

# JMP® Start Statistics

A Guide to Statistics and  
Data Analysis Using JMP®

Sixth Edition



John Sall, Ann Lehman, Mia Stephens, Sheila Loring

The correct bibliographic citation for this manual is as follows: John Sall, Ann Lehman, Mia Stephens, and Sheila Loring. *JMP® Start Statistics: A Guide to Statistics and Data Analysis Using JMP®, Sixth Edition*. Cary, NC: SAS Institute Inc.

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

Copyright © 2017, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-62960-875-4 (Hardcopy)

ISBN 978-1-62960-876-1 (EPUB)

ISBN 978-1-62960-877-8 (MOBI)

ISBN 978-1-62960-878-5 (Web PDF)

All rights reserved. Produced in the United States of America.

**For a hard-copy book:** No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**For a Web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**U.S. Government Restricted Rights Notice:** The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a) and DFAR 227.7202-4 and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414

February 2017

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

# Table of Contents

## Preface xv

- The Software xv
- How to Get JMP xvi
- JMP Start Statistics, Sixth Edition* xvii
- SAS xvii
- JMP versus JMP Pro xviii
- This Book xviii

## 1 Preliminaries 1

- What You Need to Know 1
  - ...about statistics 1
- Learning about JMP 1
  - ...on your own with JMP Help 1
  - ...hands-on examples 2
  - ...using Tutorials 2
  - ...reading about JMP 2
- Chapter Organization 3
- Typographical Conventions 5

## 2 Getting Started with JMP 7

- Hello! 7
- First Session 9
  - Tip of the Day 9
  - The JMP Starter (Macintosh) 9
  - The JMP Home Window (Windows) 10
  - Open a JMP Data Table 12
  - Launch an Analysis Platform 14
  - Interact with the Report Surface 15
  - Special Tools 18
- Customize JMP 19
- Modeling Type 21
  - Analyze and Graph 22
  - Navigating Platforms and Building Context 22
  - Contexts for a Histogram 23

Contexts for the <i>t</i> -Test	23
Contexts for a Scatterplot	24
Contexts for Nonparametric Statistics	24
The Personality of JMP	25
<b>3 Data Tables, Reports, and Scripts</b>	<b>27</b>
Overview	27
The Ins and Outs of a JMP Data Table	29
Selecting and Deselecting Rows and Columns	30
Mousing around a Data Table: Cursor Forms	30
Creating a New JMP Table	32
Define Rows and Columns	33
Enter Data	35
The New Column Command	36
Plot the Data	37
Importing Data	39
Importing Text Files	41
Importing Other File Types	44
Copy, Paste, and Drag Data	46
Moving Data Out of JMP	47
Saving Graphs and Reports	48
Copy and Paste	48
Drag Report Elements	49
Save JMP Reports and Graphs	49
Create Interactive Web Reports	49
Pop-up Menu Commands	50
Juggling Data Tables	51
Data Management	51
Give New Shape to a Table: Stack Columns	52
Creating Summary Statistics	55
Create Summary Statistics with the Summary Command	55
Create Summary Statistics with Tabulate	58
Working with Scripts	60
Creating Scripts	60
Running Data Table Scripts	60
Opening and Running Stand-alone Scripts	61
<b>4 Formula Editor</b>	<b>63</b>
Overview	63
The Formula Editor Window	65
The Formula Editor and the JMP Scripting Language	66
A Quick Example: Standardizing Data	67

Making a New Formula Column	69
Using Popular Formula Functions	71
Writing Conditional Expressions	72
Summarizing Data with the Formula Editor	77
Generating Random Data	82
Local Variables and Table Variables	87
Working with Dates	89
Tips on Building Formulas	90
Examining Expression Values	90
Cutting, Dragging, and Pasting Formulas	90
Selecting Expressions	91
Exercises	91

## 5 What Are Statistics? 95

Overview	95
Ponderings	97
The Business of Statistics	97
The Yin and Yang of Statistics	97
The Faces of Statistics	98
Don't Panic	99
Preparations	101
Three Levels of Uncertainty	101
Probability and Randomness	102
Assumptions	102
Data Mining?	103
Statistical Terms	104

## 6 Simulations 109

Overview	109
Rolling Dice	111
Rolling Several Dice	114
Flipping Coins, Sampling Candy, or Drawing Marbles	114
Probability of Making a Triangle	115
Confidence Intervals	120
Data Table-Based Simulations	121
Other JMP Simulators	122
Exercises	123

## 7 Univariate Distributions: One Variable, One Sample 125

Overview	125
Looking at Distributions	128

Probability Distributions	130
True Distribution Function or Real-World Sample Distribution	131
The Normal Distribution	133
Describing Distributions of Values	134
Generating Random Data	134
Histograms	135
Stem-and-Leaf Plots	137
Dot Plots	138
Outlier and Quantile Box Plots	139
Mean and Standard Deviation	141
Median and Other Quantiles	142
Mean versus Median	142
Other Summary Statistics: Skewness and Kurtosis	143
Extremes, Tail Detail	143
Statistical Inference on the Mean	144
Standard Error of the Mean	144
Confidence Intervals for the Mean	144
Testing Hypotheses: Terminology	147
The Normal $z$ -Test for the Mean	149
Case Study: The Earth's Ecliptic	150
Student's $t$ -Test	152
Comparing the Normal and Student's $t$ Distributions	153
Testing the Mean	154
The $p$ -Value Animation	155
Power of the $t$ -Test	157
Practical Significance versus Statistical Significance	159
Examining for Normality	161
Normal Quantile Plots	162
Statistical Tests for Normality	165
Special Topic: Practical Difference	167
Special Topic: Simulating the Central Limit Theorem	170
Seeing Kernel Density Estimates	172
Exercises	173

## 8 The Difference Between Two Means 177

Overview	177
Two Independent Groups	179
When the Difference Isn't Significant	179
Check the Data	180
Launch the Fit Y by X Platform	181
Examine the Plot	182
Display and Compare the Means	183
Inside the Student's $t$ -Test	184

Equal or Unequal Variances?	185
One-Sided Version of the Test	187
Analysis of Variance and the All-Purpose <i>F</i> -Test	188
How Sensitive Is the Test?	
How Many More Observations Are Needed?	190
When the Difference Is Significant	192
Normality and Normal Quantile Plots	194
Testing Means for Matched Pairs	196
Thermometer Tests	197
Look at the Data	198
Look at the Distribution of the Difference	199
Student's <i>t</i> -Test	199
The Matched Pairs Platform for a Paired <i>t</i> -Test	200
Optional Topic: An Equivalent Test for Stacked Data	203
Two Extremes of Neglecting the Pairing Situation: A Dramatization	205
A Nonparametric Approach	211
Introduction to Nonparametric Methods	211
Paired Means: The Wilcoxon Signed-Rank Test	211
Independent Means: The Wilcoxon Rank Sum Test	213
Exercises	214
<b>9 Comparing Many Means: One-Way Analysis of Variance</b>	<b>217</b>
Overview	217
What Is a One-Way Layout?	219
Comparing and Testing Means	221
Means Diamonds: A Graphical Description of Group Means	222
Statistical Tests to Compare Means	223
Means Comparisons for Balanced Data	226
Means Comparisons for Unbalanced Data	227
Adjusting for Multiple Comparisons	232
Are the Variances Equal across the Groups?	235
Testing Means with Unequal Variances	238
Nonparametric Methods	239
Review of Rank-Based Nonparametric Methods	239
The Three Rank Tests in JMP	240
Exercises	242
<b>10 Fitting Curves through Points: Regression</b>	<b>245</b>
Overview	245
Regression	247

Least Squares	247
Seeing Least Squares	248
Fitting a Line and Testing the Slope	250
Testing the Slope By Comparing Models	252
The Distribution of the Parameter Estimates	255
Confidence Intervals on the Estimates	256
Examine Residuals	258
Exclusion of Rows	258
Time to Clean Up	260
Polynomial Models	260
Look at the Residuals	261
Higher-Order Polynomials	261
Distribution of Residuals	262
Transformed Fits	263
Spline Fit	265
Are Graphics Important?	266
Why It's Called Regression	269
What Happens When X and Y Are Switched?	271
Curiosities	274
Sometimes It's the Picture That Fools You	274
High-Order Polynomial Pitfall	275
The Pappus Mystery on the Obliquity of the Ecliptic	276
Exercises	277

## 11 Categorical Distributions 281

Overview	281
Categorical Situations	283
Categorical Responses and Count Data: Two Outlooks	283
A Simulated Categorical Response	286
Simulating Some Categorical Response Data	287
Variability in the Estimates	288
Larger Sample Sizes	290
Monte Carlo Simulations for the Estimators	291
Distribution of the Estimates	292
The $\chi^2$ Pearson Chi-Square Test Statistic	293
The $G^2$ Likelihood-Ratio Chi-Square Test Statistic	294
Likelihood Ratio Tests	295
The $G^2$ Likelihood Ratio Chi-Square Test	296
Univariate Categorical Chi-Square Tests	296
Comparing Univariate Distributions	297
Charting to Compare Results	299

Exercises	301
<b>12 Categorical Models</b>	<b>303</b>
Overview	303
Fitting Categorical Responses to Categorical Factors: Contingency Tables	305
Testing with $G^2$ and $X^2$ Statistic	305
Looking at Survey Data	306
Car Brand by Marital Status	310
Car Brand by Size of Vehicle	311
Two-Way Tables: Entering Count Data	312
Expected Values under Independence	313
Entering Two-Way Data into JMP	314
Testing for Independence	314
If You Have a Perfect Fit	316
Special Topic: Correspondence Analysis— Looking at Data with Many Levels	318
Continuous Factors with Categorical Responses: Logistic Regression	321
Fitting a Logistic Model	321
Degrees of Fit	325
A Discriminant Alternative	326
Inverse Prediction	327
Polytomous (Multinomial) Responses: More Than Two Levels	330
Ordinal Responses: Cumulative Ordinal Logistic Regression	331
Surprise: Simpson's Paradox: Aggregate Data versus Grouped Data	334
Generalized Linear Models	337
Exercises	342
<b>13 Multiple Regression</b>	<b>345</b>
Overview	345
Parts of a Regression Model	347
Regression Definitions	347
A Multiple Regression Example	348
Residuals and Predicted Values	351
The Analysis of Variance Table	354
The Whole Model F-Test	354
Whole-Model Leverage Plot	355
Details on Effect Tests	356
Effect Leverage Plots	356
Collinearity	358
Exact Collinearity, Singularity, and Linear Dependency	362
The Longley Data: An Example of Collinearity	364

The Case of the Hidden Leverage Point	366
Mining Data with Stepwise Regression	369
Exercises	373

## 14 Fitting Linear Models 377

Overview	377
The General Linear Model	379
Types of Effects in Linear Models	380
Coding Scheme to Fit a One-Way ANOVA as a Linear Model	381
Regressor Construction	384
Interpretation of Parameters	385
Predictions Are the Means	385
Parameters and Means	385
Analysis of Covariance: Continuous and Categorical Terms in the Same Model	386
The Prediction Equation	389
The Whole-Model Test and Leverage Plot	390
Effect Tests and Leverage Plots	391
Least Squares Means	393
Lack of Fit	394
Separate Slopes: When the Covariate Interacts with a Categorical Effect	396
Two-Way Analysis of Variance and Interactions	400
Optional Topic: Random Effects and Nested Effects	406
Nesting	407
Repeated Measures	409
Method 1: Random Effects-Mixed Model	409
Method 2: Reduction to the Experimental Unit	414
Method 3: Correlated Measurements-Multivariate Model	416
Varieties of Analysis	418
Closing Thoughts	418
Exercises	419

## 15 Design of Experiments 421

Overview	421
Introduction	424
Key Concepts	424
JMP DOE	425
A Simple Design	426
The Experiment	426
Enter the Response and Factors	427
Define the Model	429
Is the Design Balanced?	432
Perform Experiment and Enter Data	432

Analyze the Model	433
Flour Paste Conclusions	439
Details of the Design: Confounding Structure	439
Using the Custom Designer	440
How the Custom Designer Works	440
Choices in the Custom Designer	441
An Interaction Model: The Reactor Data	442
Analyzing the Reactor Data	444
Where Do We Go from Here?	450
Some Routine Screening Examples	452
Main Effects Only (a Review)	452
All Two-Factor Interactions Involving a Single Factor	453
Alias Optimal Designs	455
Response Surface Designs	456
The Odor Experiment	456
Response Surface Designs in JMP	456
Analyzing the Odor Response Surface Design	458
Plotting Surface Effects	461
Specifying Response Surface Effects Manually	462
The Custom Designer versus the Response Surface Design Platform	463
Split-Plot Designs	464
The Box Corrosion Split-Plot Experiment	465
Designing the Experiment	465
Analysis of Split-Plot Designs	467
Design Strategies	471
DOE Glossary of Key Terms	472
Exercises	476
<b>16 Bivariate and Multivariate Relationships</b>	<b>479</b>
Overview	479
Bivariate Distributions	481
Density Estimation	481
Bivariate Density Estimation	482
Mixtures, Modes, and Clusters	484
The Elliptical Contours of the Normal Distribution	485
Correlations and the Bivariate Normal	487
Simulating Bivariate Correlations	487
Correlations across Many Variables	490
Bivariate Outliers	491
Outliers in Three and More Dimensions	494
Identify Variation with Principal Components Analysis	496
Principal Components for Six Variables	499

How Many Principal Components?	501
Discriminant Analysis	502
Canonical Plot	503
Discriminant Scores	504
Stepwise Discriminant Variable Selection	507
Cluster Analysis	508
Hierarchical Clustering: How Does It Work?	508
A Real-World Example	511
Some Final Thoughts	514
Exercises	515
<b>17 Exploratory Modeling</b>	<b>517</b>
Overview	517
Recursive Partitioning (Decision Trees)	519
Growing Trees	521
Exploratory Modeling with Partition	528
Saving Columns and Formulas	530
Neural Nets	531
A Simple Example	532
Modeling with Neural Networks	535
Saving Columns	535
Profiles in Neural	537
Exercises	541
<b>18 Control Charts and Capability</b>	<b>545</b>
Overview	545
What Does a Control Chart Look Like	548
Types of Control Charts	549
Variables Charts	550
Attributes Charts	551
Specialty Charts	551
Control Chart Basics	551
Control Charts for Variables Data	552
Variables Charts Using Control Chart Builder	553
The Control Chart Builder Work Space	553
Control Chart Builder Examples	554
Control Charts for Attributes Data	557
Specialty Charts	560
Presummarize Charts	560
Levey-Jennings Charts	561
Uniformly Weighted Moving Average (UWMA) Charts	561

Exponentially Weighted Moving Average (EWMA) Chart	563
<b>Capability Analysis</b>	<b>564</b>
What Is Process Capability?	564
Capability for One Process Measurement	567
Capability for Many Process Measurements	569
Capability for Time-Ordered Data	572
<b>A Few Words about Measurement Systems</b>	<b>574</b>
Exercises	574

## **19 Mechanics of Statistics 577**

Overview	577
<b>Springs for Continuous Responses</b>	<b>579</b>
Fitting a Mean	579
Testing a Hypothesis	580
One-Way Layout	581
Effect of Sample Size Significance	581
Effect of Error Variance on Significance	582
Experimental Design's Effect on Significance	583
Simple Regression	584
Leverage	585
Multiple Regression	586
Summary: Significance and Power	586
<b>Mechanics of Fit for Categorical Responses</b>	<b>586</b>
How Do Pressure Cylinders Behave?	587
Estimating Probabilities	588
One-Way Layout for Categorical Data	589
Logistic Regression	591

## **A Answers to Selected Exercises 593**

Chapter 4, "Formula Editor"	593
Chapter 7, "Univariate Distributions: One Variable, One Sample"	596
Chapter 8, "The Difference Between Two Means"	600
Chapter 9, "Comparing Many Means: One-Way Analysis of Variance"	602
Chapter 10, "Fitting Curves through Points: Regression"	606
Chapter 11, "Categorical Distributions"	609
Chapter 12, "Categorical Models"	610
Chapter 13, "Multiple Regression"	612
Chapter 14, "Fitting Linear Models"	613
Chapter 15, "Design of Experiments"	614
Chapter 16, "Bivariate and Multivariate Relationships"	615
Chapter 17, "Exploratory Modeling"	617
Chapter 18, "Control Charts and Capability"	617

**B References and Data Sources** 619

**Technology License Notices** 625

**Index** 627



# Preface

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

## The Software

As statistical discovery software, JMP emphasizes working interactively with data and graphics in a progressive structure to make discoveries.

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

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

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

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

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

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

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

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

## How to Get JMP

There are several ways to get JMP:

- JMP is available through department or campus licenses at most colleges and universities and through site licenses in many organizations. See your software IT administrator for availability and download information.
- Individual copies of JMP for academic use are also available from <http://onthehub.com/jmp>. If you would like more information about academic licensing or would like to request an evaluation copy of JMP for classroom use, email [academic@jmp.com](mailto:academic@jmp.com).

- If you do not qualify for an academic license, a trial version of JMP is available at <http://jmp.com/trial>. Read license information at <http://jmp.com/buy>.

## ***JMP Start Statistics, Sixth Edition***

JMP Start Statistics has been updated and revised to feature JMP 13. Major enhancements have been made to JMP since the fifth edition, which was based on JMP 10. The new enhancements include DOE (Design Evaluation, new Custom Design options, and Definitive Screening designs), analysis and modeling (Generalized Regression, Partition enhancements, Model Comparison, and Formula Depot), data preparation (handling missing values and outliers, and model validation), and graphics (continued development of the interactive Graph Builder), most of which are covered in this book. In addition, the menus have been restructured, and we've added functionality for getting data into JMP (Query Builder) and sharing results (saving as Microsoft PowerPoint, saving as HTML, and creating interactive web reports).

JMP 13 also continues our focus on enhancing the user experience, with new daily Tips of the Day and expanded documentation.

We include discussion of many of these new features throughout this text.

## **SAS**

SAS, or the SAS System, is an integrated statistical software system used by universities, research institutions, and industries across the globe. JMP Statistical Discovery Software is desktop software from SAS that runs natively on Mac and Windows. JMP was originally designed as a personal analysis tool for engineers and scientists, but is now used in a variety of applications and industries worldwide.

## JMP versus JMP Pro

JMP was first released by SAS in 1989 to run on a Macintosh operating system, and became available on Windows in the early 1990s. Since then, JMP has grown into a family of products, each designed to meet particular needs.

In this book we use JMP Pro, which includes advanced tools for analytics and predictive modeling. However, JMP Pro is not required to take full advantage of the methods covered. Unless otherwise specified, the features that we discuss are available in both JMP and JMP Pro.

## This Book

### **Software Manual and Statistics Text**

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

### **Not Just the Basics**

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

JMP also includes several advanced methods not covered in this book, such as nonlinear regression, multivariate analysis of variance, tools for predictive modeling and data mining, consumer research methods, text mining, and some advanced design of experiments capabilities. If you are planning to use these features extensively, it is recommended that you refer to the Help system or the JMP documentation for the professional version of JMP.

### **Examples Both Real and Simulated**

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

also available on the Internet at <http://support.sas.com/stephens>. JMP can also import data from files that are distributed with other textbooks. See Chapter 3, “Data Tables, Reports, and Scripts,” for details about importing various types of data.

### **Acknowledgments**

Thank you to the JMP testers as well as the contributors and reviewers of earlier versions of *JMP Start Statistics*: Bradley Jones, Chris Gotwalt, Lou Valente, Tom Donnelly, Michael Benson, Avignor Cahaner, Howard Yetter, David Ikle, Robert Stine, Andy Mauromoustkos, Al Best, Jacques Goupy, and Chris Olsen for contributions to earlier versions of the book. Special thanks to Curt Hinrichs for invaluable support to the JMP Start Statistics project.





# 1

## Preliminaries

### What You Need to Know

#### ...about statistics

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

### Learning about JMP

#### ...on your own with JMP Help

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

There are several ways to access JMP Help:

- Select **JMP Help** from the **Help** menu.
- You can click the **Help** button in launch windows whenever you launch an analysis or graph platform.
- After you generate a report, click the Help tool (?) on the **Tools** menu or toolbar and click the report surface. Context-sensitive help tells about the items that you click.

## ...hands-on examples

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

☞ Steps for example analyses begin with the mouse symbol in the margin, like this paragraph.

## ...using Tutorials

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

## ...reading about JMP

JMP is accompanied by a series of built-in reference manuals, a menu reference card and a quick reference card. The newest in the series of guides, *Discovering JMP*, provides a general introduction to JMP. It contains basic examples and descriptions that give you a feel for JMP and can get you started.

*Discovering JMP* is followed by *Using JMP*, which helps new users understand JMP data tables and how to perform basic operations. *Using JMP* is followed by several books that document all of the JMP analysis and graph platforms. In addition, there are specialty books for design of experiments and the JMP scripting language. These references cover all the commands and options in JMP and have extensive examples of the **Analyze**, **Graph**, and **DOE** platforms.

The documentation is available in the following formats:

- In-product help (Select the **Help > JMP Help** menu.)
- PDF files (Select the **Help > Books** menu.)
- e-books
- Help at <http://jmp.com/support/help>
- Print books

# Chapter Organization

The chapters of this book are supported by guided actions that you can take to become familiar with JMP.

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

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

Chapters 6–19 cover the array of analysis techniques offered by JMP. Chapters begin with simple-to-use techniques and gradually work toward more complex methods. Emphasis is on learning to think about these techniques and on how to visualize data analysis at work. JMP offers a graph for almost every statistic and supporting tables for every graph. Using highly interactive methods, you can learn more quickly and discover what your data has to say.

- Chapter 6, “Simulations,” introduces you to some probability topics by using the JMP scripting language. You learn how to open and execute these scripts and to see other ways of simulating data in JMP.

- Chapter 7, “Univariate Distributions: One Variable, One Sample,” covers distributions of continuous and categorical variables and statistics to test univariate distributions.
- Chapter 8, “The Difference Between Two Means,” covers  $t$  tests of independent groups and tells how to handle paired data. The nonparametric approach to testing related pairs is also shown.
- Chapter 9, “Comparing Many Means: One-Way Analysis of Variance,” covers one-way analysis of variance, with standard statistics and a variety of graphical techniques.
- Chapter 10, “Fitting Curves through Points: Regression,” shows how to fit a regression model for a single factor.
- Chapter 11, “Categorical Distributions,” discusses how to think about the variability in single batches of categorical data. It covers estimating and testing probabilities in categorical distributions, shows Monte Carlo methods, and introduces the Pearson and Likelihood ratio chi-square statistics.
- Chapter 12, “Categorical Models,” covers fitting categorical responses to a model, starting with the usual tests of independence in a two-way table, and continuing with graphical techniques and logistic regression.
- Chapter 13, “Multiple Regression,” describes the parts of a linear model with continuous factors, talks about fitting models with multiple numeric effects, and shows a variety of examples, including the use of stepwise regression to find active effects.
- Chapter 14, “Fitting Linear Models,” is an advanced chapter that continues the discussion of Chapter 12. The chapter moves on to categorical effects and complex effects, such as interactions and nesting.
- Chapter 15, “Design of Experiments,” looks at the built-in commands in JMP used to generate specified experimental designs. It also looks at examples of how to analyze common screening and response-level designs are covered.
- Chapter 16, “Bivariate and Multivariate Relationships,” looks at ways to examine two or more response variables using correlations, scatterplot matrices, three-dimensional plots, principal components, and other techniques. Discriminant and Cluster Analysis discuss methods that group data into clumps. Outliers are discussed.
- Chapter 17, “Exploratory Modeling,” illustrates common data mining techniques—Neural Nets and Recursive Partitioning.

- Chapter 18, “Control Charts and Capability,” discusses common types of control charts for both continuous and attribute data, and introduces process capability studies.
- Chapter 19, “Mechanics of Statistics,” is an essay about statistical fitting that might prove enlightening to those who enjoy mechanics.

## Typographical Conventions

The following conventions help you relate written material in this book to information that you see on your screen.

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



- References to variable names, data table names, and some items in reports appear in Helvetica but can appear in illustrations in either a plain or boldface font. These items show on your screen as you have specified in your JMP Preferences.
- Words or phrases that are important, new, or have definitions specific to JMP are in *italics* the first time you see them.

- When there is an action statement, you can do the example yourself by following the instructions. These statements are preceded by a mouse symbol ( ) in the margin. An example of an action statement is:
  - ⓐ Highlight the Month column by clicking the area above the column name, and then select **Cols > Column Info**.
- Occasionally, special information is in a boxed side bar in Helvetica to help distinguish them from the text flow.



# Getting Started with JMP

## Hello!

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

If you are already familiar with JMP and want to dive right into statistics, you can skip ahead to Chapters 6–19. You can always return later for more details about using JMP or for more details about statistics.

## Chapter Contents

Hello!	7
First Session.	9
Tip of the Day	9
The JMP Starter (Macintosh)	9
The JMP Home Window (Windows)	10
Open a JMP Data Table.	12
Launch an Analysis Platform	14
Interact with the Report Surface	15
Special Tools	18
Customize JMP	19
Modeling Type	21
Analyze and Graph	22
Navigating Platforms and Building Context	22
Contexts for a Histogram	23
Contexts for the t-Test	23
Contexts for a Scatterplot	24
Contexts for Nonparametric Statistics	24
The Personality of JMP	25

## First Session

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

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

The active JMP application displays several items by default. You can use general JMP preferences to show only what you want to see when starting JMP.

### Tip of the Day

Both the Macintosh and Windows environments begin by showing the Tip of the Day. There are many of these handy tips. But as a rule, they are useful only if you are an advanced user. If you are just starting or not interested in the tips, deselect the **Show tips at startup** box in the lower left corner of the tip. Select **Help > Tip of the Day** to see the tips at any time.

### The JMP Starter (Macintosh)

When the application begins, the Macintosh environment shows the JMP menu bar and the JMP Starter window. Appropriate Macintosh toolbars are also attached to analysis windows and therefore vary. When a JMP data table or analysis is open, use **View > Customize Toolbar** to customize the toolbars. Drag the desired item to the toolbar.

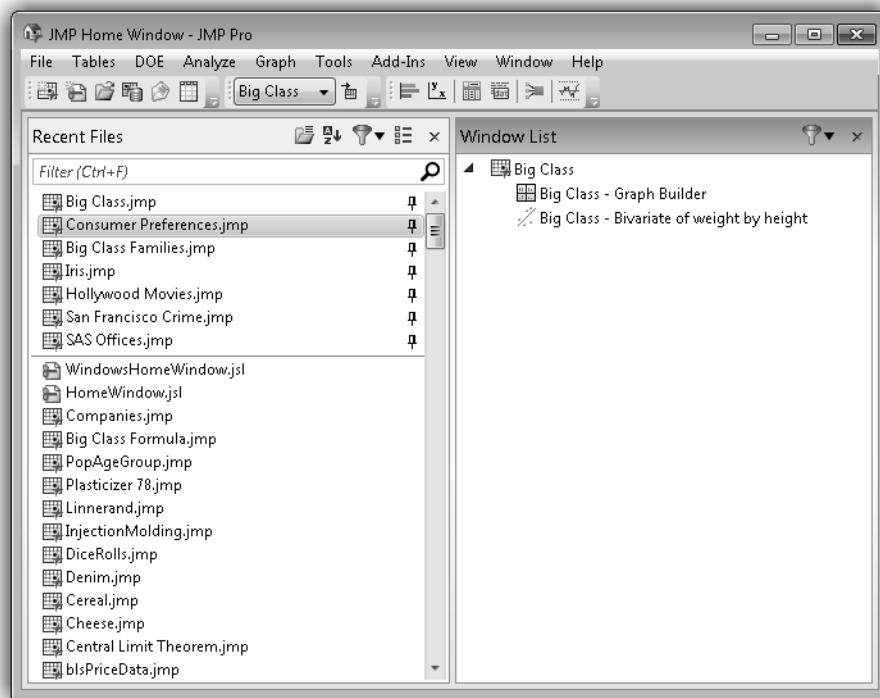
**Figure 2.1** The JMP Main Menu, Toolbar, and the JMP Starter (Macintosh)

The JMP Starter Window displays most of the commands found in the main menu and toolbars. You might find the JMP Starter helpful if you are not familiar with JMP or data analysis because the Starter briefly describes each option and report. On Windows, you can see the JMP Starter using **View > JMP Starter**.

## The JMP Home Window (Windows)

On Windows, opening JMP displays the JMP Home Window (**Figure 2.2**). It might show behind the Tip of the Day – close the Tip window or click the JMP Home Window to bring it to the front.

**Note:** You can open the Home Window on Macintosh by selecting **View > Home Window**.

**Figure 2.2** JMP Home Window (Windows)

The JMP Home Window is completely customizable. You can resize its panes or choose which panes to keep open. Once you begin using JMP, importing, opening, or creating tables, and doing analyses, the JMP Home Window becomes an invaluable desk organizer. However, closing the JMP Home Window when nothing else is open automatically closes the JMP session after asking you if you are ready to exit JMP.

You can always close JMP by selecting **File > Exit JMP** on Windows or **JMP > Quit JMP** on the Macintosh.

**Note:** A home window is also available on Macintosh, but it doesn't display by default. Select **Window > JMP Home** to show the JMP Home Window. Closing the home window on Macintosh doesn't close JMP.

So, get your toes wet by opening a JMP data table and doing a simple analysis.

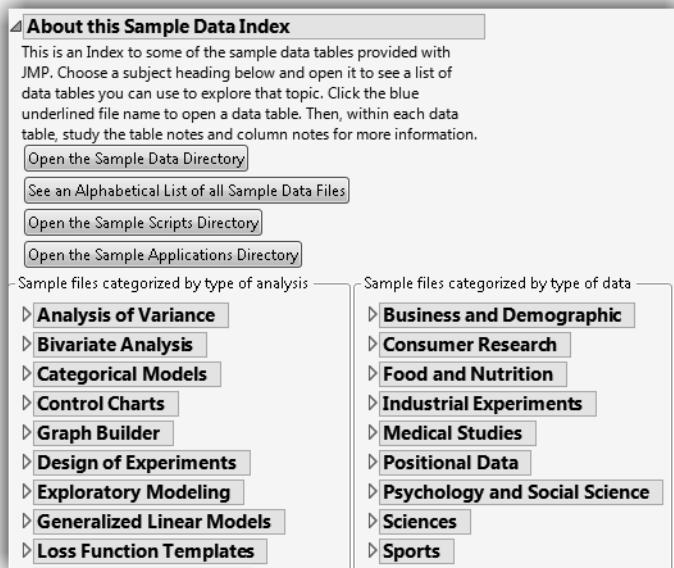
## Open a JMP Data Table

Begin by starting JMP, if you haven't already. Instead of starting with a blank file or importing data from text files, open a JMP data table from the collection of sample data tables that comes with JMP.

The JMP sample data is most easily accessed by selecting **Sample Data Library** from the **Help** menu. You can also access the sample data by opening the Sample Data Index window from the **Help > Sample Data** menu.

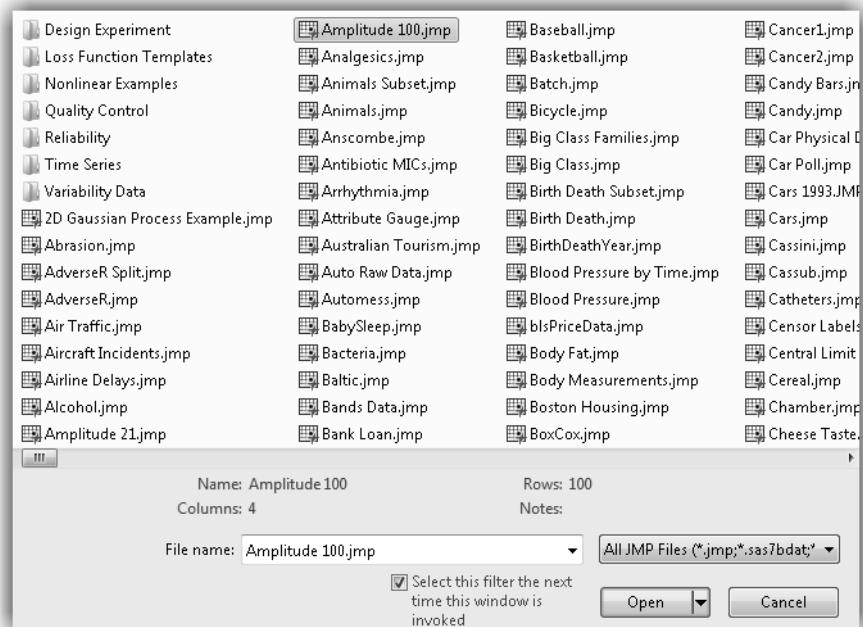
- ☞ Select **Sample Data** from the **Help** menu (**Help > Sample Data**) to see the window in **Figure 2.3**.

**Figure 2.3** Top Portion of the Sample Data Index from the JMP Help Menu



The data tables are organized in outlines by subject matter and appropriate type of analysis. You can also select a table from a complete alphabetical list of tables or see the Open File window for the JMP sample library.

- ☞ Click **Open the Sample Data Directory** in the Sample Data Index window to see all the folders and JMP tables in the JMP sample library.
- ☞ When the Open File window appears (**Figure 2.4**), select Big Class.jmp and click **Open** on the window or just double-click Big Class.jmp to open it.
- ☞ Close the Sample Data Index window.

**Figure 2.4** Open File Window (Windows)

You should now see the JMP table in **Figure 2.5** with columns titled name, age, sex, height, and weight.

**Figure 2.5** Partial Listing of the Big Class Data Table

		name	age	sex	height	weight
▼	Big Class	KATIE	12	F	59	95
►	Distribution	LOUISE	12	F	61	123
►	Bivariate	JANE	12	F	55	74
►	Oneway	JACLYN	12	F	66	145
►	Logistic	LILLIE	12	F	52	64
►	Contingency	TIM	12	M	60	84
►	Fit Model	JAMES	12	M	61	128
▼	Columns (5/0)	ROBERT	12	M	51	79
►	name	BARBARA	13	F	60	112
►	age	ALICE	13	F	61	107
►	sex	SUSAN	13	F	56	67
►	height	JOHN	13	M	65	98
►	weight	JOE	13	M	63	105
▼	Rows	MICHAEL	13	M	58	95
All rows	40	DAVID	13	M	59	79
Selected	0	JUDY	14	F	61	81
Excluded	0	ELIZABETH	14	F	62	91
Hidden	0	LESLIE	14	F	65	142
Labelled	0					

Chapter 3, “Data Tables, Reports, and Scripts,” describes details of the data table, but for now let’s try an analysis.

## Launch an Analysis Platform

What are the distributions of the weight and age columns in the table? That is, how many of each weight value and how many of each age value are there in the Big Class table?

- ~ Select **Analyze > Distribution**.

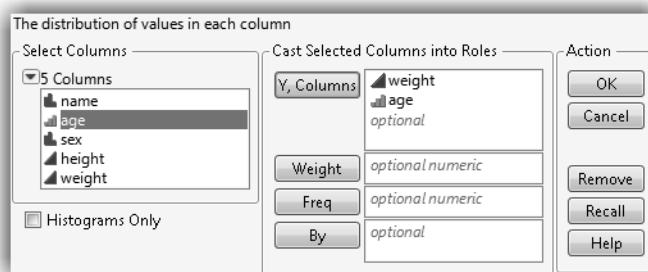
This is called *launching* the Distribution platform. The launch window appears, and prompts you to select variables to analyze.

- ~ Click on weight to highlight it in the variable list on the left of the window.
- ~ Click **Y, Columns** to add weight to the list of variables on the right of the window. These are the variables to be analyzed.
- ~ Similarly, select the age variable and click **Y, Columns**.
- ~ Click **OK**.

