

The example in the previous section showed a plane fit to the whole model. JMP is also useful to look at hypothesis tests for each regressor and to test whether the regressor’s parameter is significantly different from zero.

One way to view this test is to evaluate the difference between the current fit and the fit that occurs if the regressor variable is removed from the model.
For example, remove the Runtime variable from the model by following these steps:
First, make sure Fitness.jmp is the active data table.
Select Analyze > Fit Model, set oxy as y and add age as a construct model effect, and click Run Model.

Click the red triangle icon and select Save Columns > Prediction Formula.

The new predicted column (labeled Pred Formula Oxy 2) is calculated using the formula

$$62.4229492 + -0.3156031 \times \text{Age}$$

To compare this fitted line with the plane in the previous example,

- Create a new data table.
- Add new columns and label them Age, Runtime, and Oxy.
- Right-click the Age column (Command-click on the Macintosh) and select Formula from the menu that appears.
- Click Numeric in the function list and select the Count function.
- Enter 30 in the from slot.
- Enter 60 in the to slot.
- Select the steps slot.
- Click Row in the functions list and select NRow.
- Highlight the entire NRow slot, then click the root button in the formula editor's keypad.

Note: Highlight the NRow root slot, then press the comma key on the computer keyboard to access the fourth (initially hidden) argument.

Click OK to close the formula editor.

Right-click the Runtime column (Command-click on the Macintosh) and select Formula from the menu that appears.

Click Numeric in the function list and select the Count function.

Enter 8 in the from slot.

Enter 14 in the to slot.

Select the steps slot.

Click Row in the functions list and select NRow and use the root button in the formula editor to set the formula as

Note: Highlight the NRow root slot, then press the comma key on the computer keyboard to access the fourth (initially hidden) argument.

Click OK to close the formula editor.

Go to the fitness.jmp data table.

Right-click (Command-click on the Macintosh) the Pred Formula Oxy2 column and select Formula from the menu that appears.

When the formula editor opens, select Edit > Copy to copy the formula \(62.4229492 + -0.3156031 \times \text{Age}\) to the clipboard.
Click the newly-created untitled data table to make it active.

Right-click (Command-click on the Macintosh) the Oxy column and select **Formula** from the menu that appears.

Select **Edit > Paste** to paste the formula into the formula editor.

Click **OK** to close the formula editor.

Add 49 rows to the table.

Select **Tables > Concatenate** and choose to concatenate the new untitled table to the existing FitPlane table.

Observations 81 to 129 are the new grid, which has an Age slope but is flat along Runtime.

To examine the two regressor grid variables:

Select **Graph > Spinning Plot**.

Add Age, Runtime, and Oxy as the **Y, Columns** to spin.

Click **OK**.

Both grids represent least squares regression planes, but one plane has a slope of zero in the orientation of the Runtime axis. **Figure 10.7** shows the spinning plot from an angle.

**Figure 10.7** Three-Dimensional Plot with Regression Planes

Observed Values (x)

Oxy by Age regression plane (squares)

Oxy by Age and Runtime regression plane (diamonds)

Click the home button in the spin control panel.

This view is shown to the left in **Figure 10.8**. Notice this subset (Age-only model) regression showing as a line instead of a plane. The view is edge-on for Runtime, which eliminates it from the visual model.

Rotate the plot so that it appears like the one on the right of **Figure 10.8**.

This view shows a straight-line fit for a regression using Runtime as the only regressor.
Figure 10.8  Comparison of Three-Dimensional Views

Rotate the plot to look like the one in Figure 10.9.

The view in Figure 10.9 shows the bivariate regression plane edge-on and represents the linear combination of the effects fit by the plane.

Figure 10.9  Rotating to See Relationships

Leverage Plots: The Hypothesis-Eyed View

Rotate the plot from the previous example so that both the bivariate and subset regression planes show edge-on, as in Figure 10.10.

Figure 10.10 shows the subset regression plane oriented in the Z direction. This view best dramatizes the difference between the fit of the bivariate regression and the subset regression. Notice how much better the fit is for the bivariate regression plane than for the plane that represents the Age variable only.