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

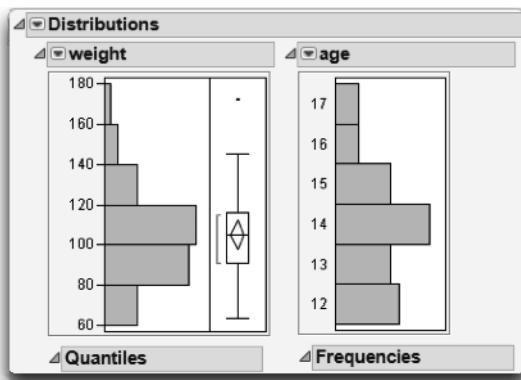
You should now see the completed launch window shown in **Figure 2.6**.

**Figure 2.6** Distribution Platform Launch Window



- ~ Click **OK** to close the window and perform the analysis.

The resulting analysis window shows the distribution of the two variables, weight and age (graphs are shown in **Figure 2.7**).

**Figure 2.7** Histograms for weight and age from the Distribution Platform

## Interact with the Report Surface

All JMP reports start with a basic analysis that you can work with interactively. This lets you dig into a more detailed analysis or customize the presentation. The report is a live object, not a dead transcript of calculations.

### Highlight Rows

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

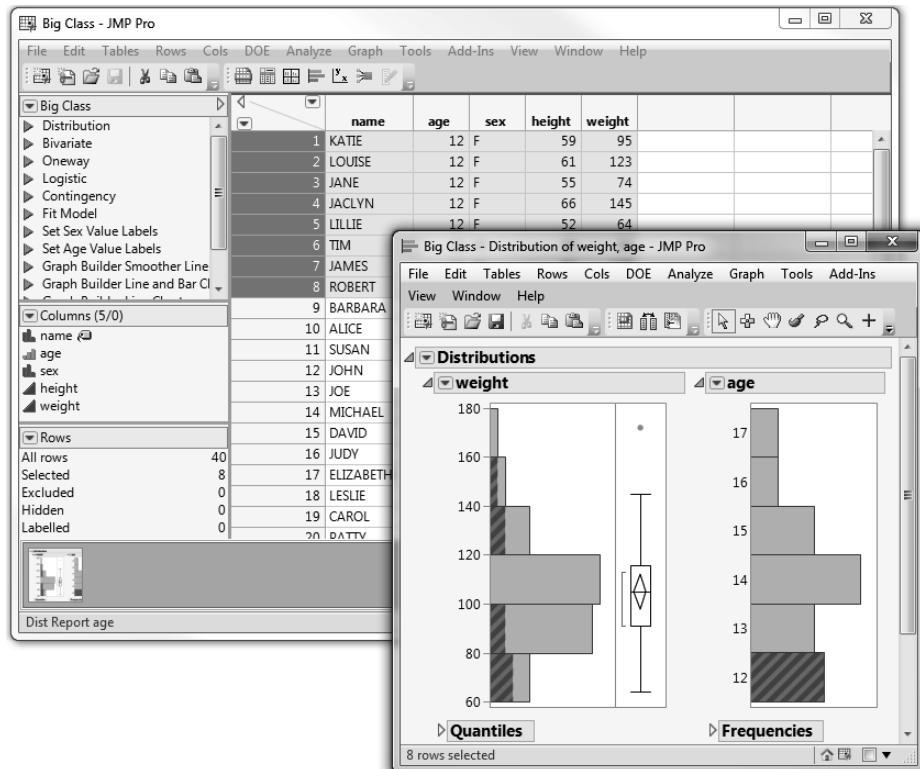
The bar is highlighted, along with portions of the bars in the other histogram and rows in the data table that correspond to the highlighted histogram bar, as shown in **Figure 2.8**. This is the dynamic linking of rows in the data tables to plots. Later, you see other ways of selecting and working with row attributes in a table.

**Note:** You might need to resize and move windows around to see both data tables and analyses at the same time.

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

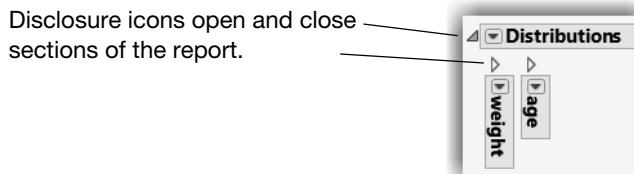
- ☞ Click on the point in the plot. The point highlights, and the corresponding row is highlighted in the data table.
- ☞ Move the mouse over that point to see the label, LAWRENCE, appear.

The point for Lawrence is away from the other weight points and is sometimes referred to as an “outlier.”

**Figure 2.8** Highlighted Bars and Data Table Rows

### Disclosure Icons

Each report title is part of an analysis presentation outline. Click on the gray triangle (disclosure icon) on the side of each report title to alternately open and close the contents of that outline level.

**Figure 2.9** Disclosure Icons

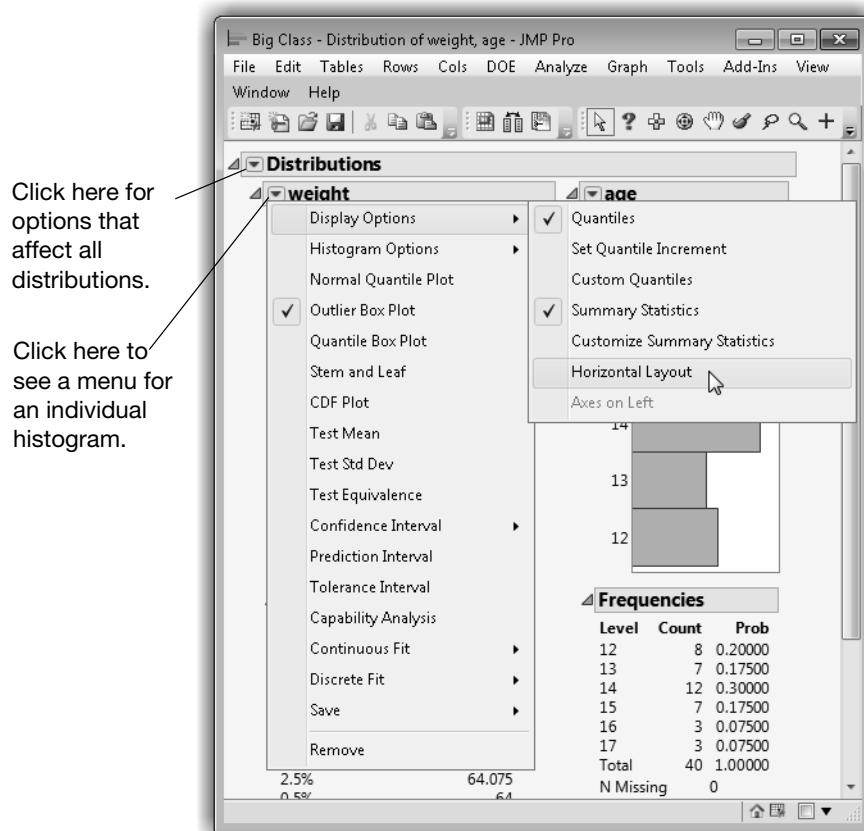
### Contextual Menus

When there are presentation options, the small red triangle to the left of the title on a title bar gives you access to menu commands and for that part of the analysis report (and enables you to remove or keep options). This red triangle menu has

commands specific to the platform. The red triangles on the title bars of each histogram contain commands that influence only the histogram and the corresponding analysis. For example, you can change the orientation of the graphs in the Distribution platform by selecting or deselecting **Display Options > Horizontal Layout** (**Figure 2.10**).

- ☞ Click the red triangle next to weight and select **Display Options > Horizontal Layout**.

**Figure 2.10** Display Options Menus



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

## Menus and Toolbars (Windows)

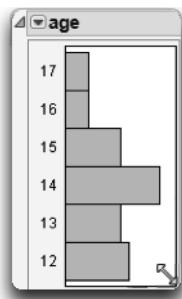
**Important:** In Windows environments, all windows have a JMP menu and toolbar. These might be hidden depending on the size of the window. To view a hidden menu, click Alt; or move your mouse above the gray space that is over the window's title bar, as illustrated here.



Press the Alt key, or move your mouse here to see the menu.

## Resizing Graphs

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



## Special Tools

When you need to do something special, select a tool in the Tools menu and click or drag inside the analysis. See the note above for displaying hidden menus.

The grabber (⌘) is for grabbing objects.

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

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

- ⇨ Select the brush and drag the histogram. To change the size of the rectangle, press Option and drag (Macintosh) or press Alt and drag (Windows).

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

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

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

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

The question mark () is for getting help on the report or graph.

- ~ Select the question mark tool and click on different areas in the Distribution report.

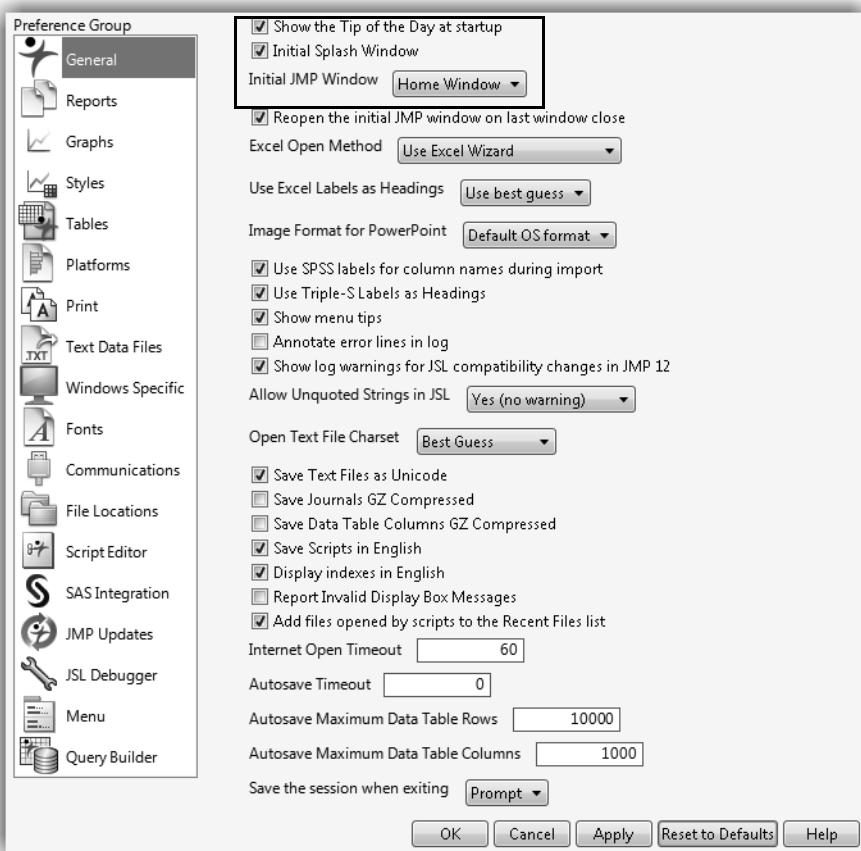
The selection tool () is for selecting an area to copy so that you can paste its contents into another application. Hold down the Shift key to select multiple report sections. Refer to Chapter 3, “Data Tables, Reports, and Scripts,” for details.

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

## Customize JMP

Want to have larger markers, different colors, or other graphs or statistical output each time you use JMP? You can customize your JMP experience using the **Preferences** command on the **File** menu (on the **JMP** menu on Macintosh).

You can set general preferences (shown in **Figure 2.11**) to change what you see when you open JMP each time. When you become more familiar with JMP operations, you might want to change other preferences, which are grouped under **Preference Group**.

**Figure 2.11** Default JMP Preferences**Notes:**

- You can change the look and feel of JMP reports using the preferences in **Styles** and **Graphs**. For this book, we have turned off the **Styles** options **Shade Table Headings** and **Table Headings Column Borders**.
- For each analysis, default graphical and statistical output are displayed. For a particular analysis, you can change these settings using red triangle options or by right-clicking on the report. To change the default settings, select the **Platforms** preference group, select the desired platform from the list, and select the options you'd like to change. For example, the default layout for Distribution reports is vertical, a setting that you can change using the **Horizontal Layout** option. In this book, we sometimes change the default layout for histograms and other output using red triangle options.

## Modeling Type

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

Here is a brief description of the three main modeling types:

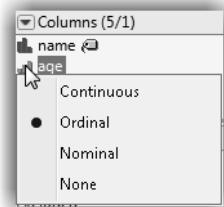
- *Continuous* (■) are numeric values used directly in an analysis.
- *Ordinal* (■) values are category labels, but their order is meaningful.
- *Nominal* (■) values are treated as unordered, categorical names of levels.

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

You can change the modeling type using the Columns panel at the left of the data grid. Notice the ■ beside the column heading for age. This icon is on a pop-up menu.

- ☞ Click the ■ to open the menu for choosing the modeling type for a column.

The different modeling types tell JMP ahead of time how you want the column treated so that you don't have to say it again every time you do another analysis. Modeling types also help reduce the number of JMP commands that you need to learn. Instead of two distribution platforms, one for continuous variables and a different one for categorical variables, a single command performs the anticipated analysis based on the modeling type that you assigned.



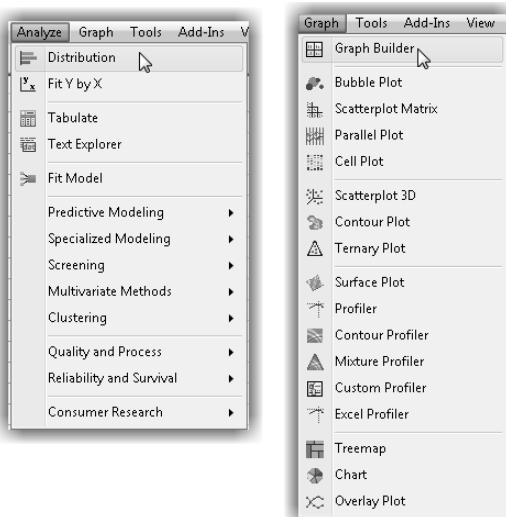
You can change the modeling type whenever you want the variable treated differently. For example, if you want to find the mean of age instead of categorical frequency counts of each age, simply change the modeling type from ordinal to continuous and repeat the analysis. You can change the modeling type in the data table as illustrated above, or in the launch window of the platform that you are using.

**Note:** The None modeling type should be selected for columns that are not used in the analysis. Other modeling types for a column (which you can select from the **Cols > Column Info** window) are Multiple Response, Unstructured Text, and Vector.

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

## Analyze and Graph

Commands in the **Analyze** and **Graph** menus, shown here, launch interactive platforms to analyze data. The **Analyze** menu is for statistics and data analysis. The **Graph** menu is for specialized plots. That distinction, however, doesn't prevent analysis platforms from being full of graphs, nor the graph platforms from computing statistics. Each platform provides a context for sets of related statistical methods and graphs. It won't take long to learn which platforms you want to use for your data.



The previous example used the Distribution platform to illustrate some of the features in JMP.

Select **Help > Books > Menu Card** for a brief description of each menu command. Select **Help > JMP Help** and refer to the *Using JMP* book for detailed documentation and examples.

## Navigating Platforms and Building Context

The first few times that you use JMP, you might have navigational questions: How do I get a particular graph? How do I produce a histogram? How do I get a *t*-test?

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

There are three keys for establishing the context:

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

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

## Contexts for a Histogram

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

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

## Contexts for the *t*-Test

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

If you want the *t*-test to test a single variable's mean against a hypothesized value, you are focusing on a univariate distribution. In this case, you would launch the Distribution platform (**Analyze > Distribution**). The Distribution red triangle menu provides the **Test Mean** command. This command gives you a *t*-test, as well as the option to conduct a *z*-test or a nonparametric test.

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

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

## Contexts for a Scatterplot

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

You might also consider the **Graph Builder** command in the **Graph** menu when you want a presentation graph. As a **Graph** menu platform, it provides only a handful of statistical options, but is interactive and flexible. For example, it can overlay multiple Ys in the same graph and support two *y*-axes.

If you have a whole series of scatterplots for many variables in mind, your context is many bivariate associations. These scatterplots are available from the **Graph** menu using **Scatterplot Matrix** or **Scatterplot 3D**. Scatterplot matrices, along with many options for exploring and analyzing bivariate associations, are available in the Multivariate platform from the **Analyze > Multivariate Methods** menu.

## Contexts for Nonparametric Statistics

There is not a separate platform for nonparametric statistics. However, there are many standard nonparametric statistics in JMP, positioned by context. When you test a mean in the Distribution platform, there is an option to do a (nonparametric) Wilcoxon signed-rank test. When you do a *t*-test or one-way ANOVA in the Fit Y by X platform, you also have optional nonparametric tests, including the Wilcoxon rank sum. (Wilcoxon rank sum is equivalent to the Mann-Whitney *U*-test). If you want a nonparametric measure of association, like Kendall’s  $\tau$  or Spearman’s correlation, look in the Multivariate platform from the **Analyze > Multivariate Methods** menu.

# The Personality of JMP

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

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

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

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

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

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

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



# 3 Data Tables, Reports, and Scripts

## Overview

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

People often ask about the largest data table that JMP can handle. In general, JMP can handle a data table that is about half the size of the available memory in the machine. For example, if you have 8 GB of available memory, JMP can manage a data table that is (approximately) 4 GB.

Commands in the **File**, **Edit**, **Tables**, **Rows**, and **Cols** menus give you a broad range of options for data handling and file management, such as data entry and import, data cleanup, data table manipulation and restructuring, creating numeric summaries, and running JSL (JMP Scripting Language) scripts.

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

**Note:** To open the sample data tables used in the examples, use **Help > Sample Data Library**. You can also select **Help > Sample Data** and click either **Open the Sample Data Directory** or **See an Alphabetical List of all Sample Data Files**.

## Chapter Contents

Overview .....	27
The Ins and Outs of a JMP Data Table.....	29
Selecting and Deselecting Rows and Columns.....	30
Mousing around a Data Table: Cursor Forms.....	30
Creating a New JMP Table .....	32
Define Rows and Columns .....	33
Enter Data.....	35
The New Column Command .....	36
Plot the Data.....	37
Importing Data .....	39
Importing Text Files .....	41
Importing Other File Types.....	44
Copy, Paste, and Drag Data.....	46
Moving Data Out of JMP.....	47
Saving Graphs and Reports .....	48
Copy and Paste .....	48
Drag Report Elements .....	49
Save JMP Reports and Graphs .....	49
Create Interactive Web Reports.....	49
Pop-up Menu Commands .....	50
Juggling Data Tables .....	51
Data Management.....	51
Give New Shape to a Table: Stack Columns .....	52
Creating Summary Statistics.....	55
Create Summary Statistics with the Summary Command .....	55
Create Summary Statistics with Tabulate .....	58
Working with Scripts.....	60
Creating Scripts .....	60
Running Data Table Scripts .....	60
Opening and Running Stand-alone Scripts.....	61

## The Ins and Outs of a JMP Data Table

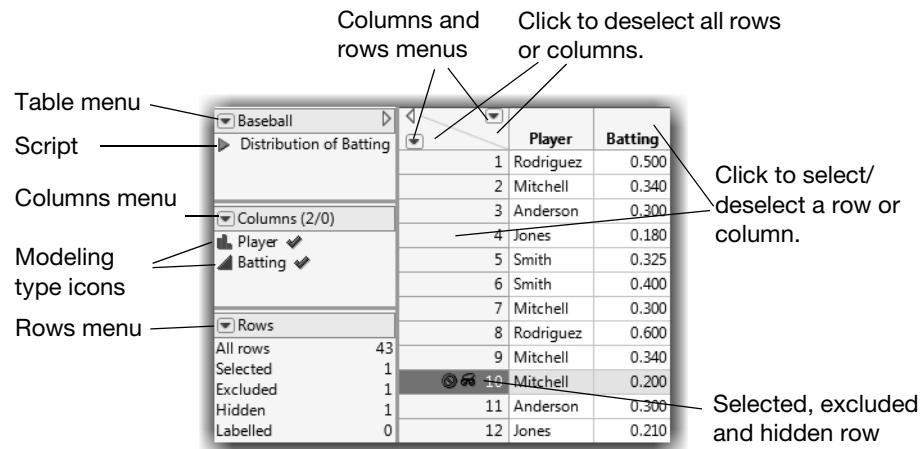
JMP displays data as a data grid, often called a data table. From the data table, you can do table management tasks such as editing cells; creating, rearranging, or deleting rows and columns; subsetting the data; sorting; and combining tables.

**Figure 3.1** identifies active areas of a JMP data table.

There are a few basic things to keep in mind:

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

**Figure 3.1** Active Areas of a JMP Spreadsheet



## Selecting and Deselecting Rows and Columns

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

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

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

To extend a selection of rows or columns, drag across the range of rows or columns (in the selection area). You can also hold down Shift and click the first and last row or column of the range. Hold down Ctrl and click (⌘-click on the Macintosh) to make a non-contiguous selection. To select both rows and columns at the same time, drag across table cells in the data grid.

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

## Mousing around a Data Table: Cursor Forms

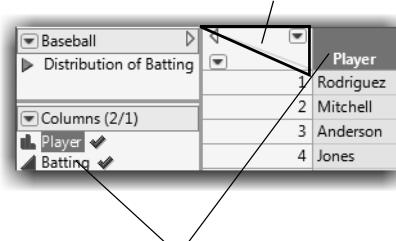
To navigate in the data table, you need to understand how the cursor works in each part of the spreadsheet.

- ☞ To experiment with the different cursor forms, select **Help > Sample Data Library** and open Baseball.jmp.
- ☞ Move the mouse around on the surface areas, and see how the cursor changes form.

### Arrow cursor (»)

When a data table is the active window, the cursor is a standard arrow except when it is on a red triangle menu icon, a gray disclosure icon, or a modeling type icon (to the left of a column name in the Columns panel). It is also a standard arrow when it is in the upper left corner of the data grid where it is used to deselect rows and columns.

Click to deselect all columns.



Click to select columns. Type to edit column name.

**I-beam cursor (I)**

The cursor is an I-beam when it is over selected text in the data grid, column names or in the Columns panel. It signals that the selected text is editable. Double-click and start typing to replace the existing text.

Player
1 Rodriguez
2 Mitchell
3 Anderson
4 Jones

**Selection cursor (+)**

Move the cursor around. The cursor becomes a large thick plus sign when you move it into a column or row selection area. It is used to select items in JMP data tables and reports.

Use the selection cursor to select a single row or column. Hold down Shift and click a beginning and ending row (or a beginning and ending column) to select an entire range. Hold down Ctrl and click (⌘-click on the Macintosh) to select multiple rows or columns that are not contiguous.

Player	Batting
1 Rodriguez	0.500
2 Mitchell	0.340
3 Anderson	0.300
4 Jones	0.180

Click and drag to select.

The selection cursor appears in a report window when you select it from the tools menu. It is used to select areas of reports to copy and paste to other locations. See “Copy, Paste, and Drag Data” on page 46 for details about using the selection cursor for cut and paste operations.

**Double Arrow cursor (↔)**

The cursor changes to a double arrow when placed on a column or row boundary. To change the width of a data table column, click and drag this cursor left or right. To change the height of a data table row, drag this cursor up or down.

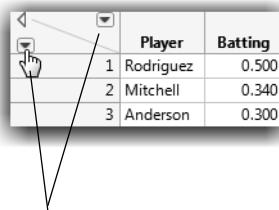
**List Check cursor (‡)**

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

### Pointer cursor (☞)

The cursor changes to a finger pointer over any red triangle menu icon or gray disclosure icon. That signifies you are over a clickable item. Click a disclosure icon to open or close a window panel or report outline; click red triangle icons to open menus.

- ☞ After you finish exploring, select **File > Close** to close the Baseball.jmp data table.



	Player	Batting
1	Rodriguez	0.500
2	Mitchell	0.340
3	Anderson	0.300

Click to see the menu.  
There are many tips that help explain what you see.

## Creating a New JMP Table

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

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

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

<i>Blood Pressure Study</i>				
	Month	Control	Placebo	300mg
	<i>March</i>	165	163	166
	<i>April</i>	162	159	165
	<i>May</i>	164	158	161
	<i>June</i>	162	161	158
	<i>July</i>	166	158	160
	<i>August</i>	163	158	157
				150

## Define Rows and Columns

JMP data tables have rows and columns, which represent *observations* and *variables* in statistical terms. In JMP, the rows always represent observations, and the columns always represent variables. A cell in the data table grid is defined by the row and column that it is in. The raw data in **Figure 3.2** are arranged as five columns (month and four treatment groups) and six rows (months March through August). The first line in the notebook describes each column of values. These descriptions can be used as column names in a JMP data table. To enter this data into JMP, you first need a blank data table.

- ☛ Select **File > New > Data Table** (or **File > New > New Data Table** on the Macintosh) to create a new empty data table, with one column and no rows.

### Add Columns

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

- ☛ Select **Cols > New Columns** and type 5 in the **Number of columns to add** box, and click **OK**.

The default column names are Column 1, Column 2, and so on. You can change them by entering the column names that you want at the top of the columns in the new table.

To edit a column name, first click the column selection area and begin entering the name of the column. Press the Enter or Return key when you are finished, and repeat for the other columns. Or, click Tab to move to the column header for the next column.

- ☛ Enter the names from the data journal in **Figure 3.2** (Month, Control, Placebo, 300 mg, and 450 mg) into the column headers of the new table.

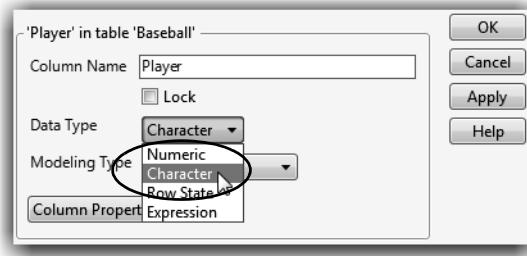
Highlight the column and then begin typing

### Set Column Characteristics

Columns can have different characteristics, such as modeling type and data type. By default, the modeling types are continuous, so the columns expect numeric data. However, in this example, the Month column holds a non numeric character variable.

- ✓ Right-click the column name area for Month and select **Column Info** to see the Column Info window in **Figure 3.3**.
- ✓ In the Column Info window, use the Data Type menu to change Month to a character variable (**Figure 3.3**), and then click **OK**. **Note:** If you don't change the data type, and enter non numeric data into a new column, the data type automatically changes.

**Figure 3.3** Column Info Window



### Add Rows

Adding new rows is easy.

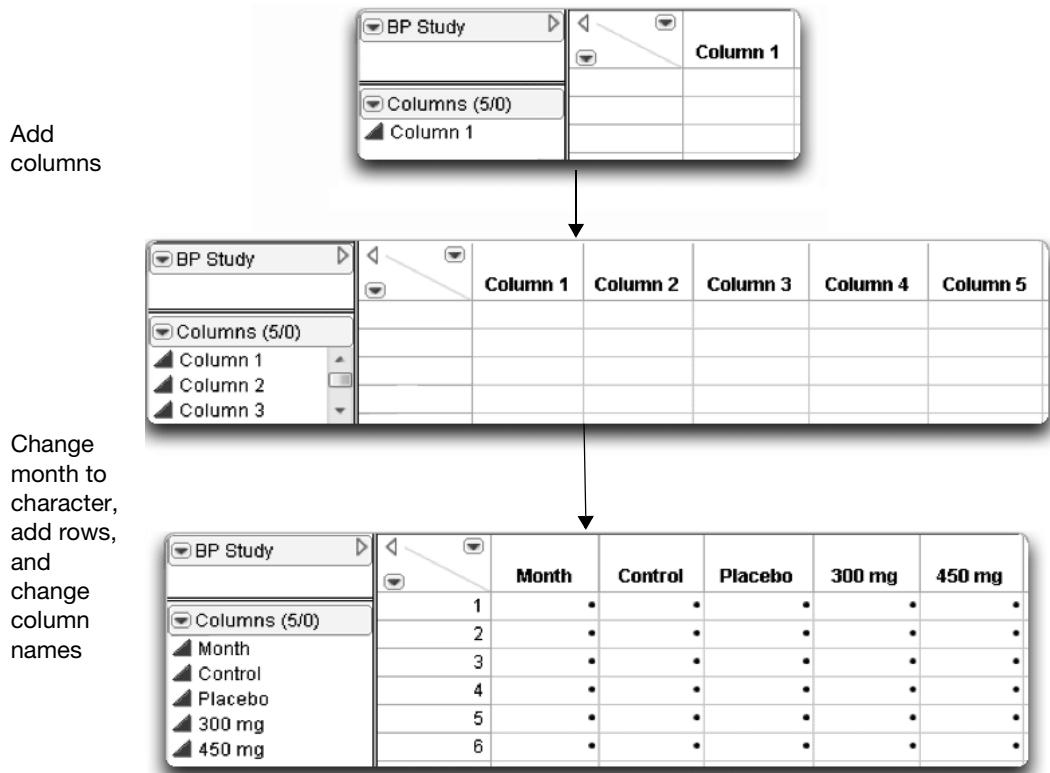
- ✓ Select **Rows > Add Rows** and ask for six new rows.

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

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

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

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

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

## Enter Data

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

- Click on a cell and start entering the appropriate data from the notebook (**Figure 3.2**).

The Tab and Return keys are useful keyboard tools for data entry:

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

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

**Figure 3.5** Finished Blood Pressure Study Table

The screenshot shows a data table window titled "BP Study". The table has columns for Month, Control, Placebo, 300 mg, and 450 mg. The data is as follows:

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

The left sidebar shows the structure of the table: "Columns (5/0)" with entries for Month, Control, Placebo, 300 mg, and 450 mg; and "Rows" with "All rows" and a count of 6.

## The New Column Command

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

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

- ❖ Begin by selecting **Cols > New Columns**, which displays a New Column window like the one shown previously in **Figure 3.3**.

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

- ❖ Enter a new name, Location, in the **Column Name** area.
- ❖ Because the actual names of the locations are characters, select **Character** from the **Data Type** menu.

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

When you click **OK**, the new column appears in the table. Enter the data “Accurate Readings Inc.” for March and April and “Most Reliable Measurements Ltd.” for the other months.

**Note:** You can also add a new column by double-clicking in the column header next to your last column.

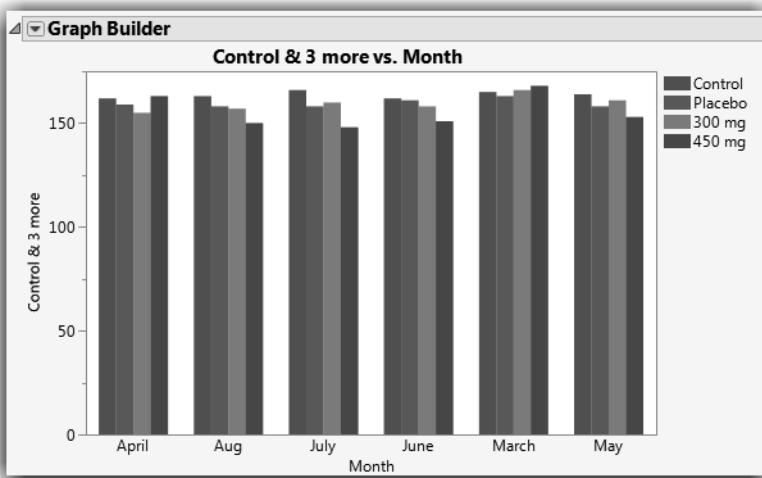
## Plot the Data

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

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

- ☞ Select **Graph > Graph Builder**.
- ☞ Drag Month to the X zone.
- ☞ Select Control, Placebo, 300 mg, and 450 mg and drag them to the Y zone.
- ☞ Click the bar chart icon above the graph.
- ☞ Click **Done** to see the graph in **Figure 3.6**.

**Figure 3.6** Initial Bar Chart



Notice that there's a problem with the values on the *x*-axis. JMP plots data in alphanumeric order. To change the ordering of labels on a graph, we can assign a column property.

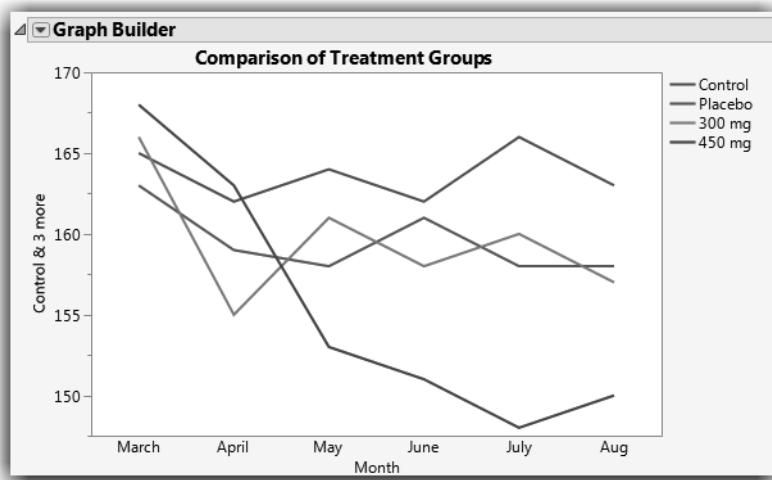
- ☞ Return to the data table, right-click the Month column name, and select **Column Info**.

- ✓ Select the **Column Properties** menu to view the available properties that can be assigned to the column.
- ✓ Select **Value Ordering** and click the **Move Up** and **Move Down** buttons to reorder the values.
- ✓ Click **OK**.
- ✓ Regenerate the graph and confirm that the months are in the correct order.

Now, let's explore the data using a different graph. Because the data are time ordered, we can graph the data using a line chart.

- ✓ From the bar chart you just created, click the line chart icon above the graph to see the graph shown in **Figure 3.7**.
- ✓ Double-click the title at the top, type Comparison of Treatment Groups, and then click **Done**.

**Figure 3.7** Line Chart with New Title



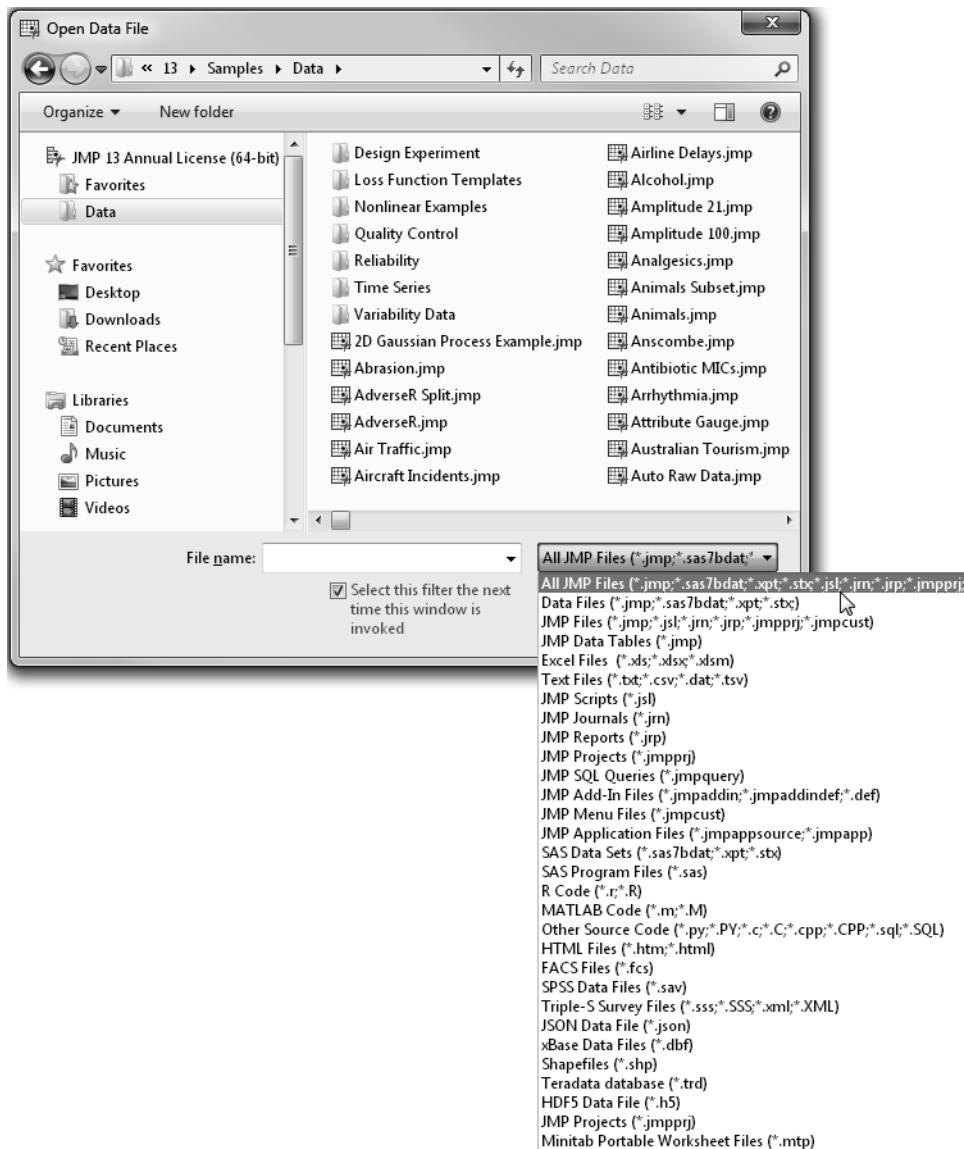
## Importing Data

The **File > Open** command enables you to locate the file that you want to open and then read the file into a JMP data table or as simple text in a JMP script window.