Now suppose this view was altered slightly so that the y-axis is vertical but the restricted regression plane is still the horizontal edge. This view of the regression and the data is so revealing that it has a special name and a place in the text report. It is called a leverage plot and is an hypothesis-eyed view of the data.
Make sure the FitPlane table is the active table, and select Analyze > Fit Model. Select Oxy as Y and Age and Runtime as model effects. Click Run Model. Examine the three-dimensional plot for the direction of the response as shown on the left in Figure 10.10 to see the leverage plot configuration for Runtime as shown on the right.

A leverage plot is important because:
- The distance from each point to the line of fit represents the residual for the full model.
- The distance from each point to the horizontal line represents the residual for a model constrained by the hypothesis.

The example shown in Figure 10.10 illustrates the hypothesis that the parameter for Runtime is zero. A leverage plot can be formed for any linear hypothesis. Figure 10.11 shows the leverage plot for testing whether Age is significant. Note that Age does not relate as strongly to the response as Runtime. The 95% confidence curves fully contain the horizontal line in the leverage plot for Age, showing that the line of fit is not significantly different from the horizontal line representing the simple mean model.
Whole Model Tests

The leverage plot in Figure 10.12 shows the joint test of both the Age and Runtime effects in the model. This plot compares the full model with the model containing the intercept that fits the overall response mean only. This leverage plot is formed by plotting the actual observed values on the y-axis and the values predicted by the whole model on the x-axis. The residual for the subset model is the distance from a point to the horizontal line drawn at the sample mean.

**Figure 10.12** Leverage Plot for the Whole Model (Age and Runtime)
More and More Regressors

It’s easy to visualize two regressors predicting a response by using fitting planes. But how can this be done with more regressors when the analysis requires more than three dimensions? In actuality, the fitting, testing, and leverage plot analyses still work for more regressors.

Continuing with the previous example,

Using the same Fit Model window from the last example, add Weight, RunPulse, and MaxPulse to the model effects. If the window was accidentally closed, fill in a new one with Oxy as Y and add Age, Weight, Runtime, RunPulse, and Maxpulse as model effects.

Click Run Model to run the model.

In this case, the prediction formula is

\[ Oxy = 101.3 - 0.2323 \text{ Age} - 0.0732 \text{ Weight} - 2.688 \text{ Runtime} - 0.3703 \text{ RunPulse} + 0.3055 \text{ MaxPulse} \]

Look at the significance of each regressor with t-ratios in the Parameter Estimates table or F-ratios in the Effects Tests table (see Figure 10.13). Because each effect has only one parameter, the F-ratios are the squares of the t-ratios, and have the same significance probabilities.

The Age variable seems significant, but Weight does not. The Runtime variable seems highly significant. Both RunPulse and MaxPulse also seem significant, but MaxPulse is less significant than RunPulse.

Figure 10.13 Statistical Tables for Multiple Regression
Interpreting Leverage Plots

The leverage plots for this example multiple regression model allow visualization of the contribution of each effect. First, look at the whole-model leverage plot, shown to the right, of observed versus predicted values. This plot illustrates the test for the whole set of regressors.

The Analysis of Variance table in Figure 10.13 shows a highly significant F corresponding to this plot. The confidence curves show the strong relationship because they cross the horizontal line.

Now examine the leverage plots for the regressors. Each plot illustrates the residuals as they are and as they would be if that regressor were removed from the model.

The confidence curves in the leverage plot for Age in Figure 10.14 show that Age is borderline significant because the curves barely cross the horizontal line of the mean. Note that the significance of the Age effect is 0.03 in the text reports (Figure 10.13), which is only slightly different from the 0.05 confidence curves drawn by JMP.

The leverage plot for Weight shows that the effect is not significant. The confidence curves do not cross the horizontal line of the mean.

Figure 10.14 Leverage Plots for the Age and Weight Effects

The leverage plot for Runtime shows that Runtime is the most significant of all the regressors. The Runtime leverage line and its confidence curves cross the horizontal mean at a steep angle.
The leverage plots for RunPulse and MaxPulse shown in Figure 10.15 are similar. Each is somewhat shrunken on the x-axis. This indicates that other variables are related in a strong, linear fashion to these two regressors, which means the two effects are strongly correlated with each other.