JMP directly reads its own data tables, journals, scripts, projects, reports, add-in files, menu files, application files, and SAS transport files. In addition, JMP can read and write SAS data sets and read SAS program files.

Besides JMP and SAS files, JMP can also read text files with any column delimiter, and can import Microsoft Excel files, R code files, SPSS data files, Minitab files, shape files for maps, and many more. JMP can also connect to a variety of databases and can write SQL queries for importing data. To open database files, you must have an appropriate Open Database Connectivity (ODBC) driver installed on your system.

The example in **Figure 3.8** shows a Windows Open Data File window with all supported file types showing.

**Figure 3.8** Using the Open Data File Window to Import a File

If the incoming file is not a JMP data table, then JMP determines the file type by the extension appended to its file name and opens it accordingly. This works as long as the file has the structure indicated by its name.

## Importing Text Files

In the Open Data File window on Windows, you can first make a selection from the file type menu. In **Figure 3.9**, the file type selection is Text Files (\*.txt, \*.csv, \*.dat, \*.tsv) so that only files with those suffixes show in the list of files above. In this example, Animals.txt is selected to be opened.

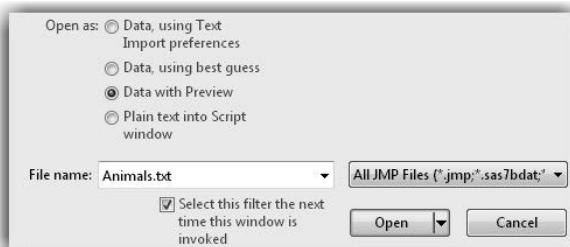
The lower area of the window has the following options:

- You can set text import preferences in your JMP Preferences file. The Data, using Text Import preferences option accesses the JMP preference file and uses those settings.
- The Data, using best guess looks at the data and attempts to determine the best way to present them in a JMP table. This is adequate for rectangular text files with no missing fields, a consistent field delimiter, and an end-of-line delimiter

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

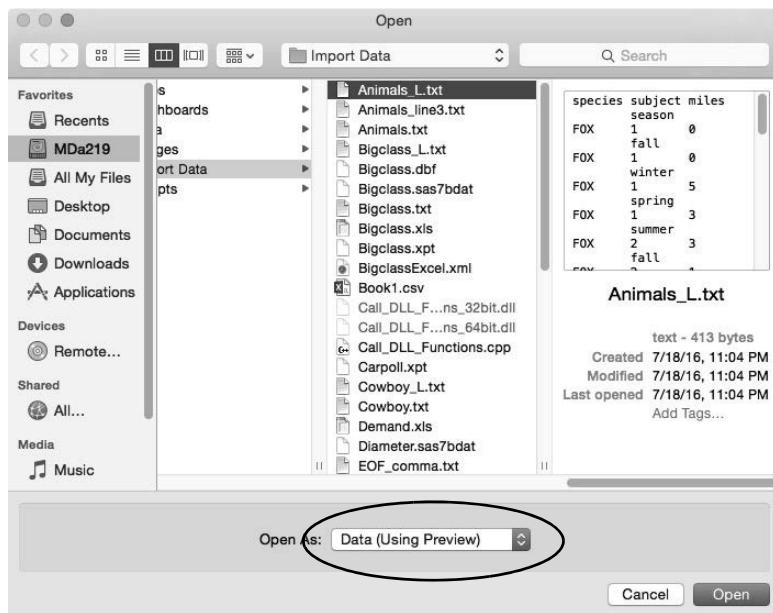
- Data with preview uses the best-guess approach but shows you a sample of the constructed JMP table rows and columns.
- Plain text into Script window writes the text file as it is into a script window without attempting to define fields or columns.

**Figure 3.9** Options for Text Files as Files of Type Selection



On the Macintosh, the import options don't appear until you select a file for the list of readable files. **Figure 3.10** shows an example of the Macintosh Open window with a text file selected. The options are in a menu at the bottom of the window. The options are the same as those described for Windows, but appear in a different order and are worded differently.

**Figure 3.10** Macintosh Import Window with Text Import Options



### Text Import with Preview

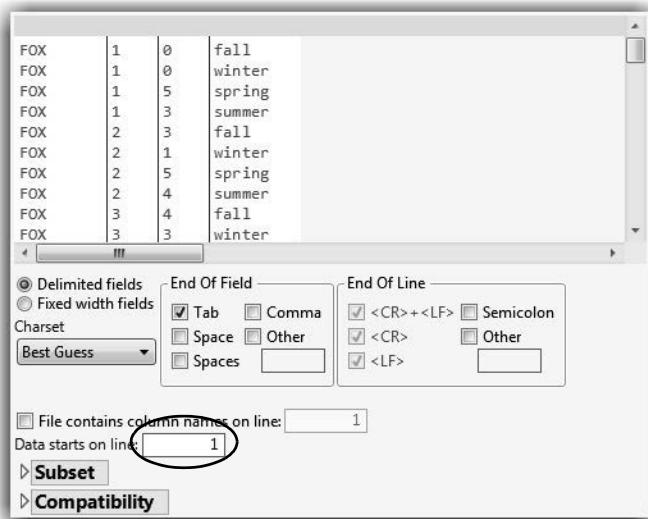
The most general and powerful text import option is **Data with Preview** (Windows) or **Data (Using Preview)** on the Macintosh. When you select this option, a preview window opens and is automatically filled in with settings from your Preferences file and shows several lines of data as shown in **Figure 3.11**.

You can identify one or more end-of-field delimiters, end-of-line delimiters, select the option to Strip enclosing quotation marks, and specify on which row to begin reading data.

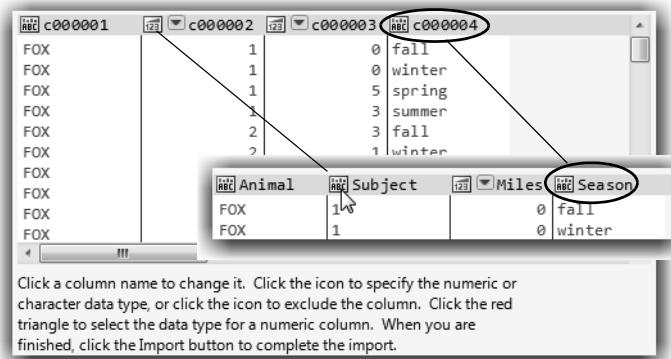
If your data has fixed-width fields, click the Fixed width fields button. This alters the window so that you can specify the widths of the fields in the input data set.

There is a check box to indicate whether the first line contains column names. You can select this option in the preferences or on the window. Then, the first line becomes column names when the data set is imported, and data are read starting on column 2, or whatever column you specify.

**Figure 3.11** Open Text Data With Preview



The example in **Figure 3.11** indicates that there are no column names in the first row of data. Clicking **Next** displays the window in **Figure 3.12** to further modify the input options. The default column names at the top of the columns are called C000001, C000002, and so on. Click and type in the name area to change them. If needed, click on the data type icon and select numeric, character, or row state.

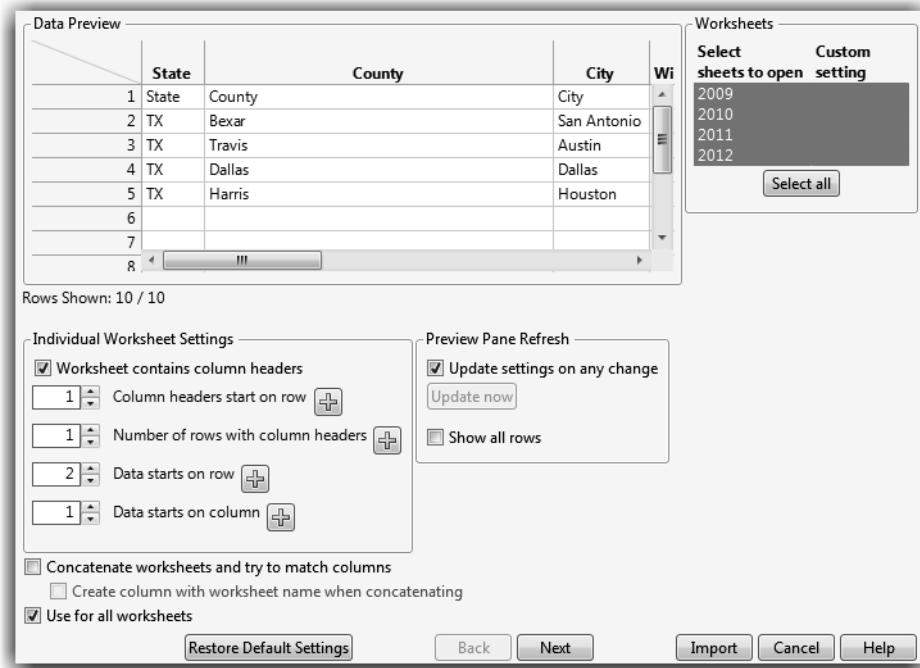
**Figure 3.12** Name Columns Using Open Text with Preview Windows

## Importing Other File Types

### Excel Spreadsheets

JMP can directly import Microsoft Excel worksheets and workbooks using the Excel Import Wizard.

When you open an Excel file in JMP, a preview of the file is displayed in the Excel Import Wizard. The wizard enables you to select import options and to open selected worksheets in separate data tables or to combine the worksheets into one data table.

**Figure 3.13** Excel Wizard

You can also open selected worksheets directly using the menu list on the Open Data File window.

**Note:** The JMP Excel Add-In is added to the Excel menu during the installation of JMP on Windows. The Add-In provides a number of options for working with JMP from Excel. For example, from an Excel worksheet, you can transfer selected cells into a new JMP data table and launch a graphing or analysis platform.



### Databases

JMP can connect to any database that has a corresponding ODBC driver on your system. Select **File > Database > Open Table** to import data from relational databases.

In addition, you can build SQL queries to select and import data from an SQL database using Query Builder. This enables you to preview data, build custom data filters, and share queries with others without writing SQL statements. Select **File > Database > Query Builder** to select a database connection and launch Query

Builder. Select **Help > JMP Help** and refer to the *Using JMP* book for details about connecting to databases.

## Copy, Paste, and Drag Data

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

### Copy

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

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

**Note:** If you use **Copy With Column Names**, the first line of information copied to the clipboard is the JMP column names.

### Paste

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

To transfer data from another application into a JMP data table, first copy the data to the clipboard from within the other application. Then use the **Paste** command to copy the values to a JMP data table. Rows and columns are automatically created as needed. If you select **Paste With Column Names**, the first line of information about the clipboard is used as column names in the new JMP data table.

To duplicate an entire row or column:

1. Select a row or column to be duplicated and select **Edit > Copy**.
2. Select an existing row or column to receive the values.

3. Select **Edit > Paste** to transfer the values.

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

### Drag

You can also move or duplicate rows and columns by dragging. Hold down the mouse in the selection area of one or more selected rows or columns and drag them to a new position in the data table. Hold down Ctrl and drag to duplicate rows and columns instead of moving them.

## Moving Data Out of JMP

Two questions that come up as you start using JMP might be “Can I get my data back out of JMP?” and “How do I get results out of JMP?”

The **Save As** command saves the active data table to a file after prompting you for a name and file type. JMP can save data in any of the following formats:

- **JMP Data Tables** saves the table in JMP format. This is the default **Save As** option.
- **Excel Workbook** saves data tables in Microsoft Excel .xlsx or .xls format. The resulting file is directly readable by most versions of Excel. You can also save data as an Excel workbook by selecting **View > Create Excel Workbook**.
- **Text Export Files** converts data from a JMP file to a standard text format, with rows and columns.

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

**Export Column Names to Text File** has an **Export Table Headers** check box to request that JMP column names be written as the first record of the text file.



**End of Field** and **End of Line** designate the characters to identify the end of each field and end of line in the saved text file. These options are described previously in the section “Importing Data” on page 39.

- **SAS Data Set** saves the data as a SAS 7 data set (.sas7bdat), readable by SAS 7 or later.
- **SAS Transport Files** converts a JMP data table to SAS transport file format and saves it in a SAS transport library. The **Append To** option appends the data table to an existing SAS transport library. If you don’t use **Append To**, a new SAS transport library is created using the name and location that you provide. If you do not specify a new filename, the SAS transport library replaces the existing JMP data table.
- Select **Database > Save Table** to save data in database formats that have ODBC drivers installed on your system.
- **JSON Data File** saves data in a JSON-formatted text file.

On the Macintosh, to save data as a JMP data table, select **File > Save As**. To save data as text, Excel, SAS or SAS Transport File, or JSON, select **File > Export**, and then select the appropriate format from the window.

## Saving Graphs and Reports

You can use standard copy and paste operations to move graphical displays and statistical reports from JMP to other applications. You can also drag JMP reports and graphs to any other application that supports drag operations, and can save JMP results in a variety of formats.

### Copy and Paste

When you copy from a report (results) window, the information is stored on the clipboard. If you want to copy part of a report window, use the selection tool () from the **Tools** menu or toolbar. To copy and paste:

- Click on the area that you want to copy, hold down the Shift key and click to extend the selected area. Select the **Copy** command to copy the selected area to the clipboard.
- Select the **Paste** command to paste the results into a JMP journal or another application.

The format used when pasting depends on the application that you paste into. If the application has a **Paste Special** command, you can select among paste formats. Rich text, which includes pictures (RTF), unformatted text (TXT), picture (PICT or WMF), bitmap (BMP), and enhanced picture (EMF) are options. On the Macintosh, PDF is available as a Paste Special option.

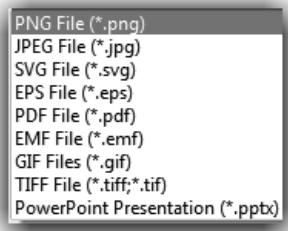
## Drag Report Elements

Any element in a JMP report window that can be selected can be dragged within the same window or into another application. When you drag report elements, they are copied and pasted to the destination area where you drop them.

## Save JMP Reports and Graphs

On Windows, you can save selected output in a variety of graphical formats selecting **Edit > Save Selection As**. Available formats are shown here.

To save the entire report, select **File > Save As**. More formatting options are available, including saving as a Microsoft PowerPoint presentation, HTML and interactive HTML with data, and Microsoft Word.



On the Macintosh, select **File > Export** to save in one of the following formats: Text, PNG, TIFF, SVG, EPS, HTML and interactive HTML with data, RTF, and Microsoft PowerPoint.

## Create Interactive Web Reports

Sometimes you want to organize your statistical output and graphs, and package these results so that you can easily share your findings with others. The **View > Create Web Report** command creates HTML output of your reports and graphics. The option also packages selected output as an interactive web report that can be viewed on most devices and browsers. You can specify the ordering of the results, customize styles and descriptive text, and add a company logo. The resulting web report organizes output as thumbnails linked to individual HTML report elements.

## Pop-up Menu Commands

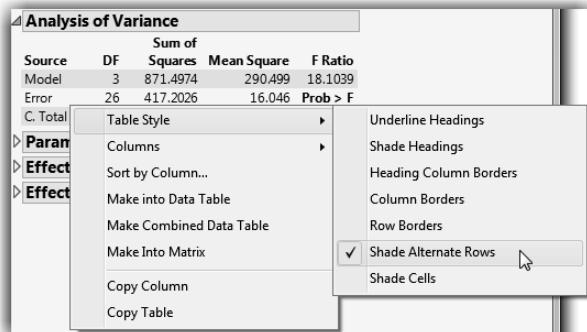
Right-click on a report window to see the pop-up menu used in the following examples. The pop-up menu changes depending on where you click (hence its name, *pop-up menu*). If your cursor is not over a display element with its own pop-up menu, right-clicking shows the menu for the whole platform.

### Context Commands for Report Tables

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

Right-click in the table area to see a variety of options, including the following:

- **Table Style** lets you enhance the appearance of a table by drawing borders or other visual styles to the table rows and columns. You can also set these styles in the Styles preferences under Report Tables.
- **Columns** lets you select which columns you want to show in the analysis table. Analysis tables often have many columns, some of which might be initially hidden. You can turn on many of these options in the Platforms Preferences.
- **Sort by Column** sorts the rows of a report table. This command displays a list of visible columns in a report, and you select one or more columns to sort by.
- **Make into Data Table** creates a JMP data table from any analysis table.
- **Make into Matrix** lets you store a report table as a matrix that is useful when you are using the JMP scripting language (JSL). When the option is selected, the window shown here appears, enabling you to designate the name of the matrix and where it should be stored.



# Juggling Data Tables

The data for analysis that reaches you is rarely neat, clean, and ready to go. Unless you can find someone else to do house-keeping on your data, reorganizing the structure of information is often necessary. Each of the following examples uses commands from the **Tables**, **Rows**, or **Cols** menus to reorganize data.

## Data Management

Suppose you have the following situation. Person A began a data entry task by entering state names in order of ascending auto theft rates. Then, Person B took over the data entry, but mistakenly entered the auto theft rates in alphabetical order by state (paying no attention to the state names that were already there).

☞ To see the result, select **Help > Sample Data Library** and open Automess.jmp.

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

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

Here is one way to solve this problem that uses JMP data management tools.

First, you need sorted Auto theft values associated with the state names as they are in the table. To do this in JMP (without having to reenter the theft values), do the following:

1. Make a copy of the Automess table.
2. Sort the Auto theft variable in the copy in ascending order.
3. Join the sorted result with the original table, keeping only the State variable from the original table and the Auto theft variable from the sorted table.

Follow these steps to correct the Automess table.

☞ With Automess.jmp active, select **Tables > Subset** and click **OK**.

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

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

- ☞ With this subset table active, right-click the Auto theft column and select **Sort > Ascending**.

Note that tables can also be sorted on multiple columns using **Tables > Sort**.

Join the incorrectly sorted Automess.jmp table with the correctly sorted Subset of Automess table.

- ☞ Select **Tables > Join**.
- ☞ When the Join window appears, note which table is listed next to the word Join (either Automess or Subset of Automess). Click the other table's name in the list of tables.
- ☞ By default, the Matching Specification box is set to **By Matching Columns**. Use the Matching Specification menu and change the specification to **By Row Number**.
- ☞ Because you don't want all the columns from both tables in the final result, click the **Select columns for joined table** check box.

The variables from both tables appear in list boxes.

- ☞ Select State from the Automess table and click **Select**.
- ☞ Select Auto theft from the Subset of Automess table and click **Select**.
- ☞ Click **OK** on the Join launch window.

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

## Give New Shape to a Table: Stack Columns

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

- ☞ To see an example of stacking columns, select **Help > Sample Data Library** and open Cheese Taste.jmp to see the table on the left in **Figure 3.14**.

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

- ☞ Select **Tables > Stack**.
- ☞ In the Stack window, select the cheeses (A, B, C, and D) from the **Select Columns** list and click **Stack Columns**. Leave everything else as is.
- ☞ Click **OK** to see the table on the lower right in **Figure 3.14**.

**Figure 3.14** Stack Columns Example

The figure displays two JMP data tables side-by-side. The left table, titled 'Cheese Taste', contains data for four cheese types (A, B, C, D) across nine response levels (1 to 9). The right table, titled 'Cheese Stacked', shows the same data after stacking the columns, resulting in three columns: Response, Label, and Data.

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

	Response	Label	Data
1	1	A	0
2	1	B	6
3	1	C	1
4	1	D	0
5	2	A	0
6	2	B	9
7	2	C	1
8	2	D	0
9	3	A	1
10	3	B	12
11	3	C	6
12	3	D	0

The Label column shows the cheeses, and the Data column is now the count variable for the response categories (1 through 9). Now use JMP to generate a contingency table.

- ✓ Select the Data column header area, and select **Cols > Preselect Role > Freq** to assign the Data column the frequency role for the analysis.

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

To see how response relates to type of cheese:

- ✓ Select **Analyze > Fit Y by X**.
- ✓ In the Fit Y by X launch window, assign Response to **Y, Response** and Label to **X, Factor**.

If you preselected its role, Data is already assigned as a **Freq** variable. If not, assign it on the launch window.

When you click **OK**, the contingency table platform appears with a mosaic plot, Crosstabs table, Tests table, and menu options. To find more information about the platform components, you can use the Help tool (?) in the **Tools** menu and click on the platform surface. A simplified version of the Crosstabs table with only counts is shown in **Figure 3.15**. (Right-click on the contingency table to modify what it displays.) The stacked Cheese data is used again later for further analysis.

**Figure 3.15** Contingency Table for the Cheese Data

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

- ✓ For practice, see whether you can use the **Split** command on the stacked data table to reproduce a copy of the Cheese Taste table.

# Creating Summary Statistics

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

**Summary** creates a JMP window that contains a summary table. This table summarizes columns from the active data table, called its *source table*. It has a single row for each level of a grouping variable that you specify.

A grouping variable divides a data table into groups according to each of its values. For example, a gender variable can be used to group a table into males and females.

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

## Create Summary Statistics with the Summary Command

The example data used to illustrate the **Summary** command is the JMP table called Companies.jmp (see **Figure 3.16**).

☞ Select **Help > Sample Data Library** and open Companies.jmp.

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

**Figure 3.16** JMP Table to Summarize

The screenshot shows the JMP software interface. On the left, there is a 'Companies' data table with columns: Type, Size Co, Sales (\$M), Profits (\$M), # Employ, profit/emp, Assets, and %profit/sales. Below it is a 'Summary' table with the same columns. To the left of the tables, a 'Columns (8/0)' list is visible, containing variables: Type, Size Co, Sales (\$M), Profits (\$M), # Employ, profit/emp, Assets, and %profit/sales. The 'Type' variable is selected in the list.

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

- ☞ Select **Tables > Summary**.
- ☞ In the Summary window, select the variable **Type** in the Columns list and click **Group** to add it to the grouping variables list (shown in **Figure 3.17**).

You can select as many grouping variables as you want. But for now, look at a single variable.

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

The new summary table appears in an active window as shown at the bottom in **Figure 3.17**. This table is linked to its source table. When you highlight rows in the summary table, the corresponding rows are also highlighted in its source table.

**Figure 3.17** Summary Window and Summary Table

The screenshot shows the 'Request Summary Statistics by Grouping Columns' dialog box and a summary table.

**Summary Window (Top):**

- Select Columns:** A list of 8 columns: Type, Size Co, Sales (\$M), Profits (\$M), # Employ, profit/emp, Assets, %profit/sales. The 'Type' column is selected.
- Statistics:** Set to 'optional'. A 'Group' button is highlighted with a red oval.
- Action:** Buttons for OK, Cancel, Remove, Recall, and Help.

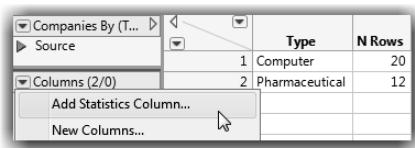
**Summary Table (Bottom):**

	Type	N Rows
1	Computer	20
2	Pharmaceutical	12

The summary table displays frequency counts (N Rows) for each level of the grouping variables. This example shows 20 computer companies and 12 pharmaceutical companies. You could have requested other statistics on the Summary window. The **Statistics** menu in the Summary window lists standard univariate descriptive statistics.

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

- ✓ Select **Add Statistics Column** command from the red triangle menu on the Columns tab at the left of the summary table.



This command displays the Summary window again.

- ☞ Select any numeric column (for example, Profits \$M) from the source table columns list.
- ☞ Select the statistic that you want (for example, **Sum**) from the **Statistics** menu on the window.
- ☞ If desired, repeat to add more statistics to the summary table.
- ☞ Click **OK** to add the columns of statistics to the summary table.

The table in **Figure 3.18** shows the sum of Profits (\$M) in the summary table grouped by Type.

**Figure 3.18** Expanded Summary Table

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

Another way to add summary statistics to a summary table is with the **Subgroup** button in the Summary window (**Figure 3.17**). This method creates a new column in the summary table for each level of the variable that you specify with **Subgroup**. The subgroup variable is usually nested within all the grouping variables.

## Create Summary Statistics with Tabulate

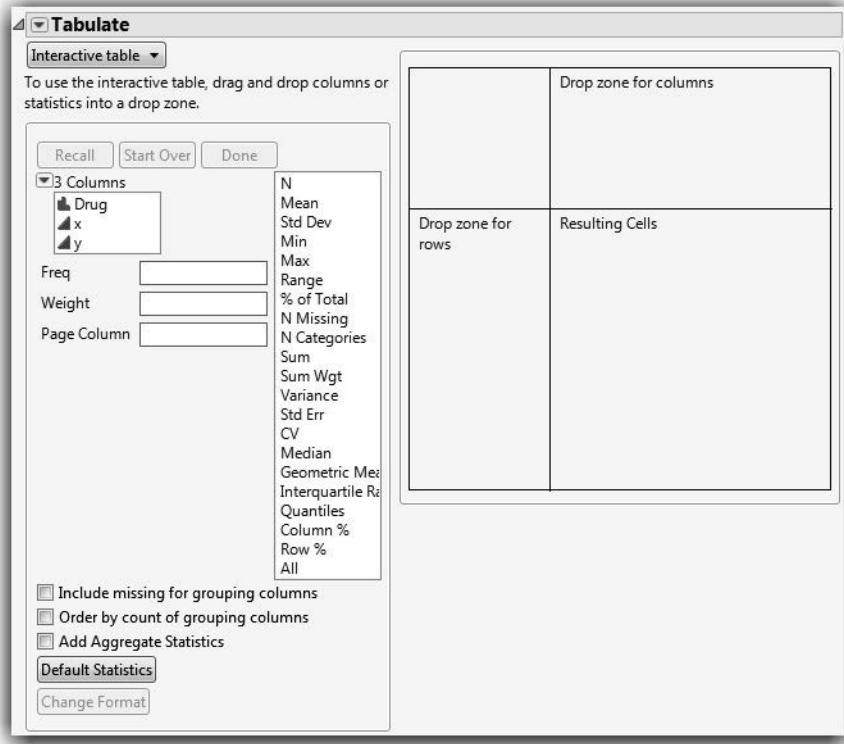
You can use **Analyze > Tabulate** to create the same table interactively.

With the Companies.jmp table active, follow these steps:

- ☞ Select **Analyze > Tabulate** to see the Tabulate window in **Figure 3.19**. You'll see the list of variables in the Companies table in the variables list on the left of the window.
- ☞ First, drag the N from the statistics panel to **Drop zone for columns**. This creates the small table shown here. So far, the table shows the total N of 32 for the whole table.
- ☞ Next drag the Type variable to the **Drop zone for rows** (which is now a small box). Now you should see a row for each type of company, with the appropriate N for each type.

N
32

Type	N
Computer	20
Pharmaceutical	12

**Figure 3.19** Tabulate Window for Dragging and Dropping Variables

Drag Profits (\$M) just to the right of N in the tabulate palette.

You see a new small gray column outlined in blue, which indicates that you want a new column. When you release the mouse, Tabulate creates the table shown in **Figure 3.20**.

Select **Make Into Data Table** from the red triangle menu next to Tabulate to create the summary data table.

**Note:** The Sum statistic is the default. If you want a different statistic, right-click on the statistics name in the Tabulate drop zone and select **Statistics**, then select the one you want. You can also drag a statistic from the list

**Figure 3.20** Add Analysis Column and Create Summary Table

The figure displays two JMP data tables side-by-side. The left table, titled 'Profits (\$M)', has columns 'Type', 'N', and 'Sum'. It contains two rows: 'Computer' with N=20 and Sum=\$48173, and 'Pharmaceutical' with N=12 and Sum=\$8280.9. The right table has columns 'Type', 'N', and 'Sum(Profits (\$M))'. It also contains two rows: 'Computer' with N=20 and Sum=\$48173, and 'Pharmaceutical' with N=12 and Sum=\$8280.9.

Type	N	Sum
Computer	20	48173
Pharmaceutical	12	8280.9

	Type	N	Sum(Profits (\$M))
1	Computer	20	48173
2	Pharmaceutical	12	8280.9

## Working with Scripts

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

Scripts are stored in two formats:

- data table scripts
- stand-alone scripts

### Creating Scripts

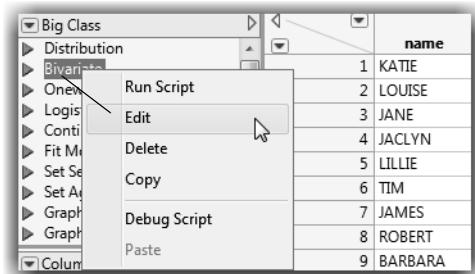
You can write your own scripts, customize existing scripts, or create scripts from any JMP platform.

To create a data table script or stand-alone script from a JMP platform, select **Save Script** from the red triangle menu on a report. This command provides several options for saving a script, such as **To Data Table** and **To Script Window**.

For information about writing JSL code, see the Scripting Index in the **Help** menu. You can also select **Help > JMP Help** and refer to the *Scripting Guide* or *JSL Syntax Reference*.

### Running Data Table Scripts

Data table scripts are listed in the Table panel, as shown here. This example is from the Big Class.jmp sample data table, showing several scripts that have been saved with it.



To work with scripts in a data table, follow these steps:

- Click the green triangle next to the script's name to run the script.
- Right-click the script name and select **Edit** to view or edit the script.

## Opening and Running Stand-alone Scripts

Stand-alone scripts are stored as simple text files with the JSL (.jsl) extension in the JMP Samples/Scripts folder.

To open and run a stand-alone script:

- ❖ Select **File > Open** and navigate to the folder that contains the script that you want.
- ❖ Double-click the script name to open it.

Some scripts are designed to run when they open. Other scripts open in a script editor window. To execute the script in the script editor window:

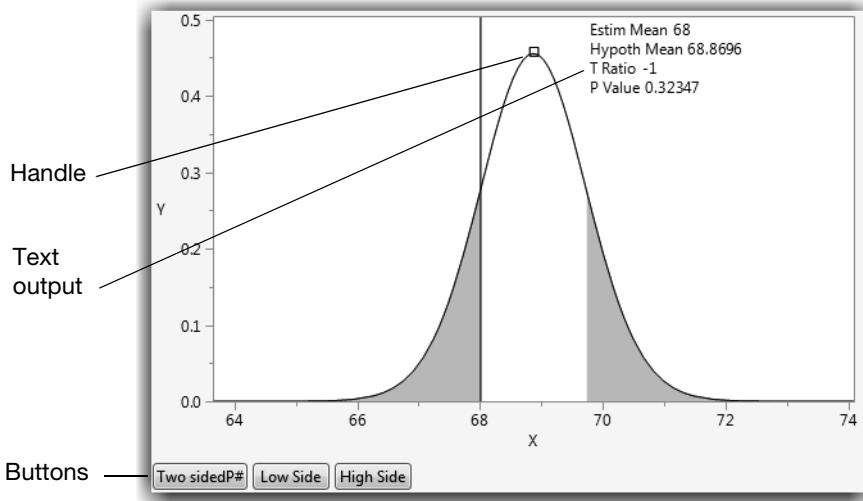
- ❖ Use the shortcut key **Ctrl-R** (Windows) or **⌘-R** (Macintosh), click the **Run Script** icon on the toolbar, or right-click in the script and select **Run Script**.

For example, use the `demoMeanTestPValue` script available in **Help > Sample Data** under the **Teaching Scripts > Teaching Demonstrations** outline.

- ❖ Run the `demoMeanTestPValue` script.

All scripts in the Sample Data window are designed to run automatically.

**Figure 3.21** illustrates several key features of a typical instructional script.

**Figure 3.21** Test an Estimated Mean with a Hypothesized Mean Window

- A *handle* is a script element that you can drag to update the display as it is dragged. In this case, the handle controls the *t*-distribution for hypothesized mean and changes as you drag it, adjusting the *t*-ratio and *p*-value.
- *Text output* is sometimes drawn directly on the graphics screen rather than displayed as reports below the window. This script shows the estimated mean, the hypothesized mean, the *t*-ratio, and the *p*-value, in the upper right corner of the window.
- *Buttons* reveal options, set conditions, or trigger actions in the script. In this example, the buttons are for a two-sided *p*, or the one-sided upper or lower *p*-value.

To practice with these elements:

- ❖ Drag the handle to different places in the window and observe how the hypothesized mean and the *p*-value change.
- ❖ Click the buttons to see how the *p*-values change.

To view the JSL code for demoMeanTestPValue, select **Help > Sample Data > Open the Sample Scripts Directory**, and double-click the script to open it.



# 4

## Formula Editor

### Overview

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

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

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

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

## Chapter Contents

Overview .....	63
The Formula Editor Window .....	65
The Formula Editor and the JMP Scripting Language .....	66
A Quick Example: Standardizing Data .....	67
Making a New Formula Column .....	69
Using Popular Formula Functions.....	71
Writing Conditional Expressions .....	72
Summarizing Data with the Formula Editor .....	77
Generating Random Data .....	82
Local Variables and Table Variables .....	87
Working with Dates.....	89
Tips on Building Formulas .....	90
Examining Expression Values.....	90
Cutting, Dragging, and Pasting Formulas.....	90
Selecting Expressions .....	91
Exercises.....	91

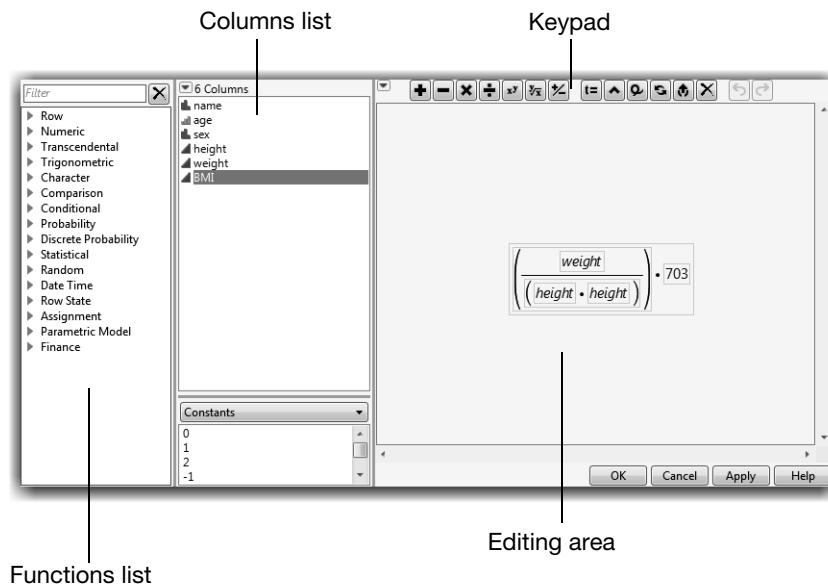
# The Formula Editor Window

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

- Select **Formula** from the **Cols** menu for one or more selected columns.
- Select **Formula** from the **Column Properties** menu in a New Column window and click the **Edit Formula** button.
- Select a column and then select **Cols > Column Info > Formula**.
- Right-click on a column header or on the column in the Columns panel and select **Formula** from the menu. This opens the Formula Editor window without first opening the Column Info window.