**Figure 10.15** Leverage Plots for the RunPulse and MaxPulse Effects

**Collinearity**

When two or more regressors have a strong correlation, they are said to be collinear. These regression points occupy a narrow band showing their linear relationship.

When a plane is fit representing collinear regressors, the plane fits the points well in the direction where they are widely scattered. However, in the direction where the scatter is very narrow, the fit is weak and the plane is unstable.

In text reports, this phenomenon translates into high standard errors for the parameter estimates and potentially high values for the parameter estimates themselves. This occurs because a small random error 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 toward the center of the plot in the x direction.

The fitness example shows collinearity geometrically in the strongly correlated regressors, RunPulse and MaxPulse. To examine these regressors, examine **Figure 10.16**, which shows rotated views of the
regression planes. Most of the points are near the intersection of the two planes. When both planes are edge-on, as shown on the right of Figure 10.16, most of the scatter is hidden. From that angle, notice that the fitting plane representing both variables holds no better than the subset plane. The angle of the fitting plane is steep, but the hold is unstable.

Geometrically, collinearity between two regressors means that the points they represent do not spread out in x space enough to provide stable support for a plane. Instead, the points cluster around the center causing the plane to be unstable. The regressors act as substitutes for each other to define one direction redundantly. This is cured by dropping one of the collinear regressors from the model. In this case, drop either MaxPulse or RunPulse from the model because both measure essentially the same thing.

Figure 10.16  Comparison of RunPulse and MaxPulse Effects

Chapter Summary

Multiple regression uses the same fitting principle as simple regression, but accounting for significance is more subtle. Each regressor opens a new dimension for fitting a hyperplane, and its significance is tested by how much the fit suffers in its absence. Graphically, leverage plots provide the best view of each effect's partial contribution. When regressors correlate to each other, they are said to be collinear, and they define directions where the fitting hyperplane is not well supported.
References

Institute Inc.

Index

JMP Introductory Guide

A

Add Columns 22
Add Rows 24
Add Statistics Column 33
All Pairs, Tukey's HSD 64
analysis methods see platforms
Analysis of Variance see Fit Y by X, Fit Model
Analysis of Variance table 66, 87, 102
analysis role 10
analysis type see modeling type 10
analyze categorical data 69-80
Analyze menu 10
Annotate cursor 27
arrow cursor 7
assign role 25

B

bar chart 26
biplot 114
box plot 48
BP Study.jmp 24

C

C. Total 66
categorical analysis see Fit Y by X, Fit Model
categorical data 69-80
categorical type 10
character column 10
Charts 25, 34
Chi-Square 77
classification variable 11
Clear Row States 37
collinear 135
Column Info 24
column name 24
columns 6
Construct Model Effects 98, 101
continuous 10, 12, 60, 83
Construct Model Effects 98, 101
Copy from Row State 123, 125
Copy to Row State 126
Count 51
Cowboy Hat.jmp 14
create subset 51-54
crossed effects 98
Cum Prob 51
cursors in data table 7
curve-fitting 83
calculate example 53
construct formula 51-53
data grid 6
data table create 19-28
density contour 117
designed experiment 97
DF 66, 87, 102
design control 50
discrete data see nominal, ordinal
display options 12, 61
distance 117
Distribution 72, 109
disclosure example 55
documentation overview 3
double-arrow cursor 8, 53
drug experiment example 19-28

E

enter data 24
Estimate 87
Exclude/Include 88, 90
Exit 38
explanatory variable 11
explore data 107-118
extend selection 9
extreme values 51, 63

F

F Ratio 66, 87, 101-102
F Statistic 101


Factor role 11
factorial analysis example 95–106
fit by groups 93
Fit Line 85, 94
Fit Model 98, 121
fit plane 127
Fit Polynomial 90
Fit Y by X 38, 59, 75
Fit Spline 91
Fitness.jmp 121
formula 52
example 53
prediction 105, 123
F-probability see Prob > F
Freq role 11
frequencies see Distribution
frequency table 12, 51