**Note:** You can also write formulas from the data table for one or more selected columns without opening the Formula Editor. To do this, right-click on the column header of one of the selected columns, select **New Formula Column**, and select from the list of available functions.

The Formula Editor window is divided into four areas: the Functions list, Columns list, formula elements, and editing area. **Figure 4.1** shows the parts of the Formula Editor, with a formula for the calculation of Body Mass Index (BMI) for the Big Class.jmp sample data. The Formula Editor editing area consists of buttons (**OK**, **Apply**, **Help**), the formula display area, and a keypad that appears above it. The formula display is an editing area that you use to construct and modify formulas.

**Figure 4.1** The Formula Editor Window

## The Formula Editor and the JMP Scripting Language

Whenever you create a formula with the Formula Editor, you are actually creating a JMP script. To see the script, double-click the formula. The script for the BMI formula in **Figure 4.1** is shown here.

$$(:weight / (:height * :height)) * 703$$

A more advanced but valuable example is recoding a continuous variable into discrete bins. For example, suppose you have a column of numeric Height values that you want to recode into three levels called Short, Medium, and Tall. You could create a new column with the formula shown to the left in **Figure 4.2**. If you double-click on the formula, you'll see the JSL statement in the middle. Columns with the original values and the recoded values are on the right in **Figure 4.2**.

**Figure 4.2** Recode Formula and JSL

height	Height Recoded
64	Medium
69	Tall
62	Medium
64	Medium
67	Tall
65	Medium
66	Tall
62	Medium
66	Tall
65	Medium
60	Short
68	Tall
62	Medium
68	Tall
70	Tall

```
If( height <= 61, "Short",
    61 < height <= 65, "Medium",
    height > 65, "Tall",
    Is Missing( height ), "Missing" )
```

## A Quick Example: Standardizing Data

The following example gives you a quick look at the basic features of the Formula Editor. Suppose you want to compute a *standardized* value. That is, for a numeric variable  $x$ , you would compute as follows:

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

where  $\bar{x}$  is mean of  $x_1, x_2, x_3, \dots, x_i$  and

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

for each row in a data table.

☞ For this example, select **Help > Sample Data Library** and open Students.jmp.

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

☞ Begin by selecting **Cols > New Columns**, which displays a New Column window like the one shown in **Figure 4.3**.

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

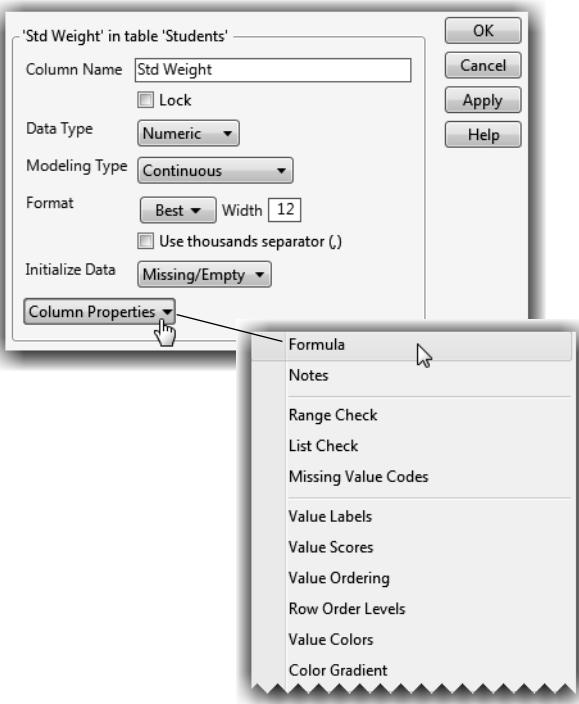
☞ Type the new name, **Std Weight**, in the **Column Name** area.

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

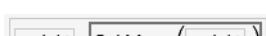
- ⓐ Click **Column Properties** and select **Formula** from menu.

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

**Figure 4.3** The New Column Window



Next, enter the formula that standardizes the weight values by following these steps.

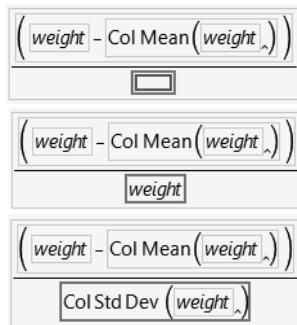
- ⓐ With “no formula” highlighted, select weight in the Columns list. 
- ⓑ Click the minus key on the keypad or on your keyboard. 
- ⓒ With the new entry highlighted, select weight in the Columns list. 
- ⓓ Select **Statistical** in the Functions list and select **Col Mean** from the menu. 
- ⓔ Click anywhere in the white space inside the boxed area to select the entire expression. 

With the entire expression selected, click the divide key on the keypad.

Select weight again from the Columns list.

With weight still highlighted, select **Statistical** in the Functions list and select **Col Std Dev**.

You have now entered your first formula.



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

Click **OK** to close the New Column window.

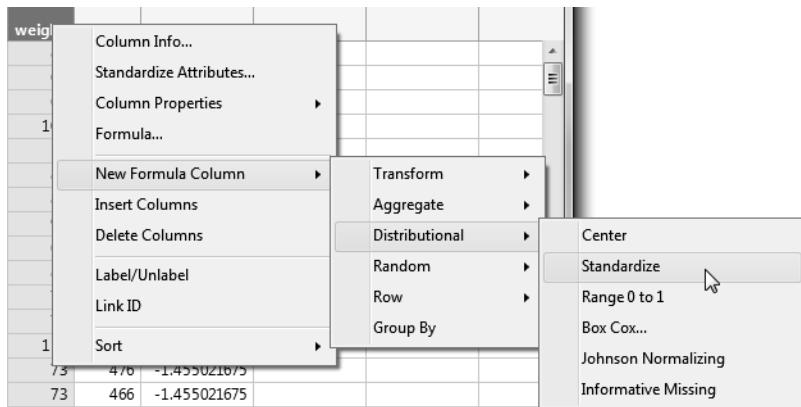
The new column fills with values. If you change any of the weight values, the calculated Std Weight values automatically recompute. If you add new weight values to the data table, the Std Weight values are automatically computed for these new values.

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

This example might be all you need to proceed. However, the rest of the chapter shows how to use the Formula Editor to create formulas with commonly used functions. For complete documentation on the Formula Editor, select **Help > JMP Help** and refer to the *Using JMP* book.

## Making a New Formula Column

Many commonly used formulas can be created using the New Formula Column shortcut. For example, to create the formula column with the standardized values shown earlier, right-click the column header for weight and select **New Formula Column > Distributional > Standardize** (Figure 4.4).

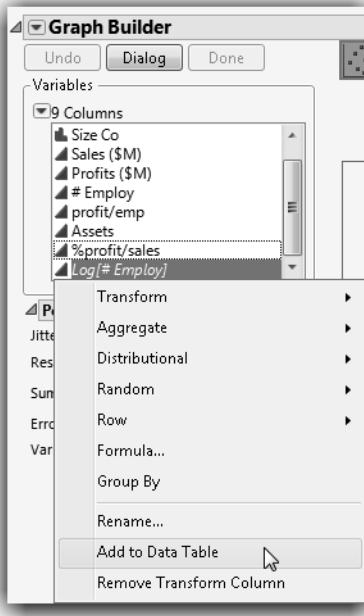
**Figure 4.4** Standardizing Data Using New Formula Column

The New Formula Column functions that are available depend on the number of columns selected and the data types. For numeric data, the function groups listed are Transform, Character, Combine, Aggregate, Distributional, Random, and Row. Date-Time functions are available for columns with a date or time format. For categorical data, the available function groups are Character, Random, and Row.

**Note:** These same options are available within launch windows. To see the available functions, right-click on one or more variables in the launch window, as shown in **Figure 4.5**. Again, the functions that are available depend on the variables selected.

For example, we use the Companies.jmp sample data. The # Employ variable represents the number of employees. We use Graph Builder to create a variable for the log of # Employ called *Log[# Employ]*. Temporary variables are added to the bottom of the list of variables in the launch window (**Figure 4.5**). The italics indicate that *Log[# Employ]* is a temporary variable.

Creating temporary variables enables you to use formula variables in graphs or analyses without adding new formula columns to the data table. If you want to add the formula column to the data table, right-click on the temporary variable in the launch window Variables list and select **Add to Data Table** as shown in **Figure 4.5**.

**Figure 4.5** Creating a Temporary Variable in Graph Builder

## Using Popular Formula Functions

There are many important Formula Editor functions, and we won't even begin to scratch the surface of their capabilities. In this section, we introduce some of the more commonly used functions and see some of the building blocks and logic used when you are building formulas in JMP. We see how to build conditional expressions, summarize data with the Formula Editor, generate random data, and work with local variables and table variables.

## Writing Conditional Expressions

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

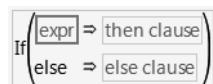
The most basic and general conditional function is the If function. Its arguments are called If, Then, and Else *clauses*.

Another general conditional function is the Match function, which is often used to recode variables.

### Using the If, Row, and Subscript functions

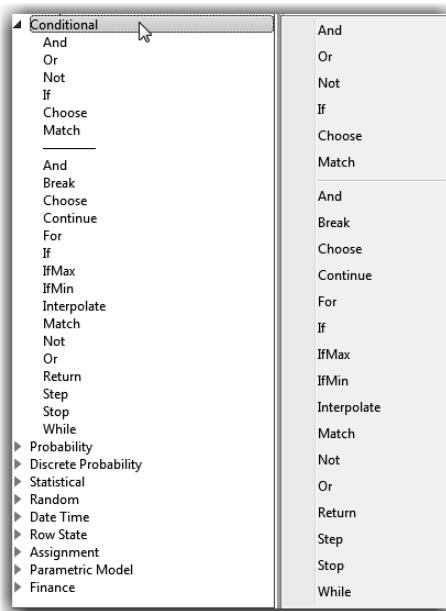
When you highlight an expression and select If, the Formula Editor creates a new conditional expression. It has one If argument (a conditional expression denoted *expr*), one then argument, and a corresponding else and else clause. A conditional is usually an expression, like a < b.

However, any expression that evaluates as a numeric value can be used as a conditional expression. Expressions that evaluate as zero or missing are false. All other numeric expressions are true. An initial If expression looks like the one shown here.



If you need more than one *then* clause, click the Insert button (the caret) on the keypad (or type a comma, its keyboard shortcut) to add a new argument. To remove unwanted arguments, click the Delete button on the keypad or press the Delete key on your keyboard.

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



$$\text{If} \begin{cases} \text{Row}() \leq 2 \Rightarrow 1 \\ \text{else } \Rightarrow \text{Fib Row}() - 1 + \text{Fib Row}() - 2 \end{cases}$$

The first two rows have the value 1.

Every other row is the sum of the previous two rows.

The preceding formula uses the following functions: If, Row, Fib, and Subscript. As an exercise to become more familiar with these functions, create a Fibonacci sequence as follows:

- ⓐ Name a blank data table column Fib.
- ⓑ Right-click on the column header and select **Formula**.
- ⓒ Select **If** from the **Conditional Functions** list.
- ⓓ Select **a<=b** from the **Comparison Functions** list.
- ⓔ Select **Row** from the **Row Functions** list.
- ⓕ Select the second argument of the conditional statement.
- ⓖ Type 2 and press Enter.
- ⓗ Type 1 in the *then clause* of the **If** function and press Enter.
- ⓘ Select the *else clause* of the **If** function.
- ⓙ Click the + key on the keypad, type Fib as the first term, and then press Enter.

$$\text{If} \begin{cases} \text{expr} \Rightarrow \text{then clause} \\ \text{else } \Rightarrow \text{else clause} \end{cases}$$

$$\text{If} \begin{cases} \boxed{\text{ }} \leq \boxed{\text{ }} \Rightarrow \text{then clause} \\ \text{else } \Rightarrow \text{else clause} \end{cases}$$

$$\text{If} \begin{cases} \text{Row}() \leq \boxed{\text{ }} \Rightarrow \text{then clause} \\ \text{else } \Rightarrow \text{else clause} \end{cases}$$

$$\text{If} \begin{cases} \text{Row}() \leq \boxed{\text{ }} \Rightarrow \text{then clause} \\ \text{else } \Rightarrow \text{else clause} \end{cases}$$

$$\text{If} \begin{cases} \text{Row}() \leq 2 \Rightarrow \text{then clause} \\ \text{else } \Rightarrow \text{else clause} \end{cases}$$

$$\text{If} \begin{cases} \text{Row}() \leq 2 \Rightarrow 1 \\ \text{else } \Rightarrow \text{else clause} \end{cases}$$

$$\text{If} \begin{cases} \text{Row}() \leq 2 \Rightarrow 1 \\ \text{else } \Rightarrow \text{else clause} \end{cases}$$

$$\text{If} \begin{cases} \text{Row}() \leq 2 \Rightarrow 1 \\ \text{else } = \text{Fib} + \boxed{\text{ }} \end{cases}$$

- ✓ Select **Subscript** from the Row Functions list.

$$\text{If} \begin{cases} \text{Row}() \leq 2 \Rightarrow 1 \\ \text{else } \Rightarrow \text{Fib}[\text{Row}() - 1] + \text{Fib}[\text{Row}() - 2] \end{cases}$$

- ✓ Select **Row** from the Row list, click the **-** key on the keypad, and then type 1.

$$\text{If} \begin{cases} \text{Row}() \leq 2 \Rightarrow 1 \\ \text{else } \Rightarrow \text{Fib}[\text{Row}() - 1] + \text{Fib}[1] \end{cases}$$

- ✓ Repeat for the second term to add Fib with lag 2.

$$\text{If} \begin{cases} \text{Row}() \leq 2 \Rightarrow 1 \\ \text{else } \Rightarrow \text{Fib}[\text{Row}() - 1] + \text{Fib}[\text{Row}() - 2] \end{cases}$$

After you create the formula and click **OK**, the formula that you entered generates the values shown in **Figure 4.6**.

**Figure 4.6** Results of the Formula Example

Fib
1
1
2
3
5
8
13
21
34
55
89
144
233
377
610
987

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

### Using the Match Function

Conditional functions are commonly used to recode variables. If functions can be used for this, but you can also use the simpler Match function.

When you select **Match** from the **Conditional** menu, the Formula Editor shows a single Match condition with an empty expression and an empty *then clause*. The Match conditional expression compares an expression to a list of clauses. The condition then returns the value of the result expression for the first matching argument that is

$$\text{Match}([\text{expr}])([\text{value} \Rightarrow \text{then clause}])$$

encountered. With **Match**, you provide the matching expression only once and then give a match for each argument.

For example, select **Help > Sample Data Library** and open Car Physical Data.jmp. The variable Type has five levels (Large, Medium, Small, Sporty, and Compact). Suppose that you want to combine Small and Compact into one category, Small/Compact. In the Formula Editor of a new column, follow these steps:

- ⓐ Select **Match** from the **Conditional** list.
- ⓑ Select Type from the list of columns. The *expr* argument in the formula is populated.
- ⓒ On the keypad at the top of the Formula Editor, click the Insert button three times.
- ⓓ Type Small and Compact in the two *value* fields.
- ⓔ Type Small/Compact in the two *then clause* fields.
- ⓕ Click on the *else clause* field and select Type from the list of columns.

$$\text{Match}(\text{Type}) \left( \begin{array}{l} \text{value} \Rightarrow \text{then clause} \end{array} \right)$$

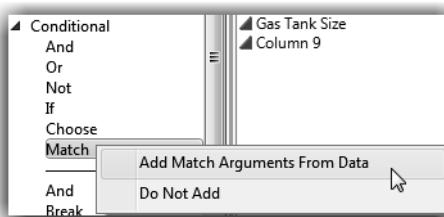
$$\text{Match}(\text{Type}) \left( \begin{array}{l} \text{value} \Rightarrow \text{then clause} \\ \text{value} \Rightarrow \text{then clause} \\ \text{else} \Rightarrow \text{else clause} \end{array} \right)$$

$$\text{Match}(\text{Type}) \left( \begin{array}{l} \text{"Small"} \Rightarrow \text{"Small/Compact"} \\ \text{"Compact"} \Rightarrow \text{"Small/Compact"} \\ \text{else} \Rightarrow \text{Type} \end{array} \right)$$

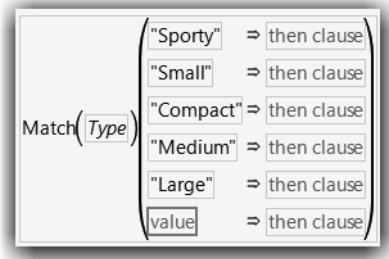
If the values in the Type column are Small or Compact, they are recoded as Small/Compact. Otherwise, the new column is populated with the original values in the Type column.

As a shortcut, you can also auto-populate the Match conditional expression with the values for the variable.

- ⓐ In the Formula Editor, select Type from the list of columns.
- ⓑ Select **Match** from the **Conditional** list.
- ⓒ Select **Add Match Arguments from Data** from the Match pop-up menu.



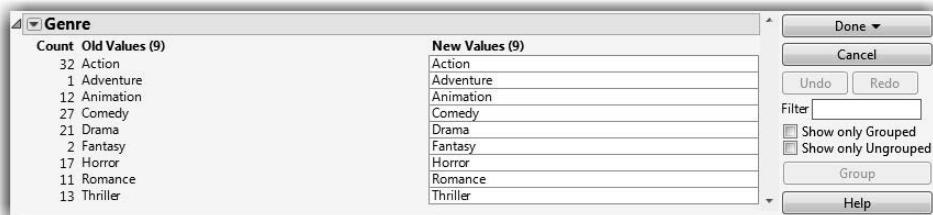
- ✓ Remove the arguments that you will not use, or click the Insert button to add new arguments.



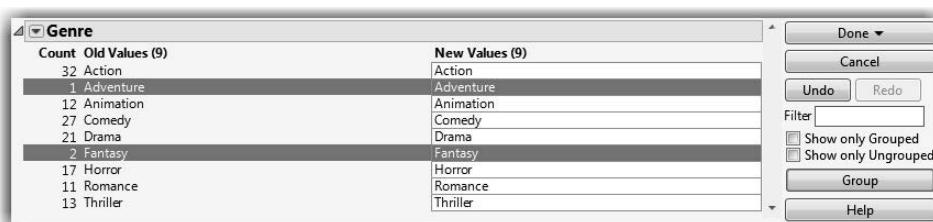
### Writing Match Statements with Recode

The Recode option in JMP provides an interface for recoding data and writing Match formulas. In this example, we use the Hollywood Movies.jmp sample data table. The file contains the **Genre** column, which has nine categories. Only two of the movies are in the Fantasy or Adventure genre. To combine these two genres, follow these steps:

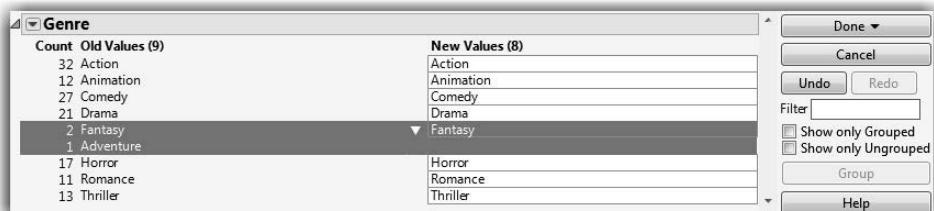
- ✓ Select the **Genre** column in the data table.  
 ✓ Select **Cols > Recode**.



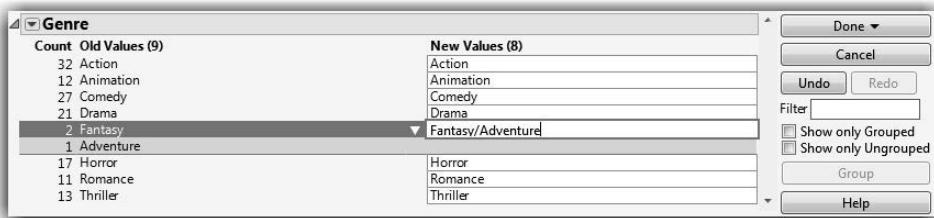
- ✓ Click **Adventure** to select it.  
 ✓ Hold down the Ctrl key and select **Fantasy**.



- ✓ Click **Group**.



✓ Change the new value to Fantasy/Adventure.



✓ Click Done and select **Formula Column**.

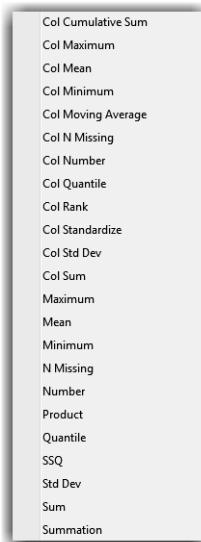
The new column with a Match formula is created.

**Note:** Recode provides many additional features for changing values in a column. Recoding is ideal for preparing categorical data for analysis.

## Summarizing Data with the Formula Editor

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

The functions with names prefaced by “Col” (Col Mean, Col Sum, and so on) always evaluate for all of the rows in a column. Thus, used alone as a column formula, these functions produce the same value for each row. You can add one or more grouping or By variables to the functions, which results in values for each combination of the By variables.



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

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

### The Quantile Function

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

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

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

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

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

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

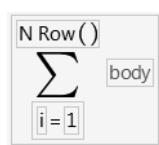
Col Quantile(age, 0.0) calculates the minimum age.

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

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

**Note:** To calculate a Quantile using **New Formula Column** from the data table, right-click on the column header and select **New Formula Column > Aggregate > Quantile**.

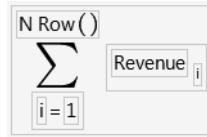
### Using the Summation Function



The Summation ( $\Sigma$ ) function uses the summation notation shown here. To calculate a sum, select **Summation** from the **Statistical** function menu and create its argument. The Summation function repeatedly evaluates the expression for the index that you apply to the body of the function from the lower summation limit to the upper summation limit. The function then adds the nonmissing results together to determine the final result. You can replace the index  $i$ , the index constant 1, and the upper limit, NRow(), with any expressions appropriate for your formula.

Use the **Subscript** function in the **Row** function category to create a subscript for the body of the summation.

For example, the summation shown here computes the total of all revenue values for row 1 through the current row number. The function fills the calculated column with the cumulative totals of the revenue column.



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

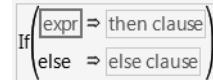
- ☛ Select **Help > Sample Data Library** and open Current Stock Averages.jmp.

Note that there is a column in the table called **Moving Average**. Use this column and its formula as a reference.

- ☛ Create a new column by selecting **Cols > New Columns**.
- ☛ Name the column **Moving Average 2** and select **Formula** from the **Column Properties** list.

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

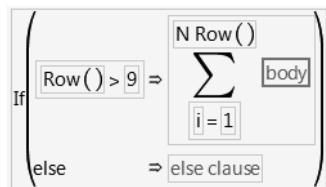
- ☛ Begin by selecting the conditional **If** function.



- ☛ With the **If** expression highlighted, select **a>b** from the **Comparison** function category, which is used to determine the row number value.
- ☛ For the left side of the comparison, select **Row** from the **Row** functions.
- ☛ Highlight the right side of the comparison and type 9. The **If** expression should now appear as **Row()>9**.

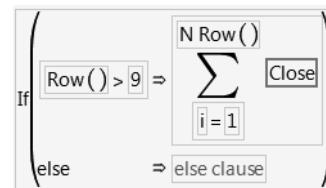


- Now highlight the **then** clause and begin the formula to compute the ten-day moving average by selecting the **Summation** function from the **Statistical** function category.  
 Highlight the body of the summation and click Close in the columns list.



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

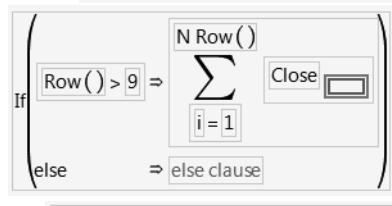
- Highlight the summation body, Close.



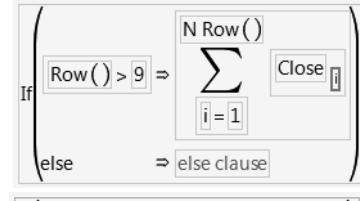
- Select **Subscript** from the **Row** function category.

An empty subscript now appears with the summation body.

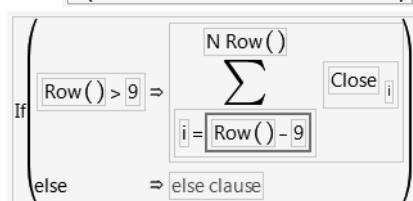
- To assign the subscript, either type the letter *i*, or drag the *i* from the lower limit of the summation into the empty subscript.



- Select the 1 in the lower summation limit and hold down the Delete key to change it to an empty term.



- Enter the expression  $\text{Row}() - 9$  inside the parentheses, using the **Row** selection in the **Row** function category.



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

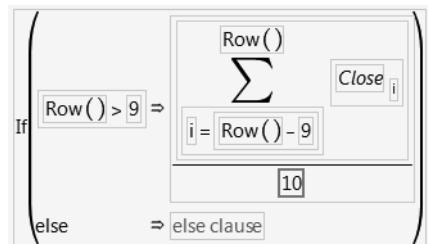
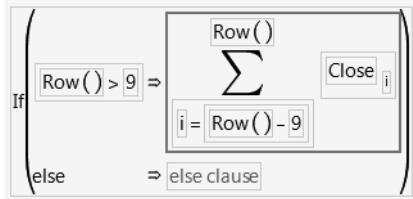
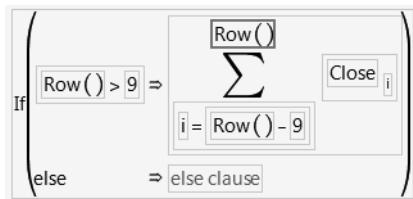
To finish the moving average formula, you want to divide the sum by 10, but not start the averaging process until you actually have 10 values to work with.

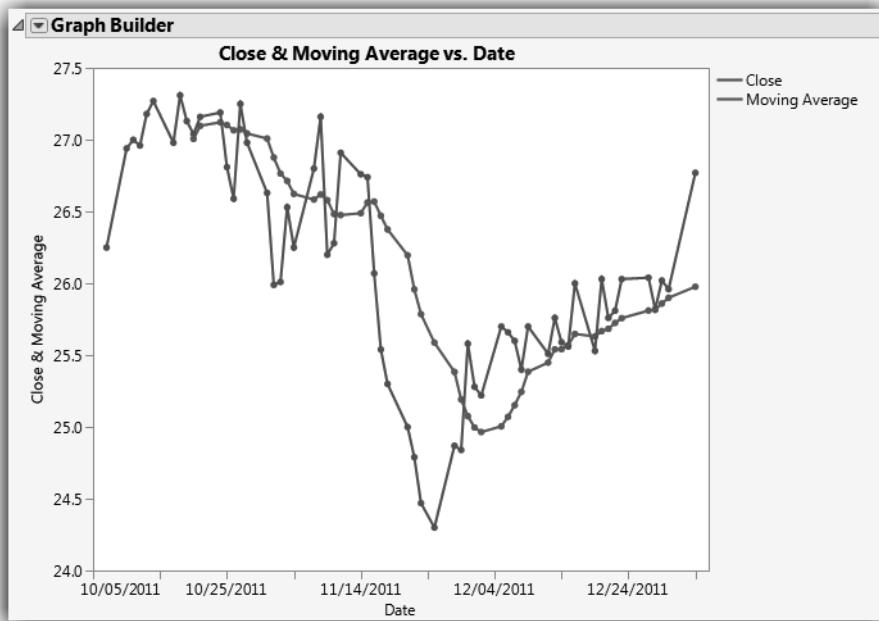
- ⓐ Click in the summation to highlight the whole summation expression.
- ⓐ Click the divide operator on the control panel, and then enter the constant “10” into the highlighted dominator that appears.
- ⓐ When you click **Apply** or close the Formula Editor, the Moving Average 2 column fills with values.

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

- ⓐ Select **Graph > Graph Builder**, and drag Date to the **X** zone and both Close and Moving Average to the **Y** zone (at the same time).
- ⓐ Right-click (or Ctrl-click on a Macintosh), and select **Smoother > Change to > Line**.
- ⓐ Click **Done** and customize the graph as desired by changing the graph title, the orientation of the axes, and so on.

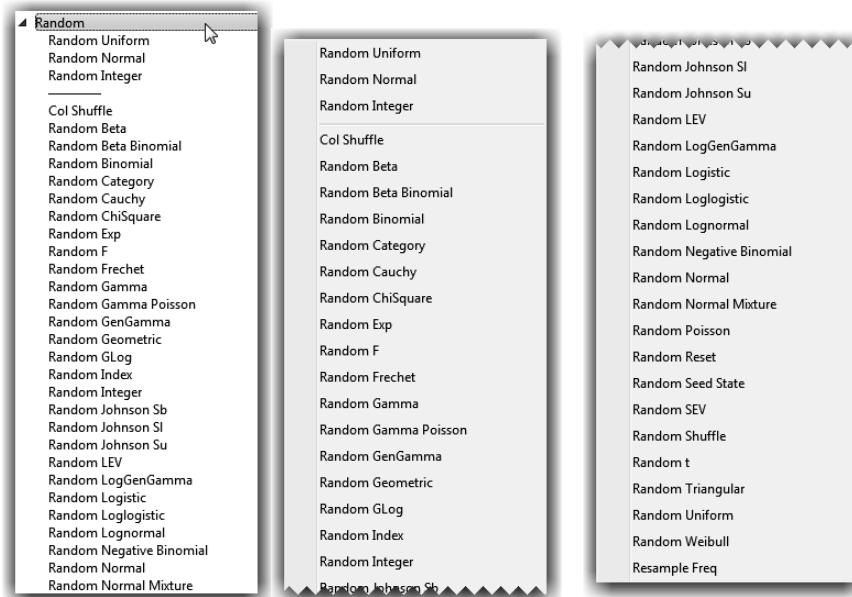
You then see the plot in **Figure 4.7**, which compares closing stock values with their ten-day moving average.



**Figure 4.7** Plot of Close and Its Moving Average over Date

## Generating Random Data

Random number functions generate real numbers by essentially “rolling the dice” within the constraints of the specified distribution. You can use the random number functions with a default “seed” that provides a pseudo-random starting point for the random series. You can also use the **Random Reset** function and give a specific starting seed. **Figure 4.8** shows JMP’s extensive random number menu.

**Figure 4.8** Random Functions Menu in Column Calculator

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

### The Uniform Distribution

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

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

- ☛ Select **File > New** to create a new data table.
- ☛ Right-click (or press Ctrl and click) on Column 1 and select **Formula**.
- ☛ When the Formula Editor window appears, select **Random Uniform** from the **Random** function menu in the function list, and then close the Formula Editor.
- ☛ Select **Rows > Add Rows** and add 500 rows.

The table fills with random uniform values between 0 and 1.

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

- ☞ Select **Cols > New Columns** to create a second column.
- ☞ Click **Random** in the function list and select **Random Uniform** from its menu.
- ☞ Click the multiply sign on the keypad and enter 10 as the multiplier.
- ☞ Select the entire formula and click the addition sign on the keypad.
- ☞ Enter 1 in the empty argument term of the plus operator.

**Note:** JMP has a **Random Integer(n)** function that selects integers from a uniform distribution from 1 to n. It could be used here for the same effect. We're using the **Random Uniform** function to illustrate how to manipulate a random number by multiplying and adding constants. You can see an example of the **Random Integer** function in "Rolling Dice" on page 111.

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

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

The final formula is

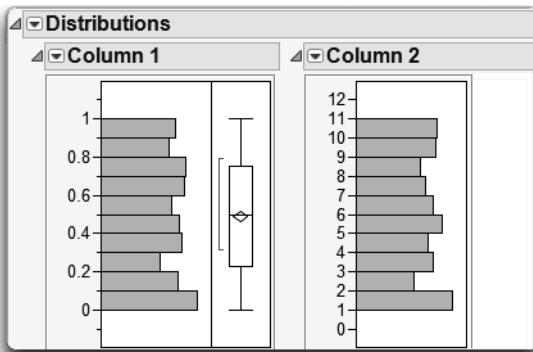
`Floor(Random Uniform()*10+1)`

- ☞ Click **OK** to close the Formula Editor.

You now have a table template for creating two uniform distributions.

- ☞ Change the modeling type of the integer column to nominal so that JMP treats it as a discrete distribution.
- ☞ Select **Analyze > Distribution**, assign both columns to **Y, Columns**, and then click **OK**.

You see two histograms similar to those shown in **Figure 4.9**. The histogram on the left represents simple uniform random numbers, and the histogram on the right shows random integers from 1 to 10.

**Figure 4.9** Example of Two Uniform Distribution Simulations

### The Normal Distribution

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

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

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

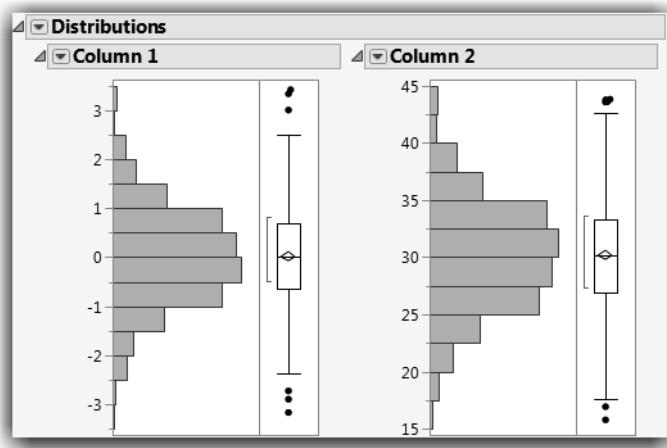
The **Random Normal()** function takes two optional arguments. The first specifies the mean of the distribution; the second specifies the standard deviation of the distribution.

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

The modified normal formula is

`Random Normal(30, 5)`

**Figure 4.10** shows the Distribution platform results for these normal simulations.

**Figure 4.10** Illustration of Normal Distributions

### The Col Shuffle Command

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

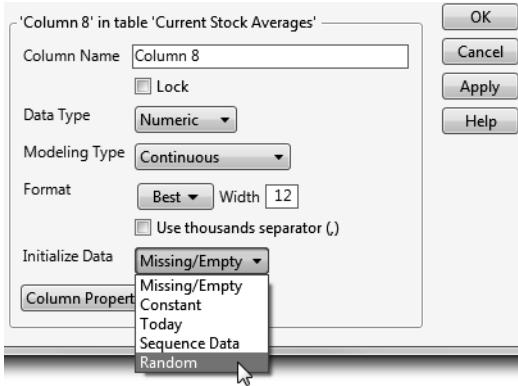
For example, to identify a 50% random sample without replacement, use the following formula:

$$\text{If} \left( \text{Row}() < \frac{\text{N Row}()}{2} \Rightarrow \text{Column 1}[\text{Col Shuffle}] \right)$$

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

### Initialize Data Options

When you add a new column to a data table, the **Initialize Data** menu in the Column Info window (shown in **Figure 4.11**) enables you to specify the type of initial data values for the new column. The default is **Missing/Empty**. The **Random** option populates the column with random data without accessing the Formula Editor or storing a formula in the column.

**Figure 4.11** Initialize Data Options

## Local Variables and Table Variables

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

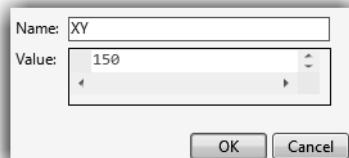
To create a new local variable, use the button on the formula editor keypad. This button adds a temporary local variable to the formula editing area, which appears as a command ending in a semicolon. Alternatively, you can select **Local Variables** from the Formula Elements menu, select **New Local**,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

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

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

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

- 1 Select **Local Variables**.
- 2 Click on **New Local**.
- 3 Fill in the resulting dialog box with the name and value.



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

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

```


$$XY = \sum_{i=1}^{N \text{ Row}()} \left( X_i - \text{Col Mean}(X) \right) \cdot \left( Y_i - \text{Col Mean}(Y) \right);$$


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

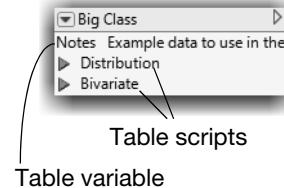

$$\frac{XY}{Xsqrd};$$


```

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

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

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



# Working with Dates

Working with dates can present a challenge. Below are some frequently encountered example formulas involving dates.

## Calculating Elapsed Time

Columns storing dates should be formatted as dates in the Column Info window. JMP dates are stored as the number of seconds since Jan 1, 1904. Creating the elapsed time in days (or minutes, hours, and so on) between two date-formatted columns requires the appropriate conversion formula. The formula shown here converts the elapsed time in seconds between Date 1 and Date 2 to days.

	Date 1	Date 2	Elapsed Time in Days
1	10/07/2011	11/12/2011	36
2	10/10/2011	11/08/2011	29

$$\frac{[(Date\ 2 - Date\ 1)]}{(60 * 60 * 24)}$$

## Age in Days

A variation of the elapsed time problem is calculating age (in days, minutes, and so on). This formula calculates the age in days between a date-formatted column and today using the `Floor` and `Today` functions.

Floor	<code>Today() - Date 2</code>
	<code>(60 * 60 * 24)</code>

**Note:** The Date Time Functions list in the Formula Editor provides a number of functions for working with dates. For example, you can use the Date Difference function to calculate the difference in two date-time values, using defined intervals (month, day, or hour).

## Fixing Unformatted Dates

If dates have been entered into JMP in a character format, they can be converted to numeric date formats using a few key functions: **Num**, **Word**, and **Substr** (all under the **Characters** function group).

Unformatted Dates	Date	Date as Numeric
1 February 25, 2011	02/25/2011	
2 February 26, 2011	02/26/2011	
3 February 27, 2011	02/27/2011	
4 February 28, 2011	02/28/2011	

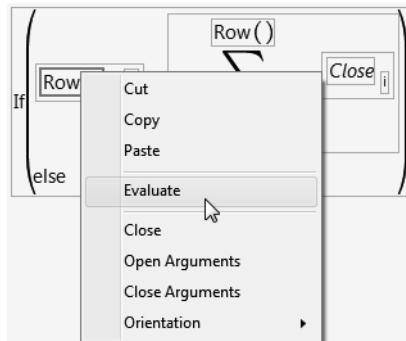
Num	<code>Word[2, Date, ", "]</code>
	<code>Substr[Word[1, Date, ", "], 1, 3]</code>
	<code>   Word[3, Date, ", "]</code>

The formula shown above parses the character string and converts the dates in the Date column to a numeric format. Note that this formula returns values in “raw” format (number of seconds since Jan 1, 1904). Change the date format in the Column Info window to display as a date.

## Tips on Building Formulas

### Examining Expression Values

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



### Cutting, Dragging, and Pasting Formulas

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

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

## Selecting Expressions

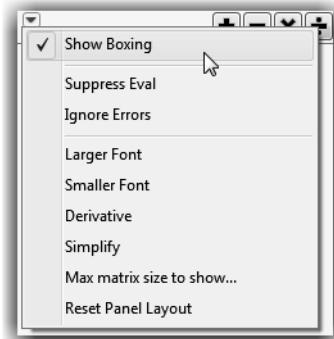
The Show Boxing red triangle option above the Formula Editor editing area tells JMP to outline terms within a formula. This outlining is called *boxing*.

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

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

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

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



## Exercises

1. The sample data table Pendulum.jmp contains the results of an experiment in a physics class. The data compare the length of a pendulum to its period (the time it takes the pendulum to make a complete swing). Calculations were made for a range of pendulums from short (2 cm) to long (20 m). Use the calculator to determine a model to predict the period of a pendulum from its length.
  - (a) Produce a scatterplot of the data by selecting **Analyze > Fit Y By X** and assigning Period to **Y, Response** and Length to **X, Factor**.
  - (b) Create a new column named Transformed Period that contains a formula to take the square root of the Period column. Produce a scatterplot of Transformed Period vs. Length. Is this the graph of a linear function?

- (c) Try other transformations (for example, natural log of Period, reciprocal of Period, or square of Period) until the scatterplot looks linear.
- (d) Find the line of best fit for the linear transformed data by selecting **Fit Line** from the pop-up menu beside the title of the scatterplot.
- (e) This line is not the fit of the original data, but of the transformed data. Substitute a term representing the transformation that you did to linearize the data. For example, if the square root transformation made the data linear, substitute  $\sqrt{\text{Period}}$  into the regression equation and solve the equation for Period.

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

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

- (f) Create a new column to use this formula to calculate the theoretical values. Next, construct another column to calculate the difference between the observed values of the students and the theoretical values.
  - (g) Examine a histogram of these differences. Does it appear that there was a trend in the observations of the students?
2. Is there a correlation among the mean, minimum, maximum, and standard deviation of a set of data? To investigate, create a new data table in JMP with these characteristics:
- (a) Create ten columns of data named Data 1 through Data 10, each with the formula `Random Uniform()`. Add 500 rows to this data table.
  - (b) Create four columns to hold the four summary statistics of interest, one column each for the mean, minimum, maximum, and standard deviation. Create a formula in each column to calculate the appropriate statistic of the ten data rows.
  - (c) Select the Multivariate platform in the **Analyze > Multivariate Methods** menu and include the four summary statistics as the Ys in the resulting window. Pressing **OK** should produce 16 scatterplots. Which statistics seem to show a correlation?
  - (d) As an extension, select two of the statistics that seem to show a correlation. Produce a single scatterplot of these two statistics using the **Fit Y by X** platform in the **Analyze** menu. From the red triangle menu beside the title

of the plot, select **Nonpar Density**. Then select **Save Density Grid** from the menu beside the density legend.

- (e) Finally, select **Graph > Scatterplot 3D** and include the first three columns of the saved density grid. You should now see a 3D scatterplot of the correlation, which contains a peak where the data points overlapped each other. Use the hand tool to move the plot around.
3. Make a data table consisting of 20 rows and a single column that holds the Fibonacci sequence (whose formula is shown on page 73). Label this column Fib.
- (a) Add a new column called Ratio, and give it the following formula to take the ratio of adjacent rows.

**Note:** This produces an error when evaluated for the first row. Click **OK** on the error window and resolve this issue.

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

What value does this ratio converge to?

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

$$\left[ \begin{array}{l} \text{If } \text{Row}() == 1 \Rightarrow 2 \\ \text{If } \text{Row}() == 2 \Rightarrow 5 \\ \text{else } \quad \Rightarrow \text{Lucas}_{\text{Row}()-1} + \text{Lucas}_{\text{Row}()-2} \end{array} \right]$$

- (c) Create a column to calculate the ratio of two successive terms of the Lucas sequence in part (b). Is it the same as the number in part (a)?
- (d) Create a Lucas sequence starting with the values 1, 2. Calculate the ratio of successive terms and compare it to the answer in part (c).
- (e) There are innumerable other properties of the Fibonacci sequence. For example, add a column that contains the following formula, and comment on the result.

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



# 5

## What Are Statistics?

### Overview

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

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

## Chapter Contents

Overview .....	95
Ponderings.....	97
The Business of Statistics .....	97
The Yin and Yang of Statistics.....	97
The Faces of Statistics .....	98
Don't Panic.....	99
Preparations .....	101
Three Levels of Uncertainty.....	101
Probability and Randomness .....	102
Assumptions .....	102
Data Mining? .....	103
Statistical Terms .....	104

# Ponderings

## The Business of Statistics

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

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

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

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

Statistics is a necessary bureaucracy of science.

## The Yin and Yang of Statistics

There are two sides to statistics.

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

their improbability. The yang mechanically cranks through the ore of knowledge and distills it to the answer.

On the other side, we find the contrapositive Yin, the moon, reflecting the light. The yin catches the shadow of the world, feeling the shape of truth under the umbra. The yin is forever looking and listening for clues, nurturing seeds of pattern and anomaly into maturing discoveries. The yin whispers its secrets to our left hemisphere. It unlocks doors for us in the middle of the night, planting dream seeds, making connections. The yin draws out the puzzle pieces to tantalize our curiosity. It teases our awareness and tickles our sense of mystery until the climax of revelation—Eureka!

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

## The Faces of Statistics

Everyone has a different view of statistics.

<b>Match the definition on this side....</b>	<b>with someone likely to have said it on this side</b>
1. The literature of numerical facts.	a. Engineer
2. An applied branch of mathematics.	b. Original meaning
3. The science of evidence in the face of uncertainty.	c. Social scientist
4. A digestive process that condenses a mass of raw data into a few high-nutrient pellets of knowledge.	d. Philosopher
5. A cooking discipline with data as the ingredients and methods as the recipes.	e. Economist
6. The calculus of empiricism.	f. Computer scientist
7. The lubricant for models of the world.	g. Mathematician
8. A calibration aid.	h. Physicist

<b>Match the definition on this side....</b>	<b>with someone likely to have said it on this side</b>
9. Adjustment for imperfect measurement.	i. Baseball fan
10. An application of information theory.	j. Lawyer
11. Involves a measurable space, a sigma algebra, and Lebesgue integration.	k. Joe College
12. The nation's state.	l. Politician
13. The proof of the pudding.	m. Businessman
14. The craft of separating signal from noise.	n. Statistician
15. A way to predict the future.	

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

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

## Don't Panic

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

### Abstract Mathematics

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

### Lingo

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

### Awkward Phrasing

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

### A Bad Reputation

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

### Uncertainty

Although we are comfortable with uncertainty in ordinary daily life, we are not used to embracing it in our knowledge of the world. We think of knowledge in terms of hard facts and solid logic, though much of our most valuable real knowledge is far from solid. We can say when we know something for sure

(yesterday, it rained), and we can say when we don't know (don't know whether it will rain tomorrow). But when we describe knowing something with incomplete certainty, it sounds apologetic or uncommitted. For example, it sounds like a form of equivocation to say that there is a 90% chance that it will rain tomorrow. Yet much of what we think we know contains just that type of uncertainty.

## Preparations

Learning a few fundamental concepts prepares you for absorbing details in upcoming chapters.

### Three Levels of Uncertainty

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

#### **Random Events**

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

#### **Unknown Parameters**

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

#### **Unknown Models**

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

Statistics is very limited in its help for certifying that a model is correct. Most statistical conclusions assume that the hypothesized model is correct. The correctness of the model is the responsibility of the subject-matter science.

Statistics might give you clues if the model is not carrying the data very well. Statistical analyses can give diagnostic plots to help you see patterns that could lead to new insights, to better models.

## Probability and Randomness

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

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

Statistical language uses the term probability on several levels:

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

## Assumptions

Statisticians are naturally conservative professionals. Like the boilerplate of official financial audits, statisticians' opinions are full of provisos such as "assuming that the model is correct, and assuming that the error is normally distributed, and assuming that the observations are independent and identically distributed, and assuming that there is no measurement error, and assuming...." Even then, the conclusions are hypothetical, with phrases like "if you say the hypothesis is false, then the probability of being wrong is less than 0.05."

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

### Ceteris Paribus

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

### Is the Model Correct?

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

### Is the Sample Valid?

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

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

## Data Mining?

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

given test significant at the 0.05 level about 5% of the time. So you could just repeat the experiment until you get what you want, discarding the rest. That's obviously bad science, but something similar often happens in published studies. This multiple-testing problem remains largely unaddressed by statistical literature and software. Exceptions include means comparisons, a few general methods that might be inefficient (Bonferroni's adjustment), and expensive, brute-force approaches (resampling methods).

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

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

## Statistical Terms

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

### **Model**

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

### **Parameters**

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

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

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

### Hypotheses

In science, the *hypothesis* is the bright idea that you want to confirm. In statistics, this is turned upside down because it uses logic analogous to a proof-by-contradiction. The so-called *null hypothesis* is usually the statement that you want to demolish. The usual null hypothesis is that some factor has no effect on the response. You are of course trying to support the opposite, which is called the *alternative hypothesis*. You support the alternative hypothesis by statistically rejecting the null hypothesis.

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

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

### Statistical Significance

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

**Significance Level, *P*-value,  $\alpha$ -level**

To reject a null hypothesis, you want small *p*-values. The *p-value* is the probability of being wrong if you declare an effect to be non-null—that is, the probability of rejecting a “true” null hypothesis. The *p*-value is sometimes labeled the *significance probability*, or sometimes labeled more precisely in terms of the distribution that is doing the measuring. The *p*-value labeled “Prob>|t|” is read as “the probability of getting a greater t (in absolute value).” The  $\alpha$ -level is your standard of the *p*-value that you claim, so that *p*-values below this reject the null hypothesis (that is, they show that there is an effect).

**Power,  $\beta$ -level**

*Power* is how likely you are to detect an effect if it is there. The more data you have, the more statistical power. The greater the real effect, the more power. The less random variation in your world, the more power. The more sensitive your statistical method, the more power. If you had a method that always declared an effect significant, regardless of the data, it would have a perfect power of 1. However, it would have an  $\alpha$ -level of 1, too, the probability of declaring significance when there was no effect. The goal in experimental design is usually to get the most power that you can afford, given a certain  $\alpha$ -level. It is not a mistake to connect the statistical term power with the common sense of power as persuasive ability. It has nothing to do with work or energy, though.

**Confidence Intervals**

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

### Biased, Unbiased

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

### Sample Mean versus True Mean

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

### Variance and Standard Deviation, Standard Error

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

### Degrees of Freedom

*Degrees of freedom* (df) is the specific name for a value that indexes some popular distributions of test statistics. It is called degrees of freedom because it relates to differences in numbers of parameters that are or could be in the model. The more parameters a model has, the more freedom it has to fit the data better. The df (degrees of freedom) for a test statistic is usually the difference in the number of parameters between two models.





# 6

## Simulations

### Overview

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

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

## Chapter Contents

Overview .....	109
Rolling Dice .....	111
Rolling Several Dice .....	114
Flipping Coins, Sampling Candy, or Drawing Marbles .....	114
Probability of Making a Triangle .....	115
Confidence Intervals .....	120
Data Table-Based Simulations .....	121
Other JMP Simulators .....	122
Exercises .....	123

## Rolling Dice

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

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

- ❖ Open the `DiceRolls.jmp` sample data table from **Help > Sample Data > Simulations**.

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

**Figure 6.1** `DiceRolls.jmp` Data Table

Use these scripts to conduct the simulation.

	Dice Roll	Average
Num Rolls	1000	
Roll Once		
Roll Many		
Plot Results		
Columns (2/0)		
Dice Roll		
Average		

```

Random Integer (6)
If Row() == 1 => Dice Roll
else
    => Dice Roll + Average Row() - 1 * (Row() - 1)
        _____
                    Row()
    
```

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

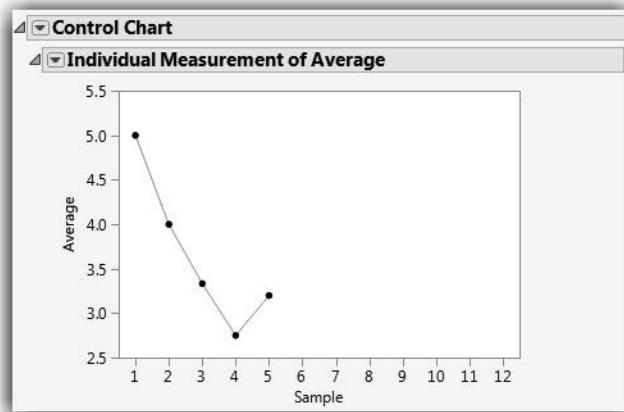
- ⌚ Click the green triangle next to the **Roll Once** script in the side panel of the data table to run the script.

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

- ⌚ Repeat this three or four times to add rows to the data table.
- ⌚ After rows have been added, run the **Plot Results** script in the side panel of the data table.

This produces the control chart of the results in **Figure 6.2**. Note that the results fluctuate fairly widely at this point.

**Figure 6.2** Plot of Results after Five Rolls



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

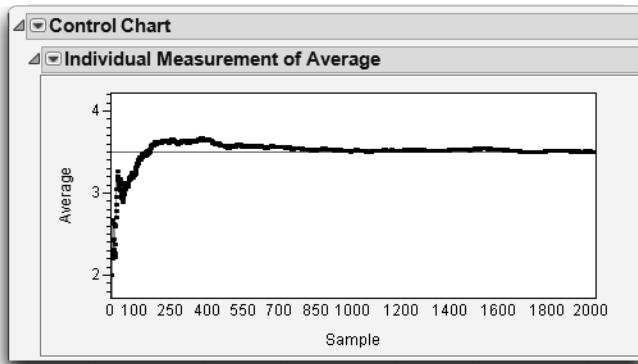
This adds many rolls at once. In fact, it adds the number of rows specified in the table variable **Num Rolls** (1000) each time it is clicked. To add more or fewer rolls at one time, adjust the value of the **Num Rolls** variable. Click the green triangle next to the **Num Rolls** script at the top of the tables panel and enter any number you want in the edit box.

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

- ⌚ Continue adding points until there are about 2,000 points in the data table.

You need to manually adjust the  $x$ -axis to see the plot in **Figure 6.3**.

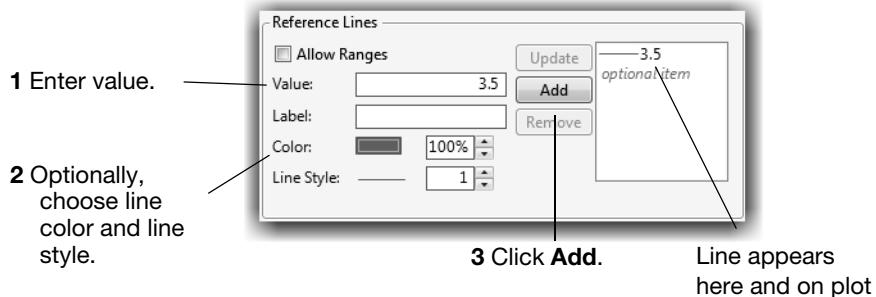
**Figure 6.3** Observed Mean Approaches Theoretical Mean



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

- ⌚ Double-click the  $y$ -axis to open the axis specification window.
- ⌚ Enter values into the window as shown in **Figure 6.4**.

**Figure 6.4** Adding a Reference Line to a Plot



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

## Rolling Several Dice

If you want to roll more than one die at a time, simply copy and paste the formula from the existing column into other columns. Adjust the running average formula to reflect the additional random dice rolls.

## Flipping Coins, Sampling Candy, or Drawing Marbles

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

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

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

$$\text{If} \begin{cases} \text{Random Uniform ()} < 0.5 \Rightarrow "H" \\ \text{else} \qquad \qquad \qquad \Rightarrow "T" \end{cases}$$

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

Extending this to sampling candies of different colors is easy. Suppose you have a bag of multi-colored candies with the distribution shown on the left in **Figure 6.5**.

Also, suppose you had a column named *t* that held random numbers from a uniform distribution. Then an appropriate JMP formula could be the middle formula in **Figure 6.5**.

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

Or, you could use a slightly more complicated formula. The formula on the right in **Figure 6.5** uses a local variable called *t* to combine the random number and candy selection into one column formula. Note that a semicolon is needed to separate the two scripting statements. This formula eliminates the need to have the extra column, *t*, in the data table.

**Figure 6.5** Probability of Sampling Different Color Candies

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

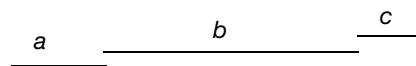
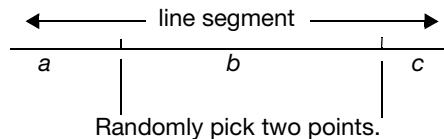
If  $t < 0.1 \Rightarrow \text{"Blue"}$   
 If  $t < 0.2 \Rightarrow \text{"Brown"}$   
 If  $t < 0.3 \Rightarrow \text{"Green"}$   
 If  $t < 0.4 \Rightarrow \text{"Orange"}$   
 If  $t < 0.6 \Rightarrow \text{"Red"}$   
 If  $t < 0.8 \Rightarrow \text{"Yellow"}$   
 else  $\Rightarrow \text{"Purple"}$

If  $t < 0.1 \Rightarrow \text{"Blue"}$   
 If  $t < 0.2 \Rightarrow \text{"Brown"}$   
 If  $t < 0.3 \Rightarrow \text{"Green"}$   
 If  $t < 0.4 \Rightarrow \text{"Orange"}$   
 If  $t < 0.6 \Rightarrow \text{"Red"}$   
 If  $t < 0.8 \Rightarrow \text{"Yellow"}$   
 else  $\Rightarrow \text{"Purple"}$

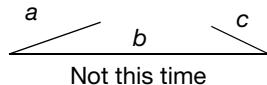
## Probability of Making a Triangle

Suppose you randomly select two points along a line segment. Then, break the line segment at those two points forming three line segments, as illustrated here. What is the probability that a triangle can be formed from these three segments? (See Isaac, 1995.) It seems clear that you cannot form a triangle if the sum of any two of the subsegments is less than the third. This situation is simulated in the triangleProbability.jsl script, found in the Sample Scripts folder. (Select **Help > Sample Data > Open the Sample Script Directory** to find the script.) In the open script, click **Edit > Run Script** to create a data table that holds the simulation results.

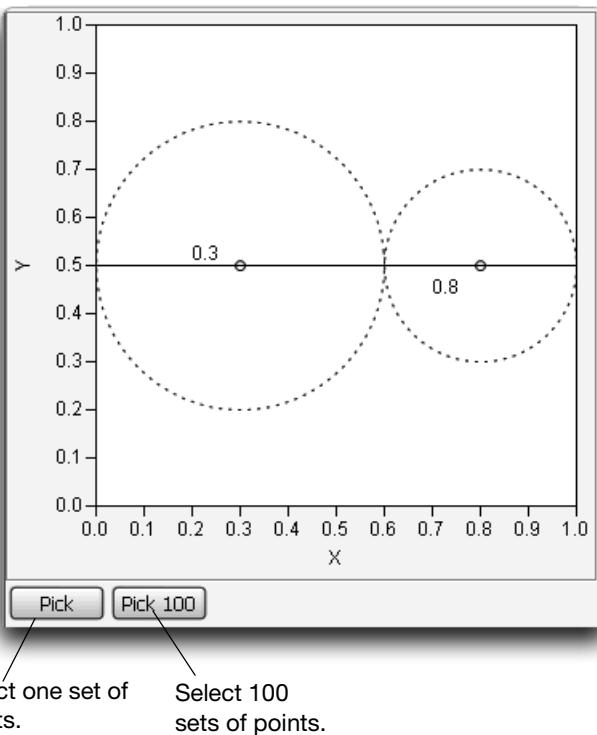
The initial window is shown in **Figure 6.6**. For each of the two selected points, a dotted circle indicates the possible positions of the “broken” line segment that they determine.



When can you form a triangle from three random subsegments?



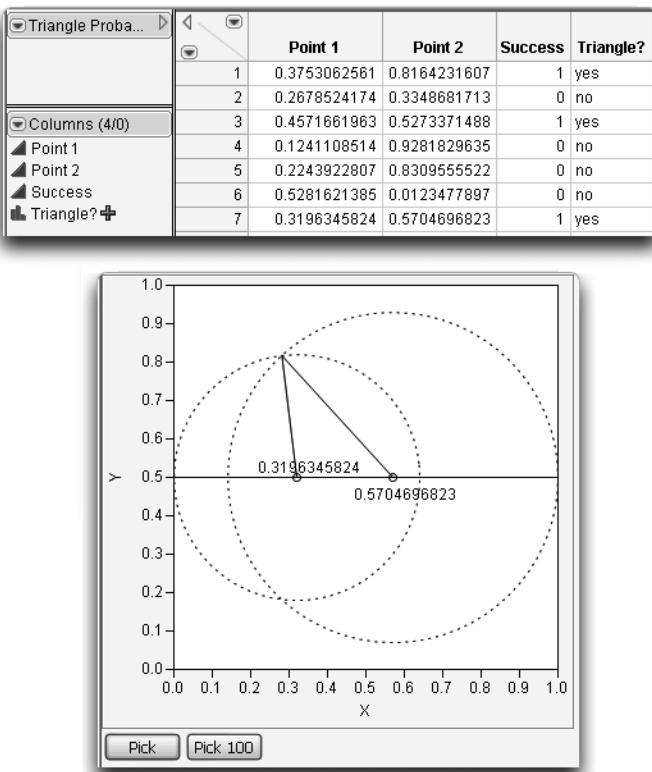
Not this time

**Figure 6.6** Initial Triangle Probability Window

To use this simulation:

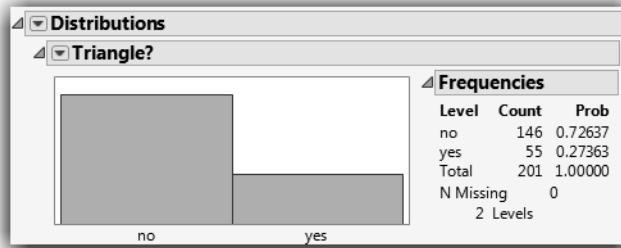
- Click the **Pick** button to select a single pair of points.

Two points are selected, and their information is added to a data table. The results after seven simulations are shown in **Figure 6.7**.

**Figure 6.7** Triangle Simulation after Seven Iterations

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

- ⓐ Click the **Pick 100** button a couple of times to generate a large number of samples.
- ⓑ When finished, select **Analyze > Distribution** and select Triangle? to **Y, Columns**.
- ⓒ Click **OK** to see the distribution report in **Figure 6.8**.

**Figure 6.8** Triangle Probability Distribution Report

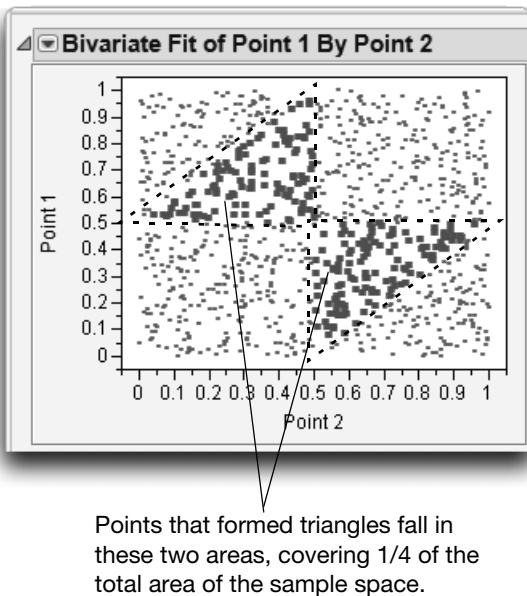
It appears (in this case) that about 26% of the samples result in triangles. To investigate whether there is a relationship between the two selected points and their formation of a triangle,

- ⓐ Select **Rows > Color or Mark by Column** to see the column and color selection window.
- ⓐ Select the **Triangle?** column on the window and be sure to select **Save to Column Property**.
- ⓐ Click **OK**.

This puts a different color on each row depending on whether it formed a triangle (Yes) or not (No). Examine the data table to see the results.

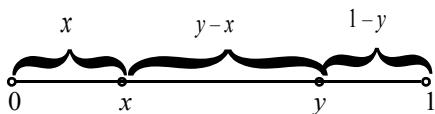
- ⓐ Select **Analyze > Fit Y By X** and assign Point 1 to **Y, Response** and Point 2 to **X, Factor**.

This reveals a scatterplot that clearly shows a pattern.

**Figure 6.9** Scatterplot of Point 1 by Point 2

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

Analytically, this makes sense. If the two randomly selected points are  $x$  and  $y$ , letting  $x$  represent the smaller of the two, then we know  $0 < x < y < 1$ , and the three segments have length  $x$ ,  $y - x$ , and  $1 - y$  (see **Figure 6.10**).

**Figure 6.10** Illustration of Points

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

$$\begin{aligned}x + (y - x) &> 1 - y \\(y - x) + (1 - y) &> x \\(1 - y) + x &> y - x\end{aligned}$$

Elementary algebra simplifies these inequalities to

$$x < 0.5$$

$$y > 0.5$$

$$y - x < 0.5$$

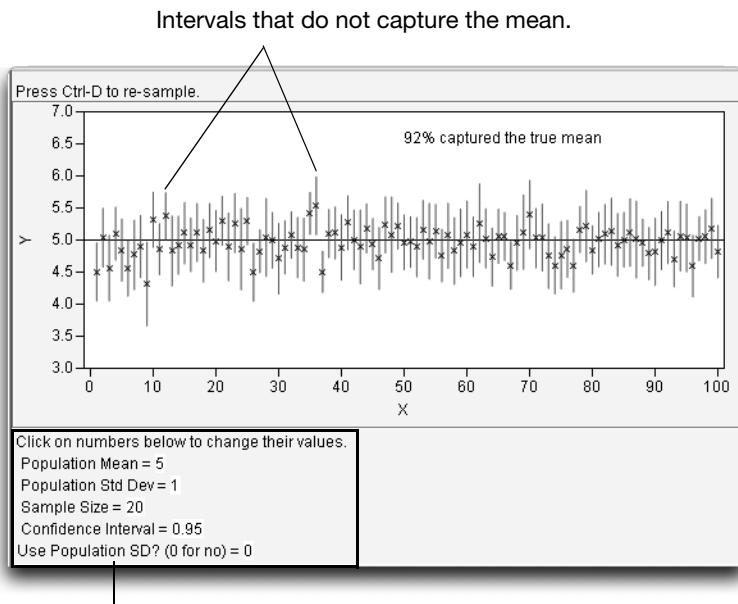
which explain the upper triangle in **Figure 6.9**. Repeating the same argument with  $y$  as the smaller of the two variables explains the lower triangle.

## Confidence Intervals

Beginning students of statistics and nonstatisticians often think that a 95% confidence interval contains 95% of a set of sample data. It is important to help students understand that the confidence measurement is on the test methodology itself.

To demonstrate the concept, use the confidence script from **Help > Sample Data > Teaching Script > Teaching Demonstrations**. Its output is shown in **Figure 6.11**.

**Figure 6.11** Confidence Interval Script



The script draws 100 samples of sample size 20 from a normal distribution with a mean of 5 and a standard deviation of 1. For each sample, the mean is computed with a 95% confidence interval. Each interval is graphed, in gray if the interval captures the overall mean and in red if it doesn't. Note that the gray intervals cross the mean line on the graph (meaning they capture the mean); the red lines don't cross the mean.

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

This script can also be used to illustrate the effect of changing the confidence level on the width of the intervals.

 Change the confidence interval to 0.5.

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

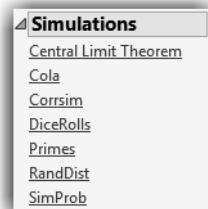
The **Use Population SD?** option enables you to use the population standard deviation in the computation of the confidence intervals (rather than the one from the sample). When this is set to “no”, all the confidence intervals are the same width.

## Data Table-Based Simulations

Some of the simulation examples in this chapter are *table templates* found in the Simulations section of **Help > Sample Data**. A table template is a table that has no rows, but has columns with formulas that use a random number function to generate a given distribution. You add as many rows as you want and examine the results with the Distribution platform and other platforms as needed.

Many popular simulations in table templates, including DiceRolls, have been added to the Simulations outline in the Teaching Resources section under **Help > Sample Data**. These simulations are described below.

- DiceRolls is the first example in this chapter.



- Primes is not actually a simulation table. It is a table template with a formula that finds each prime number in sequence, and then computes differences between sequential prime numbers.
- RandDist simulates four distributions: Uniform, Normal, Exponential, and Double Exponential. After adding rows to the table, you can use Distribution or Graph Builder to plot the distributions and compare their shapes and other characteristics.
- SimProb has four columns that compute the mean for two sample sizes (50 and 500), for two discrete probabilities (0.25 and 0.50). After you add rows, use the Distribution platform to compare the difference in spread between the samples sizes, and the difference in position for the probabilities.  
**Note:** After creating the histograms, use the **Uniform Scaling** command from the top red triangle menu. Then select the grabber (hand) tool from the tools menu and stretch the distributions.
- Central Limit Theorem has five columns that generate random uniform values taken to the 4th power (a highly skewed distribution) and finds the mean for sample sizes 1, 5, 10, 50, and 100. You add as many rows to the table as you want and plot the means to see the Central Limit Theorem unfold. You'll explore this simulation in an exercise, and we'll revisit it later in the book.
- Cola is presented in Chapter 11, "Categorical Distributions," to show the behavior of a distribution derived from discrete probabilities.
- Corrsim simulates two random normal distributions and computes the correlation between them at levels 0.50, 0.90, 0.99, and 1.00. **Note:** After adding columns, use the **Fit Y by X** platform with X as X, Response and all the Y columns as Y. Then select **Density Ellipse** from the red triangle menu on the Bivariate title bar for each plot.

## Other JMP Simulators

The JMP Scripting Language (JSL) provides a powerful way to build custom simulators to explore statistical concepts. These simulators can also be packaged as add-ins, which can be added to menus in JMP.

Here's a summary of where you can find other JSL-based simulators:

- Simulators for teaching and exploring statistical concepts are available in **Help > Sample Data > Teaching Scripts > Teaching Demonstrations**.

- A family of more comprehensive simulators, the Interactive Teaching Modules, are also found in **Help > Sample Data > Teaching Scripts**.
- To access the complete set of built-in simulators and teaching scripts, go to **Help > Sample Data > Open the Sample Scripts Directory**.
- You can find additional simulators in the JMP User Community at <http://community.jmp.com>.

## Exercises

1. Use the Central Limit Theorem simulation to explore the distribution of sample means for highly skewed data.
  - (a) Add 100 rows to the data table. Each row contains the mean for the sample size specified in the column name. So, column N=1 contains individual values, and column N=100 has means for samples of size 100.
  - (b) Use the Distribution platform to plot the distributions of the five columns.
  - (c) Describe the shape of each distribution. Specifically, what happens to the shape of the distributions as the sample size increases?
  - (d) Describe the variability, or spread, of each distribution. What happens to the spread of the distribution as the sample size increases?
2. Open the `confidence.jsl` script, and explore what happens to the width of confidence intervals as the sample size and confidence level are changed.
  - (a) Use different values for the sample size (that is, 5, 10, 50, and 100). What happens to the widths of the confidence intervals as the sample size changes?
  - (b) Change the confidence intervals (the confidence level) to different values (that is, 0.8, 0.9, and 0.99). What happens to the widths of the confidence intervals as the confidence level changes? How does the percentage captured by the true mean change? Conversely, how does this impact the number of times the intervals miss the true mean?

- (c) Open the Confidence Intervals for the Population Mean teaching module from **Help > Sample Data > Teaching Scripts > Interactive Teaching Modules**. Repeat steps a and b above. For this exercise, the process variable is IQ, the population mean is 100, and the standard deviation is 15.