G
Graph menu 10
Group By 93–94
grouped charts see Charts
grouped fitting 93
grouping data 32
grouping variable 11
Growth.jmp 83

H
hand cursor 8, 47
Help 3
high dimensionality 107
highlight see Select
histogram see Distribution
honestly significant difference see Tukey-Kramer
H SD
hotdog example 29–42
Hotdogs.jmp 31

I
I-beam cursor 7
independent variable 11
Index tab of JMP Starter 3
interaction 97, 103
interquartile range 48

J
JMP Starter window 5

J
Journal 89, 92
journaling analysis results 88

L
Label 37
Label/Unlabel 48, 109
lambda 91
large cross cursor 8
Least Squares Means table 103–104
Level 51, 66
leverage plot 100, 132
Likelihood Ratio 77
linked table 32
local error 92
logistic regression see Fit Y by X, Fit Model
LS Means Plot 103

M
magnifier tool 27
Mahalanobis distance 117
main effect 97–98
Markers 37, 88, 110
Mean 66
Mean of Response 65, 87, 103
Mean Square 66, 87, 102
Means Diamonds 49, 61–62
Means for Oneway Anova table 66
Means, Anova/t-test 61
median 49
modeling type 9–10, 60
modify data table 72–75
moments see Distribution
mosaic plot see Distribution
multiple comparison see Compare Means
multiple regression 119, 122–136
Multivariate 111–113

N
New Column 51, 72
computing values 53
New Data Table 5
nominal 10, 12, 60, 75, 83
normal distribution 47, 49
notation used in manuals 4
Number 66
numeric column 10
Observations 65, 87, 103
Open Data Table 5
ordinal 10, 32, 60, 83
outlier 48, 88, 107, 117
outlier box plot see Distribution

Parameter Estimates table 87
partitioning 101
pattern in data 107
Pearson Chi Square 77
percentile see quantile
plane fit 127
platforms 10
pointer cursor 8
Polynomial fit 90–91
post hoc see Compare Means
Prediction Formula 104
principal components analysis 113
Prob 51
Prob > |t| 88
Prob > F 66, 87, 102

Quantile Box Plot 49, 63
Quantile Box Plot see Distribution
Quit 18

R² 91
references 137
regression analysis 83
regression example 83
regression line 99
regression see Fit Y by X, Fit Model
Remove Fit 86, 91
report options 12
rescale axis 26
resize plot 53, 89
Response role 11
Return 25
role 10, 25, 75
Root Mean Square Error (RMSE) 65, 86, 103
rows 6
Rsquare 65, 86, 103
Rsquare Adj 65, 86, 103

Save As 24, 89
Save Predicteds 86
Save Prediction Formula 123
saving a JMP session 13
scatterplot matrix 111
scatterplot see Fit Y by X, Multivariate, Overlay Plot
Select All Rows 37
select rows and columns 9
selection tool 68
Shift-Return 25
Shift-Tab 25
shortest half 48
Show Points 85
smoothing 91
solubility study 107–118
Solubility.jmp 109
Source 66, 87, 107
spin controls 112
Spinning Plot 123, 126
example 14
spline 91
start JMP 21
statistical summaries see Distribution
Std Error 66, 87
StdErrProb 51
Students.jmp 55
Subset 51–54, 110
Sum of Squares 66, 87, 102
summarizing data 29–42
Summary 32
Summary of Fit table 65, 86, 91, 103
survey data 69–80

tab 25

tension 91
term 87
three-dimensional plots 119–136
three-dimensional plots see Spinning Plot
tick mark 61
title of report 27
toggle 61
Tools 27
t-ratio 88
t-test 88
tutorial examples
data table 19–28
drug experiment 19–28
exploratory study 107–118
multiple regression 119–136
popcorn experiment 95–106
regression analysis 83–106
school health study 55
summarizing data 29–42
survey data 69–80
typing study 57–68
Typing Data jmp 59
typing study 57–68

W-Z
weight role 11
weight-height ratio example 83
whiskers 48
White Background 123
Whole-model plot 99
X, Factor role 11
Y, Response role 11