# 7

## Univariate Distributions: One Variable, One Sample

### Overview

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

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

## Chapter Contents

Overview .....	125
Looking at Distributions .....	128
Probability Distributions .....	130
True Distribution Function or Real-World Sample Distribution .....	131
The Normal Distribution .....	133
Describing Distributions of Values .....	134
Generating Random Data .....	134
Histograms .....	135
Stem-and-Leaf Plots .....	137
Dot Plots .....	138
Outlier and Quantile Box Plots .....	139
Mean and Standard Deviation .....	141
Median and Other Quantiles .....	142
Mean versus Median .....	142
Other Summary Statistics: Skewness and Kurtosis .....	143
Extremes, Tail Detail .....	143
Statistical Inference on the Mean .....	144
Standard Error of the Mean .....	144
Confidence Intervals for the Mean .....	144
Testing Hypotheses: Terminology .....	147
The Normal z-Test for the Mean .....	149
Case Study: The Earth's Ecliptic .....	150
Student's t-Test .....	152
Comparing the Normal and Student's t Distributions .....	153
Testing the Mean .....	154
The p-Value Animation .....	155
Power of the t-Test .....	157
Practical Significance versus Statistical Significance .....	159
Examining for Normality .....	161
Normal Quantile Plots .....	162
Statistical Tests for Normality .....	165
Special Topic: Practical Difference .....	167

Special Topic: Simulating the Central Limit Theorem .....	170
Seeing Kernel Density Estimates .....	172
Exercises.....	173

# Looking at Distributions

Let's examine some actual data and start noticing aspects of its distribution.

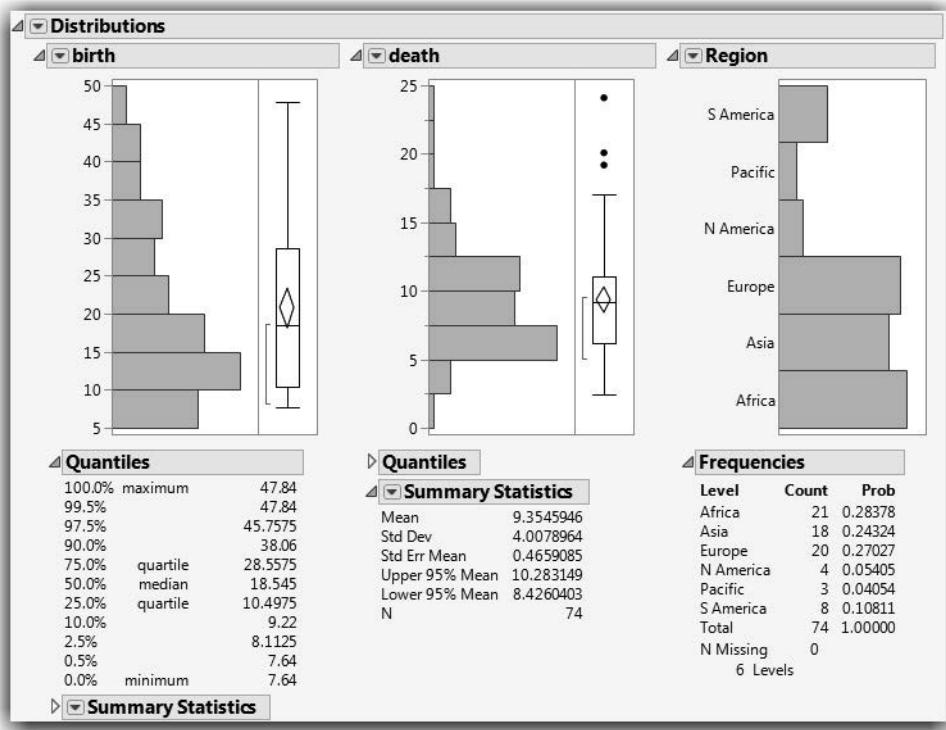
- ❖ Begin by selecting **Help > Sample Data Library** and opening Birth Death.jmp, which contains the 2009 birth and death rates of 74 nations (**Figure 7.1**).
- ❖ From the main menu bar, select **Analyze > Distribution**.
- ❖ On the Distribution launch window, assign the birth, death, and Region columns to **Y, Columns**, and then click **OK**.

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

	country	birth	death	Region
1	AFGHANISTAN	45.46	19.18	Asia
2	ALGERIA	16.90	4.64	Africa
3	ANGOLA	43.69	24.08	Africa
4	ARGENTINA	17.94	7.41	S America
5	AUSTRALIA	12.47	6.74	Pacific
6	AUSTRIA	8.65	9.98	Europe
7	BANGLADESH	24.68	9.23	Asia
8	BELGIUM	10.15	10.44	Europe
9	BRAZIL	18.43	6.35	S America
10	BULGARIA	9.51	14.31	Europe

When you see the report (**Figure 7.2**), be adventurous: scroll around and click in various places on the surface of the report. You can also right-click in plots and reports for additional options. Notice that histograms and statistical tables can be opened or closed by clicking the disclosure icon on the title bars.

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

**Figure 7.2** Histograms, Quantiles, Summary Statistics, and Frequencies

Note that there are two types of analyses:

- The analyses for **birth** and **death** are for continuous distributions. Quantiles and Summary Statistics are examples of reports that you get when the column in the data table has the continuous modeling type. The next to the column name in the Columns panel of the data table indicates that this variable is continuous.
- The analysis for **Region** is for a categorical distribution. A frequency report is an example of the type of report you get when the column in the data table has the modeling type of nominal or ordinal. or appears next to the column name in the Columns panel.

You can click on the icon and change the modeling type of any variable in the Columns panel to control which type of report you get. You can also right-click on the modeling type icon in any platform launch window to change the modeling type and redo an analysis. This changes the modeling type in the Columns panel as well.

For continuous distributions, the graphs give a general idea of the shape of the distribution. The death data cluster together with most values near the center. Distributions like this one, with one peak, are called *unimodal*. The birth data have a different distribution. There are more countries with low birth rates, with fewer countries gradually tapering toward higher birth rates. This distribution is *skewed* toward the higher rates.

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

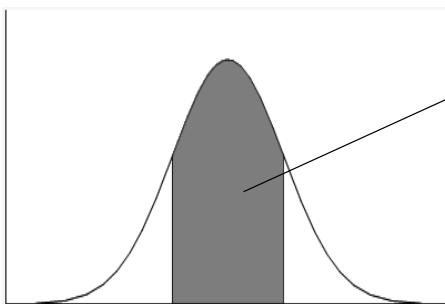
- *Quantiles* are the points at which various percentages of the total sample are above or below.
- *Summary Statistics* combine the individual data points to form descriptions of the entire data set. Two common summary statistics are the *mean* and *standard deviation*.

The report for the categorical distribution focuses on frequency counts. This chapter concentrates on continuous distributions and postpones the discussion of categorical distributions until Chapter 11, “Categorical Distributions.”

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

## Probability Distributions

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

**Figure 7.3** Continuous Distribution

The probability of being in a given interval is the proportion of the area under the density curve over that interval.

The density function describes all possible values of the random variable, so the area under the whole density curve must be 1, representing 100% probability. In fact, this is a defining characteristic of all density functions. In order for a function to be a density function, it must be nonnegative and the area underneath the curve must be 1.

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

## True Distribution Function or Real-World Sample Distribution

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

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

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

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

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

If you take a collection of values from a random process, sum them, and divide by the number of them, you have calculated a mean. You can then calculate the variance associated with this single number. There is a simple algebraic relationship between the variability of the responses (the standard deviation of the original data) and the variability of the sum of the responses divided by  $n$  (the standard error of the mean). Complete details follow in the section “Standard Error of the Mean” on page 144.

**Table 7.1.** Properties of Distribution Functions and Samples

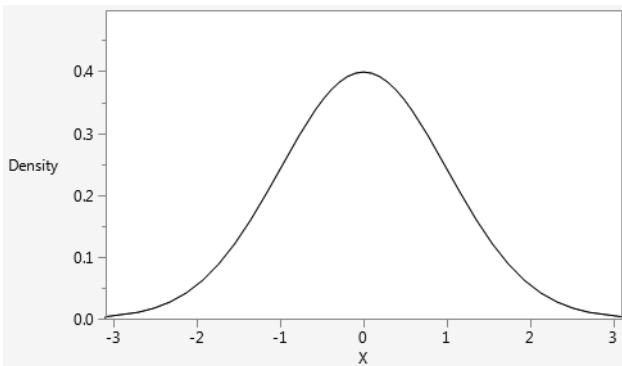
Concept	Abstract mathematical form, probability distribution	Numbers from the real world, data, sample
Mean	Expected value or true mean, the point that balances each side of the density	Sample mean, the sum of values divided by the number of values
Median	Median, the mid-value of the density area, where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

The statistic from the real world data is an estimate of the parameter from the distribution.

## The Normal Distribution

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

**Figure 7.4** Standard Normal Density Curve



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

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

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

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

## Describing Distributions of Values

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

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

## Generating Random Data

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

In Plato's theory of forms, the “true” world to be an ideal form. What you perceive as real data are only shadows that give hints at what the true data are like. Most of the time the true state is unknown, so an experience where the true state is known is valuable.

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

To create your own random data:

- ❖ Open RandDist.jmp. (Select **Help > Sample Data** and click the **Simulations** outline).

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

- ~ Select **Rows > Add Rows** and type 1000 to see a table like the one shown in **Figure 7.5**.

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

**Figure 7.5** Partial Listing of the RandDist.jmp Data Table

	Uniform	Normal	Exponential	Dbl Expon
1	0.643114	-0.40693	0.225066	0.469991
2	0.436727	-0.65254	0.093061	0.077219
3	0.82906	1.465737	0.166295	-0.5755
4	0.36994	0.406438	0.116109	-0.0719
5	0.544946	0.911773	1.229372	-2.21089
6	0.018742	1.389419	1.77889	-1.01573
7	0.878576	-0.30944	4.19619	0.932217
8	0.44195	1.538281	0.056345	0.211887
9	0.887352	-0.88501	0.036427	0.096371

- ~ To look at the distributions of the columns in the RandDist.jmp table, select **Analyze > Distribution**.
- ~ In the Distribution launch window, assign the four columns to **Y, Columns**, select **Histograms Only**, and then click **OK**.

The analysis automatically shows a number of graphs and statistical reports. To see further graphs and reports (**Figure 7.6**, for example), select an option from the red triangle menu for each analysis. The following sections examine the graphs and the text reports available in the Distribution platform.

## Histograms

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

Try the following histogram features:

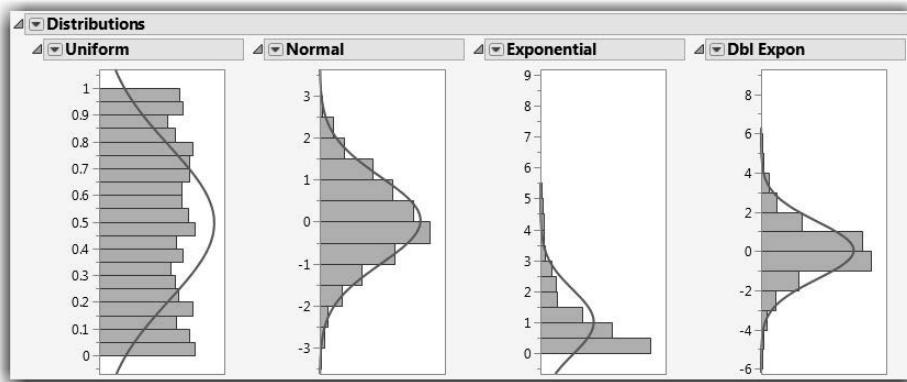
- ~ Click in a histogram bar.

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

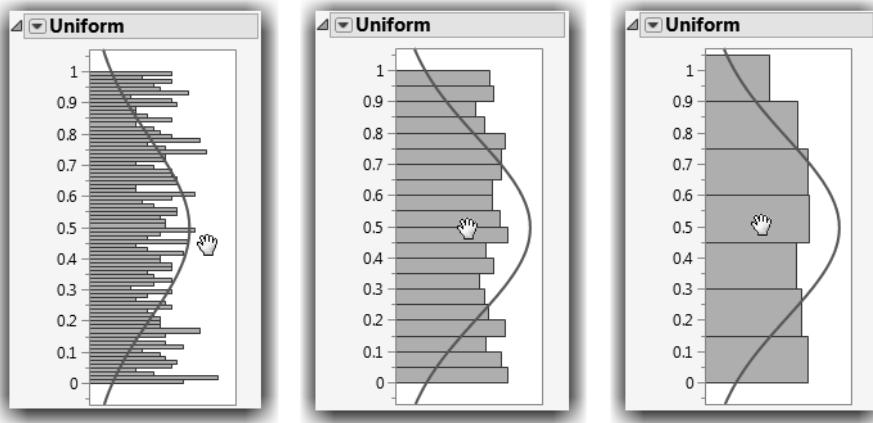
- ☞ Double-click on a histogram bar to produce a new JMP table that is a subset corresponding to that bar.
- ☞ Go back to the Distribution plots. For any histogram, select the **Normal** option from the **Continuous Fit** command (**Continuous Fit > Normal**) on the red triangle menu.

This superimposes over the histogram the normal density that corresponds to the mean and standard deviation in your sample. **Figure 7.6** shows the four histograms with normal curves superimposed on them.

**Figure 7.6** Histograms of Various Continuous Distributions



- ☞ Click the hand tool from the **Tools** menu or toolbar.
- ☞ Drag the Uniform histogram to the right, and then back to the left to see the histogram bars get narrower and wider (**Figure 7.7**).

**Figure 7.7** The Hand Tool Adjusts Histogram Bar Widths

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

Keep this data table open. You will use it later.

## Stem-and-Leaf Plots

A *stem-and-leaf plot* is a variation on the histogram. It was developed for tallying data in the days when computers were rare and histograms took a lot of time to make. Each line of the plot has a stem value that is the leading digits of a range of column values. The leaf values are made from other digits of the values. As a result, the stem-and-leaf plot has a shape that looks similar to a histogram, but also shows the data points themselves.

To see two examples, select **Help > Sample Data Library** and open the Big Class.jmp and Automess.jmp sample data tables.

- ❖ For each table, select **Analyze > Distribution**. On the launch window, the **Y, Columns** variables are weight from the Big Class.jmp sample data table and Auto theft from the Automess.jmp sample data table.
- ❖ When the histograms appear, select **Stem and Leaf** from the red triangle menu next to the histogram names.

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

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

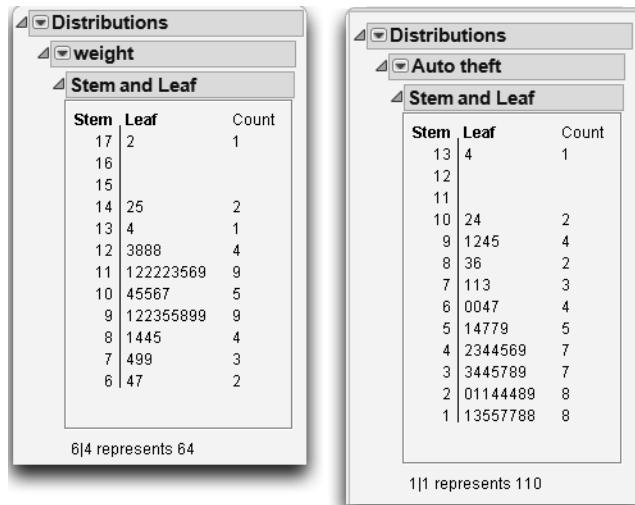
You can reconstruct the data values by joining the stem and leaf as indicated by the legend on the bottom of the plot. For example, on the bottom line of the weight plot, the values correspond to 64 and 67 (6 from the stem, 4 and 7 from the leaf). At the top, the weight is 172 (17 from the stem, 2 from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

The leaves respond to mouse clicks.

- ❖ Click on the two 5s on the bottom stem of the Auto theft plot. Hold down the Shift key to select more than one value at a time.

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

**Figure 7.8** Examples of Stem-and-Leaf Plots



## Dot Plots

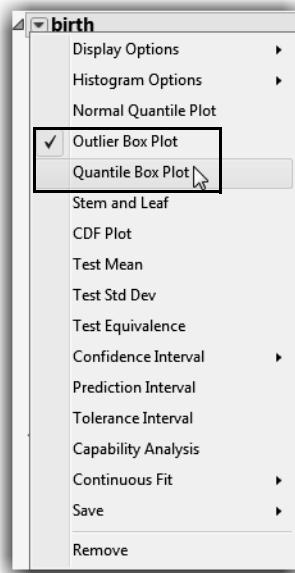
Dot plots are a variation on the histogram. Like a histogram, dot plots show how many values fall within an interval. But, instead of displaying bars, dots are drawn to represent each observation.

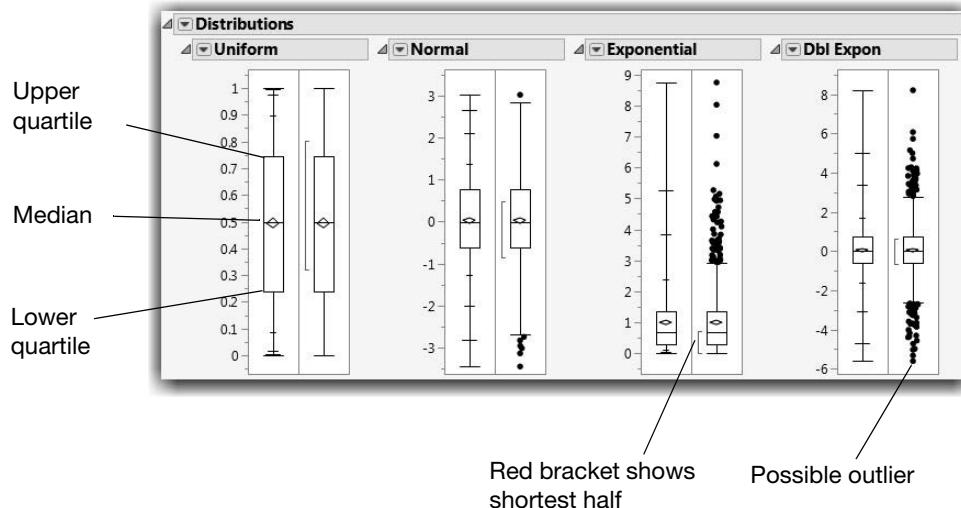
Because dot plots are primarily a teaching tool, they are available in JMP as a teaching script. To access the dot plot script, select **Help > Sample Data > Teaching Script > Teaching Demonstrations** and select Dot Plot.

## Outlier and Quantile Box Plots

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

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

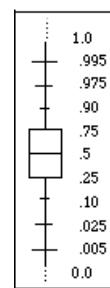


**Figure 7.9** Quantile and Outlier Box Plots

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

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

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



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

In both box plots, the mean and its 95% confidence interval are shown by a diamond. Since this experiment was created with 1,000 observations, the mean is

estimated with great precision, giving a very short confidence interval, and thus a thin diamond. Confidence intervals are discussed in the following sections.

## Mean and Standard Deviation

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

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum \frac{x_i}{n}$$

The sample mean has these properties:

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

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

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

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

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

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

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

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

## Median and Other Quantiles

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

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

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

## Mean versus Median

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

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

Suppose you have a skewed distribution, like household income in the United States. This set of data has lots of extreme points on the high end, but is limited to zero on the low end. If you want to know the income of a typical person, it makes more sense to report the median than the mean. However, if you want to track

per-capita income as an aggregating measure, then the mean income might be better to report.

## Other Summary Statistics: Skewness and Kurtosis

Certain summary statistics, including the mean and variance, are also called *moments*. Moments are statistics that are formed from sums of powers of the data's values. The first four moments are defined as follows:

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

Skewness and kurtosis can help determine whether a distribution is normal and, if not, what the distribution might be. A problem with these higher order moments is that the statistics have higher variance and are more sensitive to outliers.

 To get the skewness and kurtosis, select **Display Options > Customize Summary Statistics** from the red triangle menu next to the histogram's title. The same command is in the red triangle menu next to Summary Statistics.

## Extremes, Tail Detail

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

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

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

helpful, since the extremes are often the most informative statistics on the distribution.

## Statistical Inference on the Mean

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

### Standard Error of the Mean

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

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

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

Symbolically, this is written

$$s_{\bar{y}} = \frac{s_y}{\sqrt{n}}$$

where  $s_y$  is the sample standard deviation.

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

### Confidence Intervals for the Mean

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

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

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

- The data are normal, and
- The true standard deviation is the sample standard deviation. (We revisit this assumption later.)

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

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

Then, scale this interval by the standard deviation and add in the true mean:  
 $\mu \pm 1.96s_{\bar{y}}$ .

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

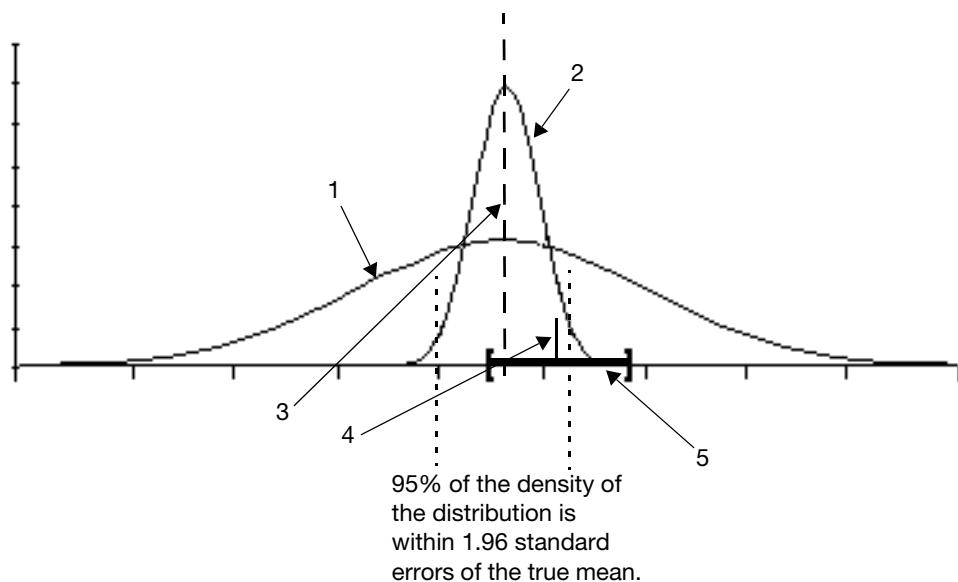
$$95\% \text{ C.I. for the mean} = \bar{x} \pm 1.96s_{\bar{y}}$$

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

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

The alpha ( $\alpha$ ) in the formula is the probability that the interval does not capture the true mean. That probability is 0.05 for a 95% interval. The Summary Statistics table reports the confidence interval as the Upper 95% Mean and Lower 95% Mean. It is represented in the quantile box plot by the ends of a diamond (see **Figure 7.11**).

**Figure 7.10** Illustration of Confidence Interval

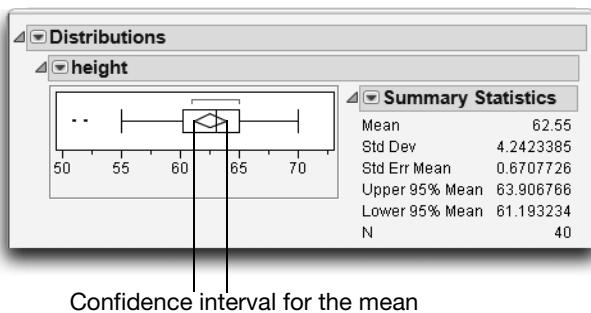


**Legend:**

1. This is the distribution of the process that makes the data with a standard deviation  $s$ .
2. This is the distribution of the mean, whose standard deviation is  $\frac{s}{\sqrt{n}}$ .
3. This is the true mean of the distribution.
4. We happened to get this mean from our random sample.

5. So, if we make an interval of 1.96 standard errors around the sample mean, we expect it to capture the true mean 95% of the time.

**Figure 7.11** Summary Statistics Report and Quantile Box Plot



If you have not done so, you should read the section “Confidence Intervals” on page 120 in Chapter 6, “Simulations,” and run the associated script.

## Testing Hypotheses: Terminology

Suppose you want to test whether the mean of a collection of sample values is significantly different from a hypothesized value. The strategy is to calculate a statistic so that if the true mean were the hypothesized value, getting such a large computed statistic value would be an extremely unlikely event. You would rather believe the hypothesis to be false than to believe that this rare coincidence happened. This is a probabilistic version of *proof by contradiction*.

The way you see an event as rare is to see that its probability is past a point in the tail of the probability distribution of the hypothesis. Often, researchers use 0.05 as a significance indicator. This means you believe that the mean is different from the hypothesized value if the chance of being wrong is only 5% (one in twenty).

Statisticians have a precise and formal terminology for hypothesis testing:

- The possibility of the true mean being the hypothesized value is called the *null hypothesis*. This is frequently denoted  $H_0$ , and is the hypothesis that you want to reject. Said another way, the null hypothesis is that the hypothesized value is not different from the true mean. The *alternative hypothesis*, denoted  $H_A$ , is that the mean is different from the hypothesized value. This can be phrased as greater than, less than, or unequal. The latter is called a *two-sided alternative*.

- The situation where you reject the null hypothesis when it happens to be true is called a *Type I error*. This declares that the difference is nonzero when it is really zero. The opposite mistake (not detecting a difference when there is a difference) is called a *Type II error*.
- The probability of getting a Type I error in a test is called the *alpha-level* ( $\alpha$ -level) of the test. This is the probability that you are wrong if you say that there is a difference. The *beta-level* ( $\beta$ -level) or *power* of the test is the probability of being right when you say that there is a difference.  $1 - \beta$  is the probability of a Type II error.
- Statistics and tests are constructed so that the power is maximized subject to the  $\alpha$ -level being maintained.

In the past, people obtained critical values for  $\alpha$ -levels and ended with a reject or don't-reject decision based on whether the statistic was bigger or smaller than the critical value. For example, researchers would declare that their experiment was significant if the test statistic fell in the region of the distribution corresponding to an  $\alpha$ -level of 0.05. This  $\alpha$ -level was specified in advance, before the study was conducted.

Computers have changed this strategy. Now, the  $\alpha$ -level isn't pre-determined, but rather is produced by the computer after the analysis is complete. In this context, it is called a *p-value* or *significance level*. The definition of a *p-value* can be phrased in many ways:

- The *p-value* is the  $\alpha$ -level at which the statistic would be significant.
- The *p-value* is how unlikely getting so large a statistic would be if the true mean were the hypothesized value.
- The *p-value* is the probability of being wrong if you rejected the null hypothesis. It is the probability of a Type I error.
- The *p-value* is the area in the tail of the distribution of the test statistic under the null hypothesis.

The *p-value* is the number that you want to be very small, certainly below 0.05, so that you can say that the mean is significantly different from the hypothesized value. The *p*-values in JMP are labeled according to the test statistic's distribution. *p*-values below 0.05 are marked with an asterisk in many JMP reports. The label “*Prob >|t|*” is read as the “probability of getting an even greater absolute *t* statistic, given that the null hypothesis is true.”

## The Normal z-Test for the Mean

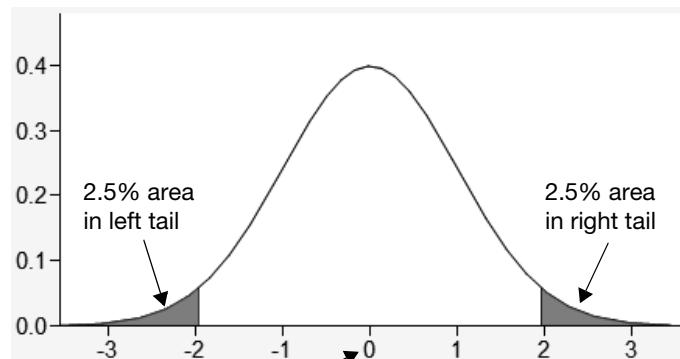
The Central Limit Theorem tells us that if the original response data are normally distributed, then, when many samples are drawn, the means of the samples are normally distributed. More surprisingly, it says that even if the original response data are not normally distributed, the sample mean still has an approximate normal distribution if the sample size is large enough. So the normal distribution provides a reference to use to compare a sample mean to a hypothesized value.

The standard normal distribution has a mean of zero and a standard deviation of one. You can center any variable to mean zero by subtracting the mean (even the hypothesized mean). You can standardize any variable to have standard deviation 1 (“unit standard deviation”) by dividing by the true standard deviation, assuming for now that you know what it is. This process is called *centering and scaling*. If the hypothesis were true, the test statistic that you construct should have this standard distribution. Tests using the normal distribution constructed like this (hypothesized mean but known standard deviation) are called *z-tests*. The formula for a *z*-statistic is

$$\text{z-statistic} = \frac{\text{estimate} - \text{hypothesized estimate}}{\text{standard deviation}}$$

You want to find out how unusual your computed *z*-value is from the point of view of believing the hypothesis. If the value is too improbable, then you doubt the null hypothesis.

To get a significance probability, you take the computed *z*-value and find the probability of getting an even greater absolute value. This involves finding the areas in the tails of the normal distribution that are greater than absolute *z* and less than negative absolute *z*. **Figure 7.12** illustrates a two-tailed *z*-test for  $\alpha = 0.05$ .

**Figure 7.12** Illustration of the Two-Tailed  $z$ -test

The null hypothesis is that the test statistic has mean 0 and standard deviation 1.

So, if the test statistic falls into one of the extreme regions, and the null hypothesis is true, you have a rare event that happens only 5% of the time. We therefore tend to doubt the null hypothesis.

## Case Study: The Earth's Ecliptic

In 1738, the Paris observatory determined with high accuracy that the angle of the earth's spin was 23.472 degrees. However, someone suggested that the angle changes over time. Examining historical documents found five measurements dating from 1460 to 1570. These measurements were somewhat different from the Paris measurement, and they were done using much less precise methods. The question is whether the differences in the measurements can be attributed to the errors in measurement of the earlier observations, or whether the angle of the earth's rotation actually changed. We need to test the hypothesis that the earth's angle has actually changed.

- ☛ Select **Help > Sample Data Library** and open Cassub.jmp (Stigler, 1986).
- ☛ Select **Analyze > Distribution** and assign Obliquity to **Y, Columns**.
- ☛ Click **OK**.

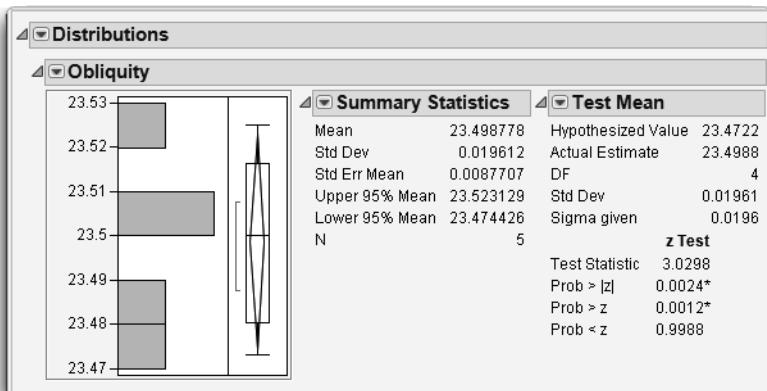
The Distribution report in **Figure 7.13** shows a histogram of the five values.

We now want to test that the mean of these values is different from the value from the Paris observatory. Our null hypothesis is that the mean is not different.

- ☛ Click on the red triangle menu next to Obliquity and select **Test Mean**.

- >Type the hypothesized value of 23.47222 (the value measured by the Paris observatory), and enter the standard deviation of 0.0196 found in the Summary Statistics table (we assume this is the true standard deviation).
- Click **OK**.

**Figure 7.13** Report of Observed Ecliptic Values



**Note:** Keep this data table open. You will use it later.

The  $z$ -test statistic has the value 3.0298. The area under the normal curve to the right of this value is reported as  $\text{Prob} > z$ . This is the probability ( $p$ -value) of getting an even greater  $z$ -value if there was no difference. In this case, the  $p$ -value is 0.0012. This is an extremely small  $p$ -value. If our null hypothesis were true (for example, the measurements were the same), our measurement would be a highly unlikely observation. Rather than believe the unlikely result, we reject  $H_0$  and claim the measurements are different.

Notice that, here, we are interested only in whether the mean is greater than the hypothesized value. We therefore look at the value of  $\text{Prob} > z$ , a one-sided test. Our null hypothesis stated above is that the mean is not different, so we test that the mean is different in either direction. For this two-sided test, we need the area in both tails. This statistic is two-sided and listed as  $\text{Prob} > |z|$ , in this case 0.0024.

The one-sided test  $\text{Prob} < z$  has a  $p$ -value of 0.9988, indicating that you are not going to prove that the mean is less than the hypothesized value. The two-sided  $p$ -value is always twice the smaller of the one-sided  $p$ -values.

## Student's *t*-Test

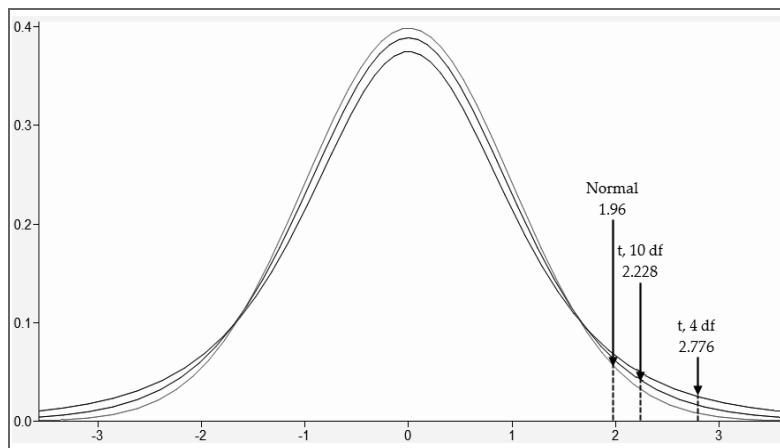
The *z*-test has a restrictive requirement. It requires the value of the true standard deviation of the response, and thus the standard deviation of the mean estimate, be known. Usually, this true standard deviation value is unknown and you have to use an estimate of the standard deviation.

Using the estimate in the denominator of the statistical test computation requires an adjustment to the distribution that was used for the test. Instead of using a normal distribution, statisticians use a *Student's t-distribution*. The statistic is called the *Student's t-statistic* and is computed by the formula shown here.  $x_0$  is the hypothesized mean, and  $s$  is the sample standard deviation of the sample data. In words, you can say

$$t\text{-statistic} = \frac{\text{sample mean} - \text{hypothesized value}}{\text{standard error of the mean}}$$

A large sample estimates the standard deviation very well, and the Student's *t*-distribution is remarkably similar to the normal distribution, as illustrated in **Figure 7.14**. However, in this example there were only five observations.

There is a different *t*-distribution for each number of observations, indexed by a value called *degrees of freedom*. Degrees of freedom is the number of observations minus the number of parameters estimated in fitting the model. In this case, five observations minus one parameter (the mean) yields  $5 - 1 = 4$  degrees of freedom. As you can see in **Figure 7.14**, the quantiles for the *t*-distribution spread out farther than the normal when there are few degrees of freedom.

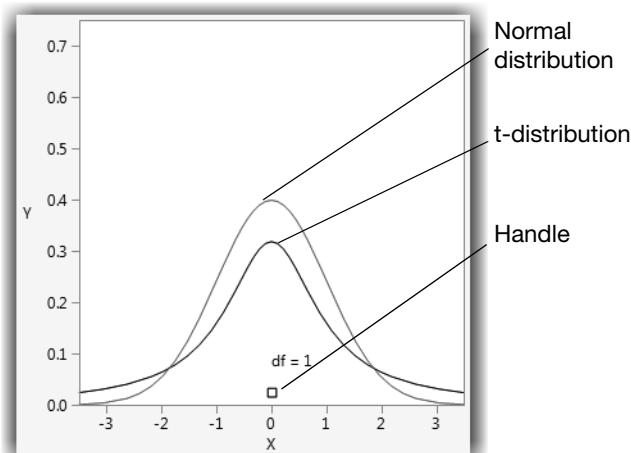
**Figure 7.14** Comparison of Normal and Student's *t* Distributions

## Comparing the Normal and Student's *t* Distributions

JMP can produce an animation to show you the relationships in **Figure 7.14**. This demonstration uses the `Normal vs. t.jsl` script.

- ☞ To run the script, select **Help > Sample Data > Teaching Scripts > Teaching Demonstrations > Normal vs. t.**

You should see the window shown in **Figure 7.15**.

**Figure 7.15** Normal vs *t* Comparison

The small square located just above 0 is called a *handle*. It is dragable, and adjusts the degrees of freedom associated with the black *t*-distribution as it moves. The normal distribution is drawn in red.

- ☞ Drag the handle up and down to adjust the degrees of freedom of the *t*-distribution.

Notice both the height and the tails of the *t*-distribution. At what number of degrees of freedom do you feel that the two distributions are close to identical?

## Testing the Mean

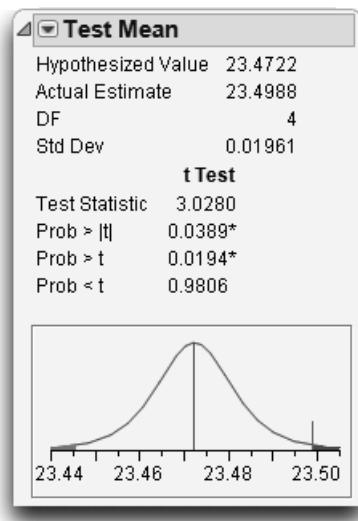
We now reconsider the ecliptic case study, so return to the Cassub.jmp Distribution of Obliquity report window. It turns out that for a 5% two-tailed test, the *t*-quantile for 4 degrees of freedom is 2.776, which is far greater than the corresponding *z*-quantile of 1.96 (shown in **Figure 7.14**). That is, the bar for rejecting  $H_0$  is higher, due to the fact that we don't know the standard deviation. Let's do the same test again, using this different value. Our null hypothesis is still that there is no change in the values.

- ☞ Select **Test Mean** and again enter 23.47222 for the hypothesized mean value. This time, do not fill in the standard deviation.
- ☞ Click **OK**.

The Test Mean report (shown here) now displays a *t*-test instead of a *z*-test (as in the Obliquity report in **Figure 7.13** on page 151).

When you don't specify a standard deviation, JMP uses the sample estimate of the standard deviation. The significance is smaller, but the *p*-value of 0.0389 still looks convincing, so you can reject  $H_0$  and conclude that the angle has changed. When you have a significant result, the idea is that under the null hypothesis, the expected value of the *t*-statistic is zero. It is highly unlikely (probability less than  $\alpha$ ) for the *t*-statistic to be so far out in the tails.

Therefore, you don't put much belief in the null hypothesis.



**Note:** You might have noticed that the test window offers the options of a Wilcoxon signed-rank nonparametric test. Some statisticians favor nonparametric tests because the results don't depend on the response having a normal distribution. Nonparametric tests are covered in more detail in Chapter 9, "Comparing Many Means: One-Way Analysis of Variance."

## The *p*-Value Animation

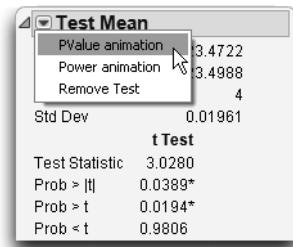
Figure 7.12 on page 150 illustrates the relationship between the two-tailed test and the normal distribution. Some questions might arise after looking at this picture.

- How would the *p*-value change if the difference between the truth and my observation were different?
- How would the *p*-value change if my test were one-sided instead of two sided?
- How would the *p*-value change if my sample size were different?

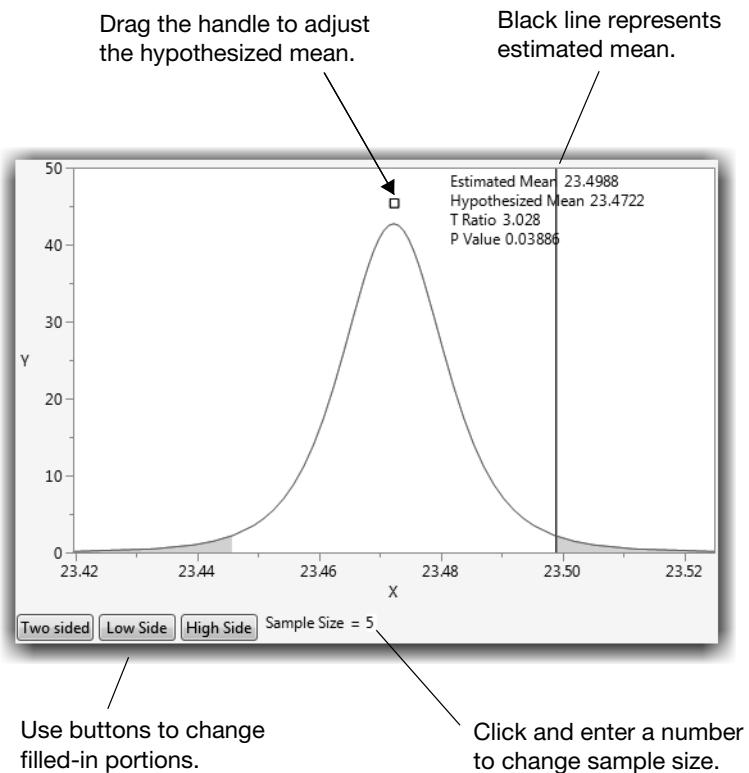
To answer these questions, JMP provides an animated demonstration, written in JMP scripting language. Often, these scripts are stored as separate files or are included in the Sample Scripts folder.

However, some scripts are built into JMP. This *p*-value animation is an example of a built-in script.

- ☞ Select **PValue animation** from the red triangle menu next to **Test Mean**, as shown here.



The *p* value animation script produces the window in Figure 7.16.

**Figure 7.16** *p*-Value Animation Window for the Ecliptic Case Study

The black vertical line represents the mean estimated by the historical measurements. You can drag the handle around the window. In this case, the handle represents the true mean under the null hypothesis. To reject this true mean, there must be a significant difference between it and the mean estimated by the data.

The *p*-value calculated by JMP is affected by the difference between this true mean and the estimated mean. You can see the effect of a different true mean by dragging the handle.

- ☞ Drag the handle left and right. Observe the changes in the *p*-value as the true mean changes.

As expected, the *p*-value decreases as the difference between the true and hypothesized mean increases.

The effect of changing this mean is also illustrated graphically. As shown previously in **Figure 7.12**, the shaded area represents the region where the null hypothesis is rejected. As the area of this region increases, the  $p$ -value of the test also increases. This demonstrates that the closer your estimated mean is to the true mean under the null hypothesis, the less likely you are to reject the null hypothesis.

This demonstration can also be used to extract other information about the data. For example, you can determine the smallest difference that your data would be able to detect for specific  $p$ -values. To determine this difference for  $p = 0.10$ :

- ☞ Drag the handle until the  $p$ -value is as close to 0.10 as possible.

You can then read the estimated mean and hypothesized mean from the text display. The difference between these two numbers is the smallest difference that would be significant at the 0.10 level. Any smaller difference would not be significant.

To see the difference between  $p$ -values for two-sided and one-sided tests, use the buttons at the bottom of the window.

- ☞ Click the **High Side** button to change the test to a one-sided  $t$ -test.

The  $p$ -value decreases because the region where the null hypothesis is rejected has become larger. It is all piled up on one side of the distribution, so smaller differences between the true mean and the estimated mean become significant.

- ☞ Repeatedly click the **Two Sided** and **High Side** buttons.

What is the relationship between the  $p$ -values when the test is one- and two-sided? To edit and see the effect of different sample sizes:

- ☞ Click on the values for sample size beneath the plot and enter different values.

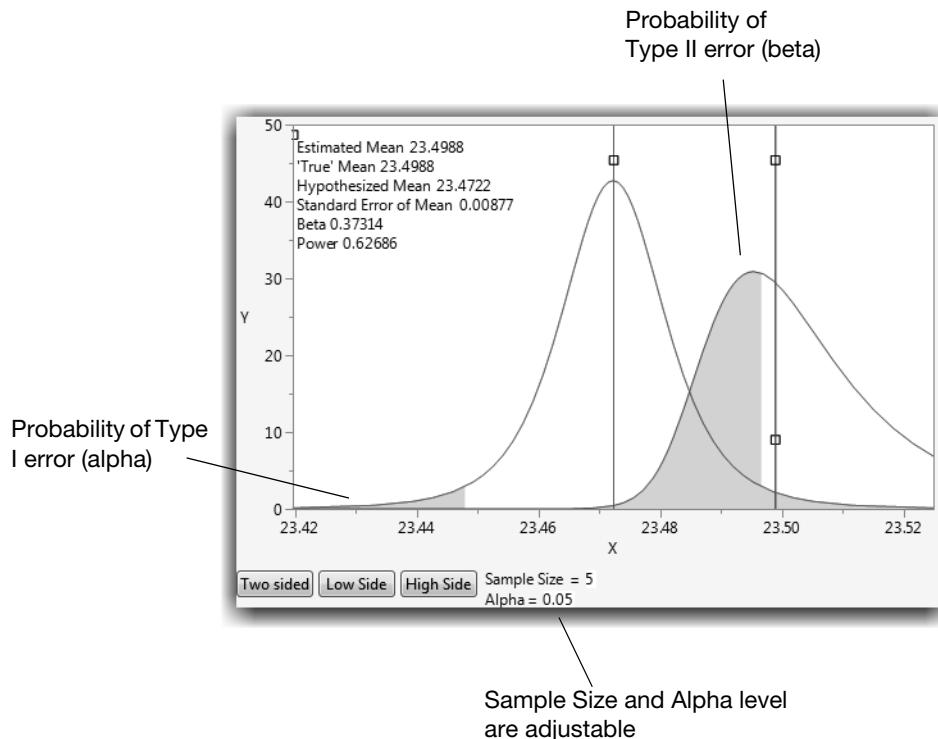
What effect would a larger sample size have on the  $p$ -value?

## Power of the $t$ -Test

As discussed in the section “Testing Hypotheses: Terminology” on page 147, there are two types of errors that a statistician is concerned with when conducting a statistical test—Type I and Type II. JMP contains a built-in script to graphically demonstrate the quantities involved in computing the power of a  $t$ -test.

- ✓ Select **Power animation** from the red triangle menu next to Test Mean to display the window shown in **Figure 7.17**.

**Figure 7.17** Power Animation Window



The probability of committing a Type I error (reject the null hypothesis when it is true), often represented by  $\alpha$ , is shaded in red. The probability of committing a Type II error (not detecting a difference when there is a difference), often represented as  $\beta$ , is shaded in blue. Power is  $1 - \beta$ , which is the probability of detecting a difference. The case where the difference is zero is examined below.

There are three handles in this window. One handle is for the estimated mean (calculated from the data). One handle is for the true mean (an unknowable quantity that the data estimates). The other handle is for the hypothesized mean (the mean assumed under the null hypothesis). You can drag these handles to see how their positions affect power.

**Note:** Click on the values for sample size and alpha beneath the plot to edit them.

- ☞ Drag the true mean (the top handle on the blue line) until it coincides with the hypothesized mean (the red line).

This simulates the situation where the true mean is the hypothesized mean in a test where  $\alpha=0.05$ . What is the power of the test?

- ☞ Continue dragging the true mean around the graph.

Can you make the probability of committing a Type II error (Beta) smaller than the case above, where the two means coincide?

- ☞ Drag the true mean so that it is far away from the hypothesized mean.

Notice that the shape of the blue distribution (around the true mean) is no longer symmetrical. This is an example of a *non central t-distribution*.

Finally, as with the *p*-value animation, these same situations can be further explored for one-sided tests using the buttons along the bottom of the window.

- ☞ Explore different values for sample size and alpha.

## Practical Significance versus Statistical Significance

This section demonstrates that a *statistically* significant difference can be quite different from a *practically* significant difference. Dr. Quick and Dr. Quack are both in the business of selling diets, and they have claims that appear contradictory. Dr. Quack studied 500 dieters and claims,

A statistical analysis of my dieters shows a statistically significant weight loss for my Quack diet.

Dr. Quick followed the progress of 20 dieters and claims,

A statistical study shows that on average my dieters lose more than three times as much weight on the Quick diet as on the Quack diet.

So which claim is right?

- ☞ To compare the Quick and Quack diets, select **Help > Sample Data Library** and open Diet.jmp.

**Figure 7.18** shows a partial listing of the Diet data table.

**Figure 7.18** Partial Listing of the Diet Data

	Quack's Weight Change	Quick's Weight Change
14	10.6	-8.2
15	3.9	-3.8
16	4.1	-10.2
17	-6.1	6.8
18	4.3	-7.9
19	15.7	-14.9
20	11.8	-2.2
21	12.4	.
22	-1.4	.
23	14	.
24	-20.6	.

- ⓐ Select **Analyze > Distribution**, assign both variables to **Y, Columns**, and then click **OK**.
- ⓑ Select **Test Mean** from the red triangle menu next to each histogram title bar to compare the mean weight loss for each diet to zero.

You should use the one-sided  $t$ -test because you are interested only in significant weight loss (not gain).

If you look closely at the means and  $t$ -test results in **Figure 7.19**, you can verify both claims!

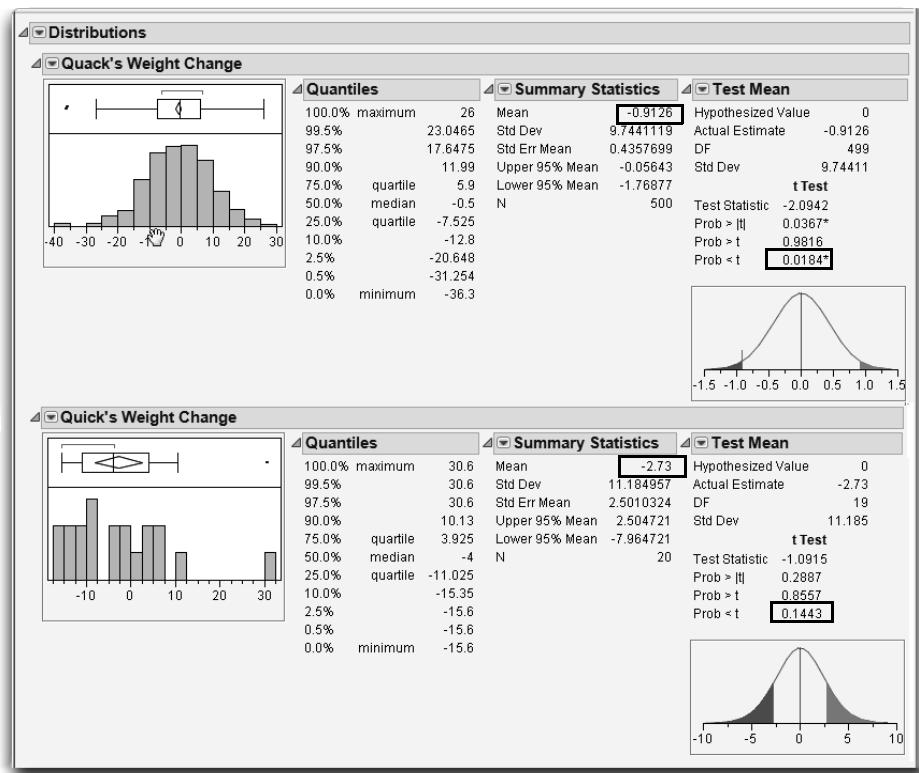
Quick's average weight loss of 2.73 is more than three times the 0.91 weight loss reported by Quack, and Quack's weight loss was significantly different from zero. However, Quick's larger mean weight loss was not significantly different from zero. Quack might not have a better diet, but the doctor has more evidence—500 cases compared with 20 cases. So even though the diet produced a weight loss of less than a pound, it is statistically significant. Significance is about evidence, and having a large sample size can make up for having a small effect.

**Note:** If you have a large enough sample size, even a very small difference can be significant. If your sample size is small, even a large difference might not be significant.

Looking closer at the claims, note that Quick reports on the estimated difference between the two diets, whereas Quack reports on the significance of his results. Both are somewhat empty statements. It is not enough to report an estimate without a measure of variability. It is not enough to report a significance without an estimate of the difference.

The best report in this situation is a confidence interval for the estimate, which shows both the statistical and practical significance. The next chapter presents the tools to do a more complete analysis on data like the Quick and Quack diet data.

**Figure 7.19 Reports of the Quick and Quack Example**



## Examining for Normality

Sometimes you might want to test whether a set of values is from a particular distribution. Perhaps you are verifying assumptions and want to test that the values are from a normal distribution.

## Normal Quantile Plots

*Normal quantile plots* show all the values of the data as points in a plot. If the data are normal, the points tend to follow a straight line.

- ⌚ Return to the four RandDist.jmp histograms that you opened in “Generating Random Data” on page 134.
- ⌚ Hold down the Ctrl key and select **Normal Quantile Plot** from one of the red triangle menus next to a histogram.

**Note:** Holding down the Ctrl key while selecting a command broadcasts that command to other analyses.

The histograms and normal quantile plots for the four simulated distributions are shown later in **Figure 7.21** and **Figure 7.22**.

The  $y$  (vertical) coordinate is the actual value of each data point. The  $x$  (horizontal) coordinate is the normal quantile associated with the rank of the value after sorting the data.

If you are interested in the details, the precise formula used for the normal quantile values is

$$\Phi^{-1}\left(\frac{r_i}{N+1}\right)$$

where  $r_i$  is the rank of the observation being scored,  $N$  is the number of observations, and  $\Phi^{-1}$  is the function that returns the normal quantile associated with the probability argument  $p$ , where  $p$  equals

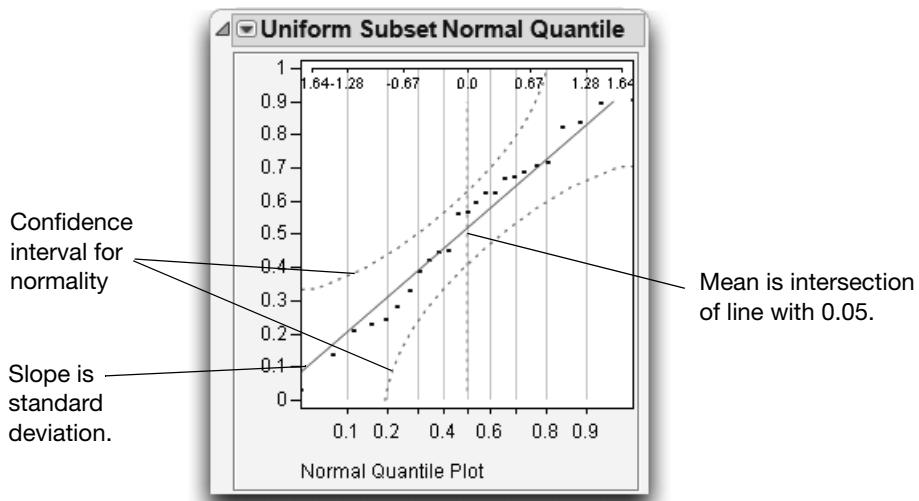
$$\frac{r_i}{N+1}$$

The normal quantile is the value on the  $x$ -axis of the normal density that has the portion  $p$  of the area below it. For example, the quantile for 0.5 (the probability of being less than the median) is 0.5, because half (50%) of the density of the standard normal is below 0.5. The technical name for the quantiles JMP uses is the *van der Waerden* normal scores; they are computationally cheap (but good) approximations to the more expensive, exact expected normal order statistics.

**Figure 7.20** shows the normal quantile plot with the following components:

- A red straight line, with confidence limits, shows where the points tend to lie if the data were normal. This line is purely a function of the sample mean and standard deviation. The line crosses the mean of the data at the normal quantile of 0.5. The slope of the line is the standard deviation of the data.
- Dashed lines surrounding the straight line form a confidence interval for the normal distribution. If the points fall outside these dashed lines, you are seeing a significant departure from normality.
- If the slope of the points is small (relative to the normal), then you are crossing a lot of (ranked) data with little variation in the real values. Therefore, you encounter a dense cluster. If the slope of the points is large, then you are crossing a lot of real values with few (ranked) points. Dense clusters make flat sections, and thinly populated regions make steep sections. (See upcoming figures for examples.)

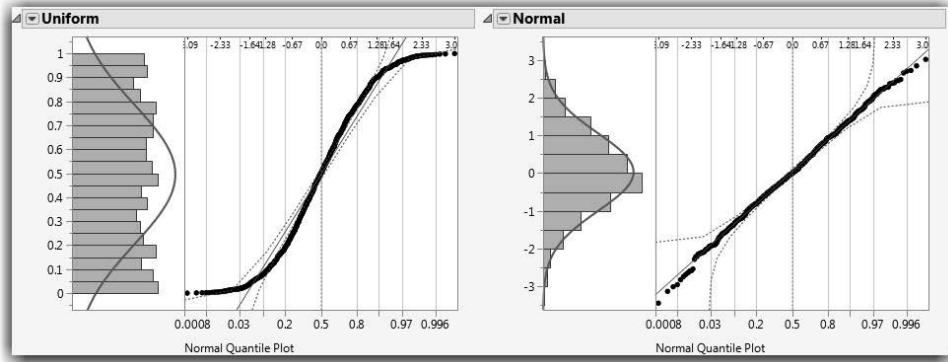
**Figure 7.20** Normal Quantile Plot Explanation



The middle portion of the uniform distribution (left plot in **Figure 7.21**) is steeper (less dense) than the normal. In the tails, the uniform is flatter (more dense) than the normal. In fact, the tails are truncated at the end of the range, where the normal tails extend infinitely.

The normal distribution (right plot in **Figure 7.21**) has a normal quantile plot that follows a straight line. Points at the tails usually have the highest variance and are most likely to fall farther from the line. Because of this, the confidence limits flair near the ends.

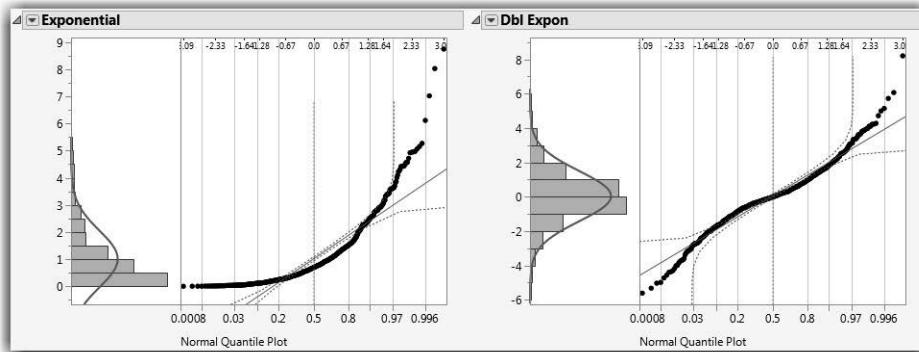
**Figure 7.21** Uniform Distribution (left) and Normal Distribution (right)



The exponential distribution (**Figure 7.22**) is skewed – that is, one-sided. The top tail runs steeply past the normal line; it spreads out more than the normal. The bottom tail is shallow and much denser than the normal.

The middle portion of the double exponential (**Figure 7.22**) is denser (more shallow) than the normal. In the tails, the double exponential spreads out more (is steeper) than the normal.

**Figure 7.22** Exponential Distribution and Double Exponential Distribution



## Statistical Tests for Normality

A widely used test that the data are from a specific distribution is the *Kolmogorov test* (also called the *Kolmogorov-Smirnov test*). The test statistic is the greatest absolute difference between the hypothesized distribution function and the empirical distribution function of the data. The empirical distribution function goes from 0 to 1 in steps of  $1/n$  as it crosses data values. When the Kolmogorov test is applied to the normal distribution and adapted to use estimates for the mean and standard deviation, it is called the *Lilliefors test* or the *KSL test*. In JMP, Lilliefors quantiles on the cumulative distribution function (cdf) are translated into confidence limits in the normal quantile plot. Therefore, you can see where the distribution departs from normality by where it crosses the confidence curves.

Another test of normality produced by JMP is the *Shapiro-Wilk test* (or the *W-statistic*), which is implemented for samples as large as 2000. For samples greater than 2000, the KSL (Kolmogorov-Smirnov-Lilliefors) test is done. The null hypothesis for this test is that the data are normal. Rejecting this hypothesis would imply the distribution is non-normal.

- ⓐ Look at the Birth Death.jmp data table again or re-open it if it is closed.
- ⓐ Select **Analyze > Distribution**, assign birth and death to **Y, Columns**, and then click **OK**.
- ⓐ Select **Fit Distribution > Continuous Fit > Normal** from the red triangle menu next to Birth.
- ⓐ Select **Goodness of Fit** from the red triangle menu next to Fitted Normal.
- ⓐ Repeat for the death distribution.

The results are shown in **Figure 7.23**.

The conclusion is that neither distribution is normal.

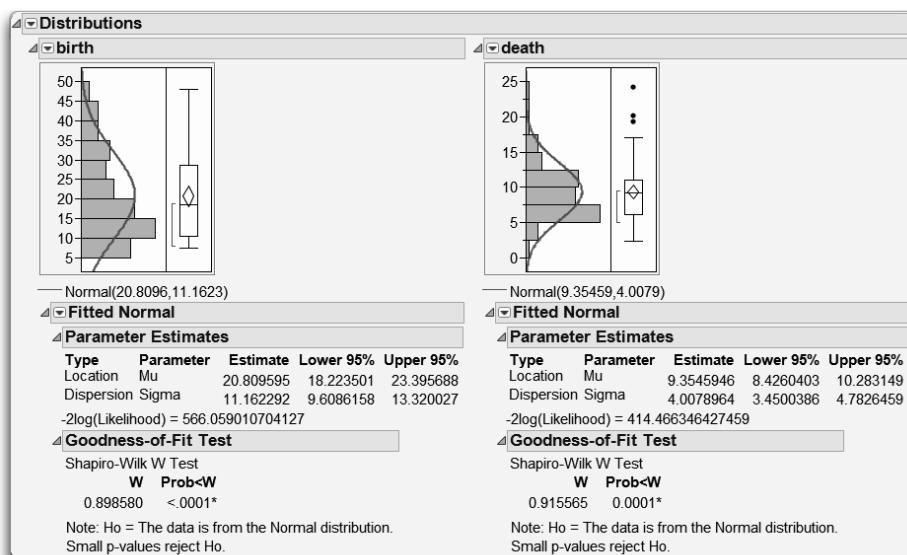
This is an example of an unusual situation where you hope the test fails to be significant, because the null hypothesis is that the data are normal.

If you have a large number of observations, you might want to reconsider this tactic. The normality tests are sensitive to small departures from normality. Small departures do not jeopardize other analyses because of the Central Limit Theorem, especially because they are also probably highly significant. All the

distributional tests assume that the data are independent and identically distributed.

Some researchers test the normality of residuals from model fits, because the other tests assume a normal distribution. We strongly recommend that you do not conduct these tests, but instead rely on normal quantile plots to look for patterns and outliers.

**Figure 7.23** Test Distributions for Normality



So far we have been doing statistics correctly, but a few remarks are in order.

- In most tests, the null hypothesis is something that you want to disprove. It is disproved by the contradiction of getting a statistic that would be unlikely if the hypothesis were true. But in normality tests, you want the null hypothesis to be true. Most testing for normality is to verify assumptions for other statistical tests.
- The mechanics for any test where the null hypothesis is desirable are backward. You can get an undesirable result, but the failure to get it does not prove the opposite—it only says that you have insufficient evidence to prove it is true. “Special Topic: Practical Difference” on page 167 gives more details about this issue.

- When testing for normality, it is more likely to get a desirable (inconclusive) result if you have very little data. Conversely, if you have thousands of observations, almost any set of data from the real world appears significantly non-normal.
- If you have a large sample, the estimate of the mean is distributed normally even if the original data is not. This result, from the Central Limit Theorem, is demonstrated in a later section beginning on page 170.
- The test statistic itself doesn't tell you about the nature of the difference from normality. The normal quantile plot is better for this.

## Special Topic: Practical Difference

Suppose you really want to show that the mean of a process is a certain value. Standard statistical tests are of no help. The failure of a test to show that a mean is *different* from the hypothetical value does not show that it *is* that value. It says only that there is not enough evidence to confirm that it isn't that value. In other words, saying "I can't say the result is different from 5" is not the same as saying "The result must be 5."

You can never show that a mean is exactly some hypothesized value, because the mean could be different from that hypothesized value by an infinitesimal amount. No matter the sample size, you might have a value that is different from the hypothesized mean by an amount that is so small that it is quite unlikely to get a significant difference even if the true difference is zero.

So instead of trying to show that the mean is exactly equal to a hypothesized value, you need to choose an interval around that hypothesized value and try to show that the mean is not outside that interval. This can be done.

There are many situations where you want to Ctrl a mean within some specification interval. For example, suppose that you make 20-amp electrical circuit breakers. You need to demonstrate that the mean breaking current for the population of breakers is between 19.9 and 20.1 amps. (Actually, you probably also require that most individual units be in some specification interval, but for now we just focus on the mean.) You'll never be able to prove that the mean of the population of breakers is exactly 20 amps. You can, however, show that the mean is close—within 0.1 of 20.

The standard way to do this is the *TOST method*, an acronym for Two One-Sided Tests (Westlake 1981, Schuirmann 1981, Berger and Hsu 1996):

1. First, you do a one-sided *t*-test that the mean is the low value of the interval, with an upper tail alternative.
2. Then you do a one-sided *t*-test that the mean is the high value of the interval, with a lower tail alternative.
3. If both tests are significant at some level  $\alpha$ , then you can conclude that the mean is outside the interval with a probability less than or equal to  $\alpha$ , the significance level. In other words, the mean is not practically different from the hypothesized value, or, in still other words, the mean is practically equivalent to the hypothesized value.

**Note:** Technically, the test works by a union intersection rule, whose description is beyond the scope of this book.

For example, a material coating process requires the mean coating weight to be  $20.4 \pm 0.2$  units.

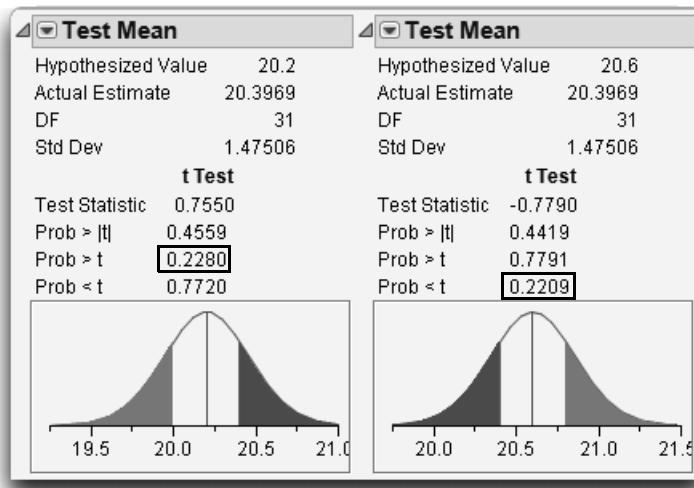
- ☞ Select **Help > Sample Data Library** and open Quality Ctrl/Coating.jmp.
- ☞ Select **Analyze > Distribution**, assign Weight to **Y, Columns**, and then click **OK**.

When the report appears,

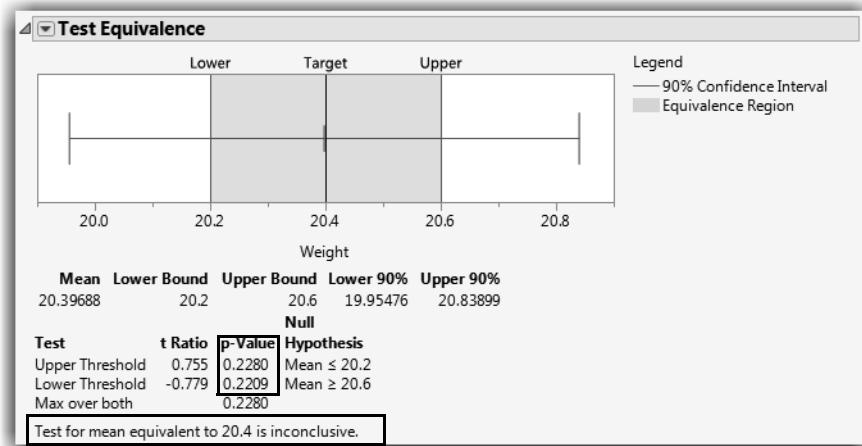
- ☞ Select **Test Mean** from the red triangle menu next to Weight, type 20.2 as the hypothesized value, and then click **OK**.
- ☞ Select **Test Mean** again, enter 20.6 as the hypothesized value, and then click **OK**.

This tests the null hypothesis that the mean Weight is between 20.2 and 20.6 (that is,  $20.4 \pm 0.2$ ) with a protection level ( $\alpha$ ) of 0.05.

The *p*-value for the hypothesis from below is approximately 0.228, and the *p*-value for the hypothesis from above is also about 0.22. Since both of these values are far above the  $\alpha$  of 0.05 that we were looking for, we declare it not significant. We cannot reject the null hypothesis. The conclusion is that we have not shown that the mean is practically equivalent to  $20.4 \pm 0.2$  at the 0.05 significance level. We need more data.

**Figure 7.24** Compare Test for Mean at Two Values

The Test Equivalence option in the Distribution platform applies the TOST method, directly conducting the two one-sided  $t$ -tests. You enter the hypothesized value, the threshold, and the confidence level.

**Figure 7.25** Test Equivalence

## Special Topic: Simulating the Central Limit Theorem

The Central Limit Theorem, which we visited in a previous chapter, says that for a very large sample size, the sample mean is very close to normally distributed, regardless of the shape of the underlying distribution. That is, if you compute means from many samples of a given size, the distribution of those means approaches normality, even if the underlying population from which the samples were drawn is not.

You can see the Central Limit Theorem in action using the template called **Central Limit Theorem.jmp**. in the sample data library.

- ⇨ Select **Help > Sample Data Library** and open **Central Limit Theorem.jmp**.
- ⇨ Click on the plus sign next to column **N=1** in the Columns panel to view the formula.
- ⇨ Do the same thing for the rest of the columns, called **N=5**, **N=10**, and so on, to look at their formulas (**Figure 7.26**).

**Figure 7.26** Formulas for Columns in the Central Limit Theorem Data Table

<input checked="" type="checkbox"/> Central Limit Theorem Notes The uniform**4 generates a highly skewed distribution.	Local $\left( \{j\}, \sum_{j=1}^1 \text{Random Uniform() }^4 \right)$	Local $\left( \{j\}, \sum_{j=1}^5 \text{Random Uniform() }^4 \right)$
<b>Columns (5/0)</b> <ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> N=1</li> <li><input checked="" type="checkbox"/> N=5</li> <li><input checked="" type="checkbox"/> N=10</li> <li><input checked="" type="checkbox"/> N=50</li> <li><input checked="" type="checkbox"/> N=100</li> </ul>	1	5
	Local $\left( \{j\}, \sum_{j=1}^{10} \text{Random Uniform() }^4 \right)$	Local $\left( \{j\}, \sum_{j=1}^{50} \text{Random Uniform() }^4 \right)$
	10	50
	Local $\left( \{j\}, \sum_{j=1}^{100} \text{Random Uniform() }^4 \right)$	
	100	

Looking at the formulas might help you understand what's going on. The expression raising the uniform random number values to the fourth power creates a highly skewed distribution. For each row, the first column, **N=1**,

generates a single uniform random number to the fourth power. For each row in the second column,  $N=5$ , the formula generates a sample of five uniform numbers, takes each to the fourth power, and computes the mean. The next column does the same for a sample size of 10, and the remaining columns generate means for sample sizes of 50 and 100.

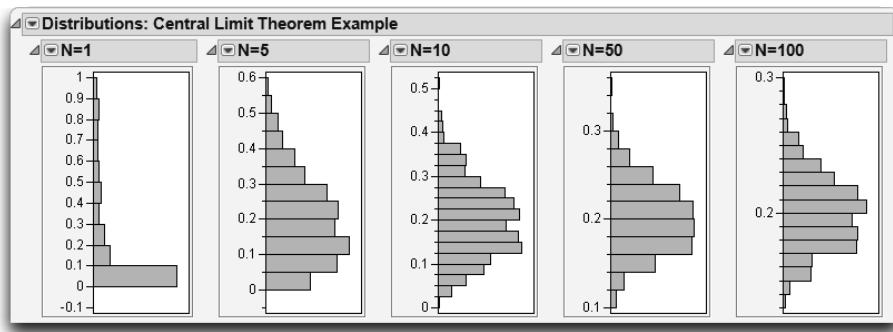
- ☞ Add 500 rows to the data table using **Rows > Add Rows**.

When the computations are complete:

- ☞ Select **Analyze > Distribution**. Select all the variables, assign them to **Y, Columns**, and then click **OK**.

Your results should be similar to those in **Figure 7.27**. When the sample size is only 1, the skewed distribution is apparent. As the sample size increases, you can clearly see the distributions becoming more and more normal.

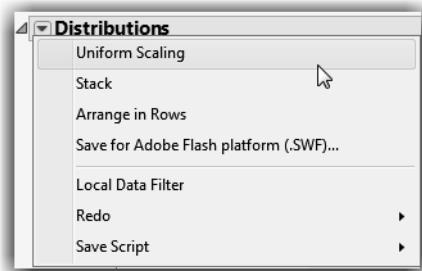
**Figure 7.27** Example of the Central Limit Theorem in Action



The distributions also become less spread out, since the standard deviation ( $s$ ) of a mean of  $n$  items is  $\frac{s}{\sqrt{n}}$ .

- ☞ To see this dramatic effect, select **Uniform Scaling** from the red triangle menu next to Distribution.

**Note:** The Sampling Distribution of Sample Means teaching module provides a more flexible interface for exploring the central limit theorem. The collection of teaching modules can be found under **Help > Sample Data > Teaching Scripts > Interactive Teaching Modules**.



## Seeing Kernel Density Estimates

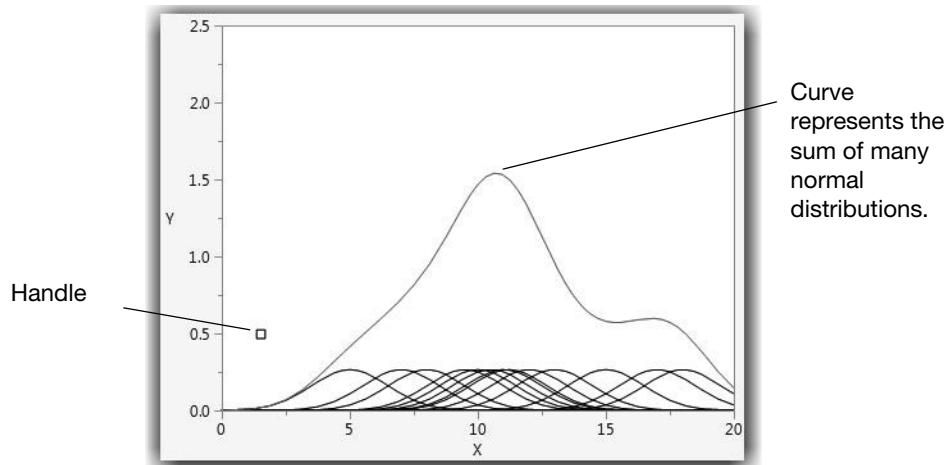
The idea behind kernel density estimators is not difficult. In essence, a normal distribution is placed over each data point with a specified standard deviation. Each of these normal distributions is then summed to produce the overall curve.

JMP can animate this process for a simple set of data. For details about using scripts, see “Working with Scripts” on page 60.

- ❖ Open the `demoKernel.jsl` script. Select **Help > Sample Data** and click the **Open Sample Scripts Directory** to see the sample scripts library.
- ❖ Select **Edit > Run Script** to run the `demoKernel` script.

You should see a window like the one in **Figure 7.28**.

**Figure 7.28** Kernel Addition Demonstration



The handle on the left side of the graph can be dragged with the mouse.

- ❖ Move the handle to adjust the spread of the individual normal distributions associated with each data point.

The larger red curve is the smoothing spline generated by the sum of the normal distributions. As you can see, merely adjusting the spread of the small normal distributions controls the smoothness of the spline fit.

## Exercises

1. The sample data table Hollywood Movies.jmp contains data for all Hollywood movies released in 2011. The data table contains the name of a movie, the amount of money that it made in the United States (Domestic) and in foreign markets (in millions of dollars), the movie genre, and other information.
  - (a) Create a histogram of the movie genres. What are the levels of this variable? How many of each level are in the data set?
  - (b) Create a histogram of the domestic gross for each movie. What is the range of values for this variable? What is the average domestic gross of these movies?
  - (c) Consider the histogram that you created in part (b) for the domestic gross of these movies. You should notice several outliers in the outlier box plot. Move your mouse over the points to identify the movies. Use your cursor to draw a box around these points. Then, use **Rows > Label** to label the points. What was the top grossing movie in 2011?
  - (d) Create a subset of the data consisting of only drama movies. Create a histogram and find the average domestic and world grosses for your subset. Are there outliers in either variable?
2. The sample data table Analgesics.jmp contains pain ratings from patients after treatments from three different pain relievers. The patients are labeled only by gender in this study. The study was meant to determine whether the three pain relievers were different in the amount of pain relief the patients experienced.
  - (a) Create a histogram of the variables gender, drug, and pain. Click on the histogram bars to determine whether the distribution of gender is approximately equal among the three analgesics.
  - (b) Create a separate histogram for the variable pain for each of the three different analgesics (**Note:** Use the **By** button). Does the mean pain response seem the same for each of the three analgesics?
3. The sample data table Scores.jmp contains data for the United States from the Third International Mathematics and Science Study, conducted in 1995. The variables came from testing more than 5000 students for their abilities in Calculus and Physics, and are

separated into four regions of the United States. Note that some students took the Calculus test, some took the Physics test, and some took both. Assume that the scores represent a random sample for each of the four regions of the United States.

- (a) Produce a histogram and find the mean scores for the United States on both tests. By clicking on the bars of the histogram, can you determine whether a high calculus score correlates highly with a high Physics score?
  - (b) Find the mean scores for the Calculus test for the four regions of the country. Do they appear to be approximately equal?
  - (c) Find the mean scores for the Physics tests for the four regions of the country. Do they appear to be approximately equal?
  - (d) Suppose that from an equivalent former test, the mean score of United States Calculus students was 450. Does this study show evidence that the score has increased since the last test?
  - (e) Construct a 95% confidence interval for the mean calculus score.
  - (f) Suppose that Physics teachers say that the overall United States score on the Physics test should be higher than 420. Do the data support their claim?
  - (g) Construct a 95% confidence interval for the mean Physics score.
4. The sample data table Cereal.jmp contains nutritional information for 76 types of cereal.
    - (a) Find the mean number of fat grams for the cereals in this data set. List any unusual observations.
    - (b) Use the Distribution platform to find the two types of cereal with unusually high fiber content.
    - (c) The hot/cold variable is used to specify whether the cereal was meant to be eaten hot or cold. Find the mean amount of sugars contained in the hot cereals and the cold cereals. Construct a 95% confidence interval for each.
  5. Various crime statistics for each of the 50 states in the United States are stored in the sample data table Crime.jmp.
    - (a) Examine the distributions of each statistic. Which (if any) do not appear to follow a normal distribution?
    - (b) Which two states are outliers with respect to the robbery variable?

6. Data for the Brigham Young football team are stored in the sample data table Football.jmp.
  - (a) Find the average height and weight of the players on the team.
  - (b) The Position variable identifies the primary position of each player. Which position has the smallest average weight? Which has the highest?
  - (c) Which position has the largest average neck measurements? What position (on average) can bench press the most weight?
7. The sample data table Hot Dogs.jmp came from an investigation of the taste and nutritional content of hot dogs.
  - (a) Construct a histogram of the type of hot dogs (beef, meat, and poultry). Is there an equal number of each type considered?
  - (b) The \$/oz variable represents the cost per ounce of hot dog. Construct an outlier box plot of this variable and find any outliers.
  - (c) Construct a 95% confidence interval for the caloric content of the three types of hot dogs. Which type gives (on average) the lowest calories?
  - (d) Test the conjecture that the mean sodium content of all hot dogs is 410 grams.





# 8

## The Difference Between Two Means

### Overview

Are the mean responses from two groups different? What evidence would it take to convince you? This question opens the door to many of the issues that pervade statistical inference, and this chapter explores these issues. Comparing group means also introduces an important statistical distinction regarding how the measurement or sampling process affects the way the resulting data are analyzed. This chapter also talks about validating statistical assumptions.

When two groups are considered, there are two distinct situations that lead to two different analyses:

**Independent Groups**—the responses from the two groups are unrelated and statistically independent. For example, the two groups might be two classrooms with two sets of students in them. The responses come from different experimental units or subjects. The responses are uncorrelated, and the means from the two groups are uncorrelated.

**Matched Pairs**—the two responses form a pair of measurements coming from the same experimental unit or subject. For example, a matched pair might be a before-and-after blood pressure measurement from the same subject. These responses are correlated, and the statistical method must take that into account.

## Chapter Contents

Overview .....	177
Two Independent Groups .....	179
When the Difference Isn't Significant .....	179
Check the Data .....	180
Launch the Fit Y by X Platform .....	181
Examine the Plot .....	182
Display and Compare the Means .....	183
Inside the Student's t-Test .....	184
Equal or Unequal Variances? .....	185
One-Sided Version of the Test .....	187
Analysis of Variance and the All-Purpose F-Test .....	188
How Sensitive Is the Test? How Many More Observations Are Needed? ..	190
When the Difference Is Significant .....	192
Normality and Normal Quantile Plots .....	194
Testing Means for Matched Pairs .....	196
Thermometer Tests .....	197
Look at the Data .....	198
Look at the Distribution of the Difference .....	199
Student's t-Test .....	199
The Matched Pairs Platform for a Paired t-Test .....	200
Optional Topic: An Equivalent Test for Stacked Data .....	203
Two Extremes of Neglecting the Pairing Situation: A Dramatization .....	205
A Nonparametric Approach .....	211
Introduction to Nonparametric Methods .....	211
Paired Means: The Wilcoxon Signed-Rank Test .....	211
Independent Means: The Wilcoxon Rank Sum Test .....	213
Exercises .....	214

## Two Independent Groups

For two different groups, the goal might be to estimate the group means and determine if they are significantly different. Along the way, it is certainly advantageous to notice anything else of interest about the data.

### When the Difference Isn't Significant

A study compiled height measurements from 63 children, all age 12. It's safe to say that as they get older, the mean height for males will be greater than for females, but is this the case at age 12? Let's find out:

- ~ Select **Help > Sample Data Library** and open Htwt12.jmp to see the data shown (partially) below.

There are 63 rows and three columns. This example uses Gender and Height. Gender has the Nominal modeling type, with codes for the two categories, "f" and "m". Gender will be the X variable for the analysis. Height contains the response of interest, and so it will be the Y variable.

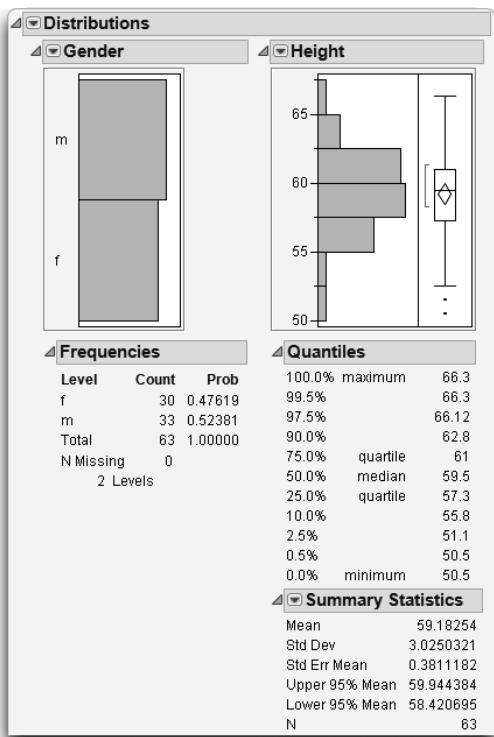
	Gender	Height	Weight
1	f	62.3	105
2	f	63.3	108
3	f	58.3	93
4	f	58.8	89
5	f	59.5	78.5
6	f	61.3	115
7	f	56.3	83.5
8	f	64.3	110.5
9	f	61.3	94

## Check the Data

To check the data, first look at the distributions of both variables graphically with histograms and box plots.

- ☞ Select **Analyze > Distribution**.
- ☞ In the launch window, assign Gender and Height to **Y, Columns**.
- ☞ Click **OK** to see an analysis window like the one shown in **Figure 8.1**.

**Figure 8.1** Histograms and Summary Tables



Every pilot walks around the plane looking for damage or other problems before starting up. No one would submit an analysis to the FDA without making sure that the data were not confused with data from another study. Do your kids use the same computer that you do? Then check your data. Does your data set have so many decimals of precision that it looks like it came from a random number generator? Great detectives let no clue go unnoticed. Great data analysts check their data carefully.

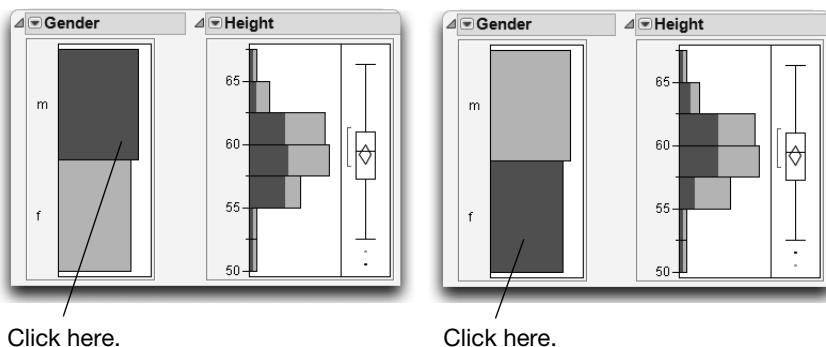
A look at the histograms for Gender and Height reveals that there are a few more males than females. The overall mean height is about 59, and there are no missing values (N is 63, and there are 63 rows in the table). The box plot indicates that two of the children seem unusually short compared to the rest of the data.

- Move the cursor to the Gender histogram, and click on the bar for “m”.

Clicking the bar highlights the males in the data table and also highlights the males in the Height histogram (See **Figure 8.2**). Now click on the “f” bar, which highlights the females and un-highlights the males.

By alternately clicking on the bars for males and females, you can see the conditional distributions of each subset highlighted in the Height histogram. This gives a preliminary look at the height distribution within each group, and it is these group means we want to compare.

**Figure 8.2** Interactive Histogram



## Launch the Fit Y by X Platform

We know to use the Fit Y by X platform because our context is comparing two variables. In this example, there are two gender groups, and we want to compare their mean weights.

You can compare these group means by assigning Height as the continuous Y variable and Gender as the nominal (grouping) X variable. Begin by launching the analysis platform:

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

- ⓐ In the launch window, assign Height to **Y** and Gender to **X**.

Notice that the role-prompts window indicates that you are doing a one-way analysis of variance (ANOVA). Because Height is continuous and Gender is categorical (nominal), the **Fit Y by X** command automatically gives a one-way layout for comparing distributions.

- ⓐ Click **OK** to see the initial graphs, which are side-by-side vertical dot plots for each group (see the left picture in **Figure 8.3**).

## Examine the Plot

The horizontal line across the middle shows the overall mean of all the observations. To identify possible outliers (students with unusual values):

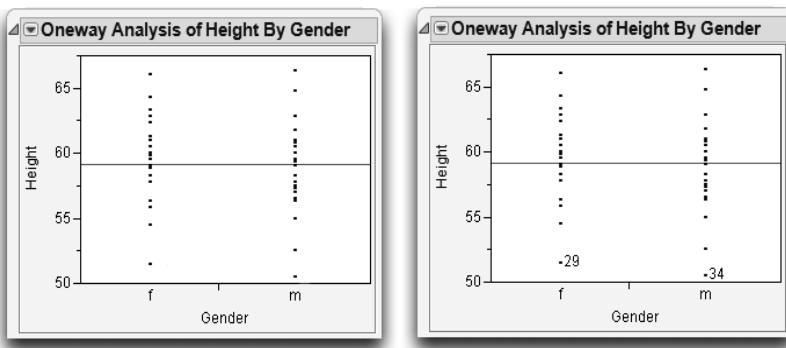
- ⓐ Click the lowest point in the “f” vertical scatter and Shift-click in the lowest point in the “m” sample.

Shift-clicking extends a selection so that the first selection does not un-highlight.

- ⓐ Select **Rows > Label/Unlabel** to see the plot on the right in **Figure 8.2**.

Now the points are labeled 29 and 34, the row numbers corresponding to each data point. Move your mouse over these points (or any other points) to see the values for Gender and Height. Click anywhere in the graph to un-highlight (deselect) the points.

**Figure 8.3** Plot of the Responses, Before and After Labeling Points



## Display and Compare the Means

The next step is to display the group means in the graph, and to obtain an analysis of them.

- ❖ Select **Means/Anova/Pooled t** from the red triangle menu next to Oneway Analysis.
- ❖ From the same menu, select **t Test**.

This adds analyses that estimate the group means and test to see if they are different.

**Note:** You don't usually select both versions of the *t*-test (shown in **Figure 8.5**). We're selecting these for illustration. To determine the correct test for other situations, see "Equal or Unequal Variances?" on page 185.

Lets discuss the first test, **Means/Anova/Pooled t**. This option automatically displays the *mean diamonds* as shown on the left in **Figure 8.4**, with summary tables and statistical test reports.

The center lines of the mean diamonds are the group means. The top and bottom of the diamonds form the 95% confidence intervals for the means. You can say the probability is 0.95 that this confidence interval contains the true group mean.

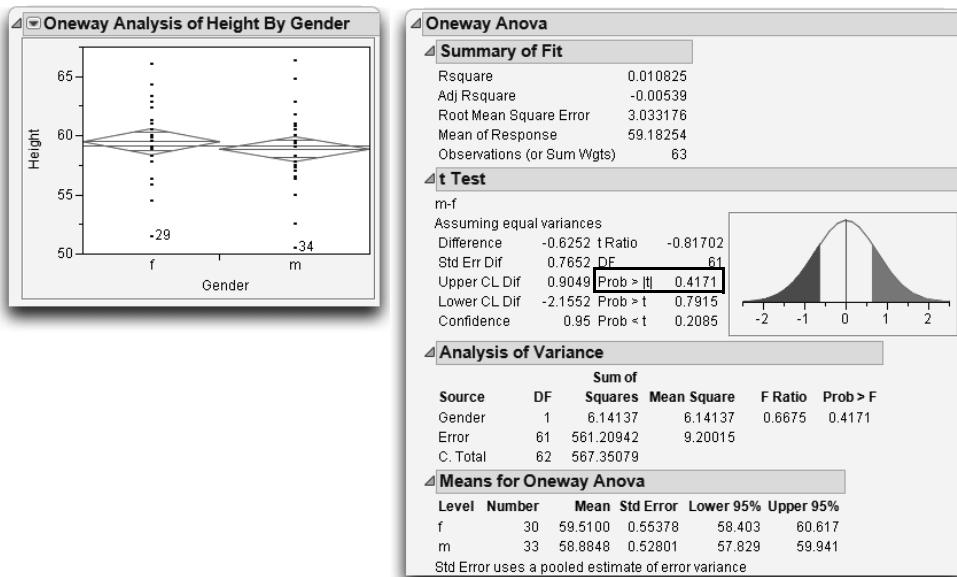
The confidence intervals show whether a mean is significantly different from some hypothesized value, but what can it show regarding whether two means are significantly different? Use the rule shown here to interpret mean diamonds.

**Interpretation Rule for Mean Diamonds:**  
If the confidence intervals shown by the mean diamonds do not overlap, the groups are significantly different (but the reverse is not necessarily true).

It is clear that the mean diamonds in this example overlap. Therefore, you need to take a closer look at the text report beneath the plots to determine if the means are really different. The report, shown in **Figure 8.4**, includes summary statistics, *t*-test reports, an analysis of variance, and means estimates.

Note that the *p*-value of the *t*-test (shown with the label **Prob>|t|** in the **t Test** section of the report) table is not significant.

**Figure 8.4** Diamonds to Compare Group Means and Pooled *t* Report



## Inside the Student's *t*-Test

The Student's *t*-test appeared in the last chapter to test whether a mean was significantly different from a hypothesized value. Now the situation is to test whether the difference of two means is significantly different from the hypothesized value of zero. The *t*-ratio is formed by first finding the difference between the estimate and the hypothesized value, and then dividing that quantity by its standard error.

$$t \text{ statistic} = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error of the estimate}}$$

In the current case, the estimate is the difference in the means for the two groups, and the hypothesized value is zero.

$$t \text{ statistic} = \frac{(\text{mean 1} - \text{mean 2}) - 0}{\text{standard error of the difference}}$$

For the means of two independent groups, the pooled standard error of the difference is the square root of the sum of squares of the standard errors of the means.

$$\text{standard error of the difference} = \sqrt{s_{\text{mean } 1}^2 + s_{\text{mean } 2}^2}$$

JMP calculates the pooled standard error and forms the tables shown in **Figure 8.4**. Roughly, you look for a *t*-statistic greater than 2 in absolute value to get significance at the 0.05 level. The *p*-value is determined in part by the degrees of freedom (DF) of the *t*-distribution. For this case, DF is the number of observations (63) minus two, because two means are estimated. With the calculated *t* (-0.817) and DF, the *p*-value is 0.4171. The label Prob>|t| is given to this *p*-value in the test table to indicate that it is the probability of getting an even greater absolute *t* statistic. Usually a *p*-value less than 0.05 is regarded as significant—this is the significance level.

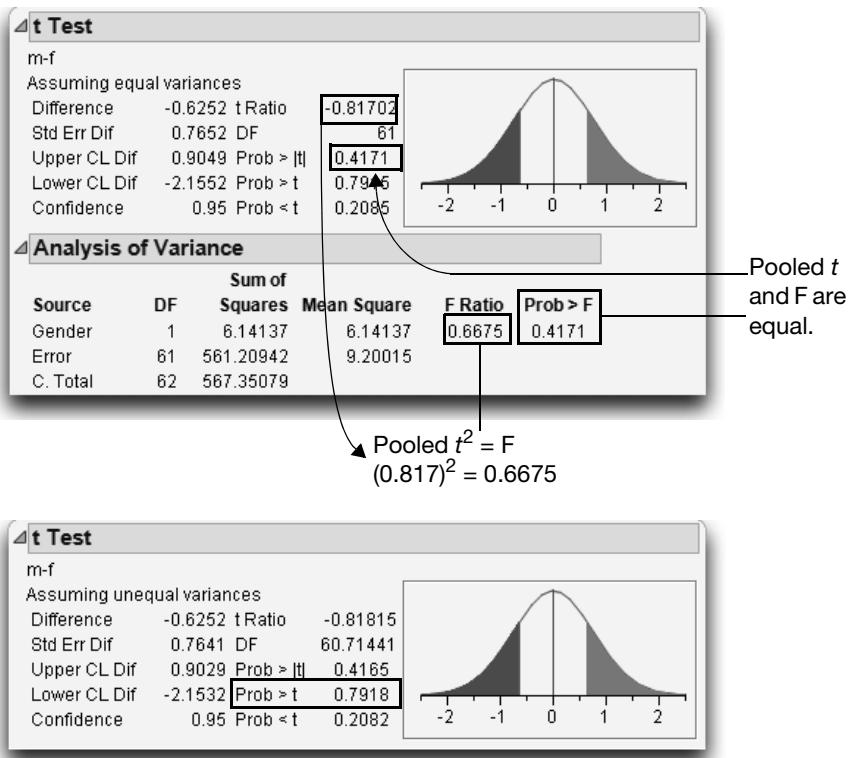
In this example, the *p*-value of 0.4171 isn't small enough to detect a significant difference in the means. Is this to say that the means are the same? Not at all. You just don't have enough evidence to show that they are different. If you collect more data, you might be able to show a significant, albeit small, difference.

## Equal or Unequal Variances?

The report shown in **Figure 8.5** shows two *t*-test reports.

- The uppermost report is labeled **Assuming equal variances**, and is generated with the **Means/Anova/Pooled t** command.
- The lower report is labeled **Assuming unequal variances**, and is generated with the **t Test** command.

Which is the correct report to use?

**Figure 8.5** *t*-Test and ANOVA Reports

In general, the unequal-variance *t*-test (also known as the *unpooled t*-test) is the preferred test. This is because the pooled version is quite sensitive (the opposite of robust) to departures from the equal-variance assumption (especially if the number of observations in the two groups is not the same), and often we cannot assume the variances of the two groups are equal. In addition, if the two variances are unequal, the unpooled test maintains the prescribed  $\alpha$ -level and retains good power. For example, you might think you are conducting a test with  $\alpha = 0.05$ , but it might in fact be 0.10 or 0.20. What you think is a 95% confidence interval might be, in reality, an 80% confidence interval (Cryer and Wittmer, 1999). For these reasons, we recommend the unpooled (**t Test** command) *t*-test for most situations. In this case, both *t*-tests are not significant.

However, the equal-variance version is included and discussed for several reasons.

- For situations with very small sample sizes (for example, having three or fewer observations in each group), the individual variances cannot be

estimated very well, but the pooled versions can be, giving better power. In these circumstances, the pooled version has slightly enough power.

- Pooling the variances is the only option when there are more than two groups, when the  $F$ -test must be used. Therefore, the pooled  $t$ -test is a useful analogy for learning the analysis of the more general, multi-group situation. This situation is covered in Chapter 9, “Comparing Many Means: One-Way Analysis of Variance.”

**Rule for  $t$ -tests:**

Unless you have very small sample sizes, or a specific a priori reason for assuming the variances are equal, use the  $t$ -test produced by the **t Test** command. When in doubt, use the **t Test** command (i.e., unpooled) version.

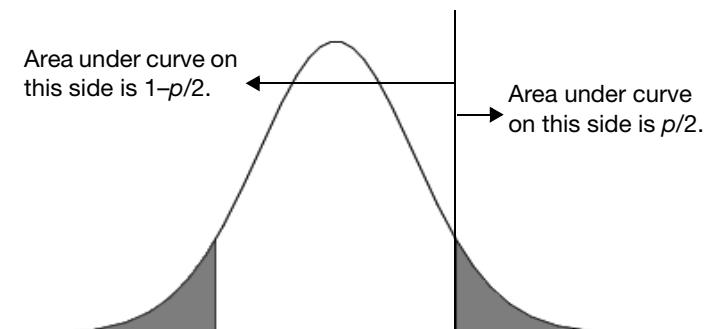
## One-Sided Version of the Test

The Student’s  $t$ -test in the previous example is for a two-sided alternative. In that situation, the difference could go either way (that is, either group could be taller), so a two-sided test is appropriate. The one-sided  $p$ -values are shown on the report, but you can get them by doing a little arithmetic on the reported two-sided  $p$ -value, forming one-sided  $p$ -values by using

$$\frac{p}{2} \text{ or } 1 - \frac{p}{2},$$

depending on the direction of the alternative.

**Figure 8.6** One- and Two-sided  $t$ -Test



The  $p$ -value presented by JMP is represented by the shaded regions in this figure. To use a one-sided test, calculate  $p/2$  or  $1-p/2$ .

In this example, the mean for males was less than the mean for females (the mean difference, using m-f, is -0.6252). The pooled t-test (top table in **Figure 8.5**) shows the *p*-value for the alternative hypothesis that females are taller is 0.2085, which is half the two-tailed *p*-value. Testing the other direction, the *p*-value is 0.7915. These values are reported in **Figure 8.5** as **Prob < t** and **Prob > t**, respectively.

## Analysis of Variance and the All-Purpose *F*-Test

As well as showing the *t*-test for comparing two groups, the top report in **Figure 8.5** shows an analysis of variance with its *F*-test. The *F*-test surfaces many times in the next few chapters, so an introduction is in order. Details will unfold later.

The *F*-test compares variance estimates for two situations, one a special case of the other. Not only is this useful for testing means, but other things, as well. Furthermore, when there are only two groups, the *F*-test is equivalent to the pooled (equal variance) *t*-test, and the *F*-ratio is the square of the *t*-ratio:  $(0.81)^2 = 0.66$ , as you can see in **Figure 8.5**.

To begin, look at the different estimates of variance as reported in the Analysis of Variance table.

First, the analysis of variance procedure pools all responses into one big population and estimates the population mean (the *grand mean*). The variance around that grand mean is estimated by taking the average sum of squared differences of each point from the grand mean.

The difference between a response value and an estimate such as the mean is called a *residual*, or sometimes the *error*.

What happens when a separate mean is computed for each group instead of the grand mean for all groups? The variance around these individual means is calculated, and this is shown in the Error line in the Analysis of Variance table. The Mean Square for Error is the estimate of this variance, called *residual variance* (also called  $s^2$ ), and its square root, called the *root mean squared error* (or  $s$ ), is the residual standard deviation estimate.

If the true group means are different, then the separate means give a better fit than the one grand mean. In other words, there will be less variance using the separate means than when using the grand mean. The change in the residual sum of squares from the single-mean model to the separate-means model leads us to

the  $F$ -test shown in the Model line of the Analysis of Variance table (“Model”, in this case, is Gender). If the hypothesis that the means are the same is true, the Mean Square for Model also estimates the residual variance.

The  $F$ -ratio is the Model Mean Square divided by the Error Mean Square:

$$F\text{-Ratio} = \frac{\text{Mean Square for the Model}}{\text{Mean Square for the Error}} = \frac{6.141}{9.200} = 0.6675$$

The  $F$ -ratio is a measure of improvement in fit when separate means are considered. If there is no difference between fitting the grand mean and individual means, then both numerator and denominator estimate the same variance (the grand mean residual variance), so the  $F$ -ratio is around 1. However, if the separate-means model does fit better, the numerator (the model mean square) contains more than just the grand mean residual variance, and the value of the  $F$ -test increases.

If the two mean squares in the  $F$ -ratio are statistically independent (and they are in this kind of analysis), then you can use the  $F$ -distribution associated with the  $F$ -ratio to get a  $p$ -value. This tells how likely you are to see the  $F$ -ratio given by the analysis if there really was no difference in the means.

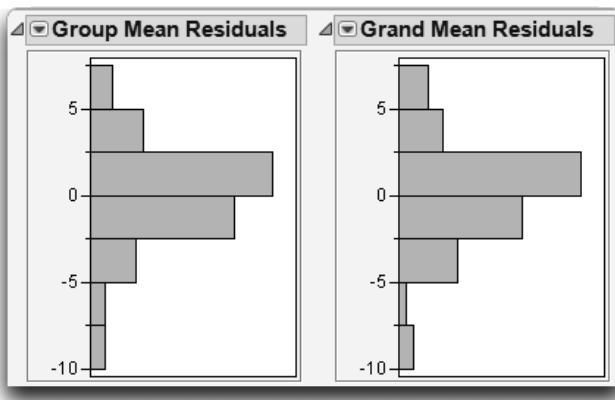
If the tail probability ( $p$ -value) associated with the  $F$ -ratio in the  $F$ -distribution is smaller than 0.05 (or the  $\alpha$ -level of your choice), you can conclude that the variance estimates are different, and thus that the means are different.

In this example, the total mean square and the error mean square are not much different. In fact, the  $F$ -ratio is actually less than one, and the  $p$ -value of 0.4171 (roughly the same as seen for the pooled  $t$ -test) is far from significant (it is much greater than 0.05).

The  $F$ -test can be viewed as whether the variance around the group means (the histogram on the left in **Figure 8.7**) is significantly less than the variance around the grand mean (the histogram on the right). In this case, the variance isn't much different. If the effect were significant, the variation showing on the left would have been much less than that on the right.

In this way, a test of variances is also a test on means. The  $F$ -test turns up again and again because it is oriented to comparing the variation around two models. Most statistical tests can be constituted this way.

**Figure 8.7** Residuals for Group Means Model (left) and Grand Mean Model (right)



**Terminology for Sums of Squares:**

All disciplines that use statistics use analysis of variance in some form. However, you may find different names used for its components. For example, the following are different names for the same kinds of sums of squares (SS):

$$\text{SS(model)} = \text{SS(regression)} = \text{SS(between)}$$

$$\text{SS(error)} = \text{SS(residual)} = \text{SS(within)}$$

## How Sensitive Is the Test?

## How Many More Observations Are Needed?

So far, in this example, there is no conclusion to report because the analysis failed to show anything. This is an uncomfortable state of affairs. It is tempting to state that we have shown no significant difference, but in statistics this is the same as saying the findings were inconclusive. Our conclusions (or lack of) can just as easily be attributed to not having enough data as to there being a very small true effect.

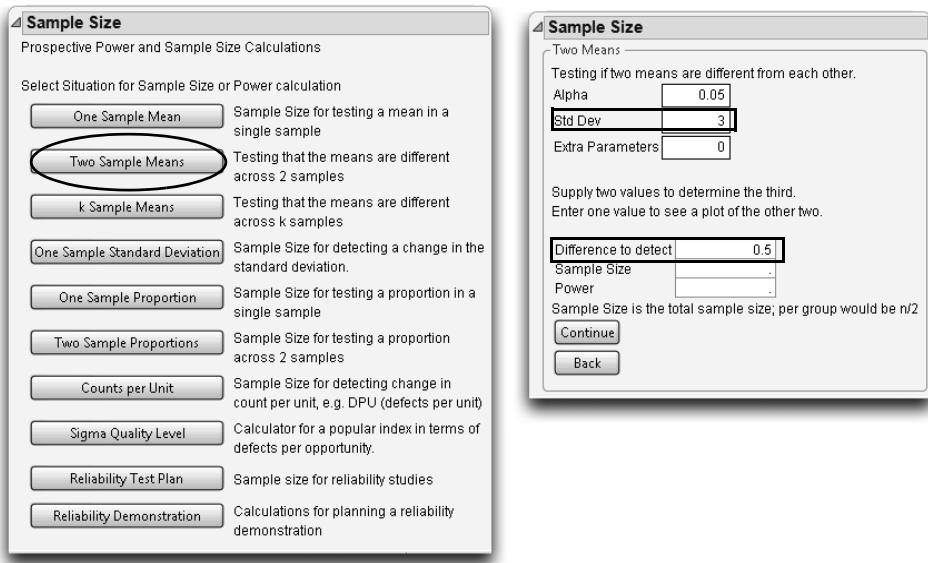
To gain some perspective on the power of the test, or to estimate how many data points are needed to detect a difference, we use the **Sample Size and Power** facility in JMP. Looking at power and sample size enables us to estimate some experimental values and graphically make decisions about the sample's data and effect sizes.

☞ Select **DOE > Design Diagnostics > Sample Size and Power**.

This command brings up a list of prospective power and sample size calculators for several situations, as shown in **Figure 8.8**. In our case, we are concerned with comparing two means. From the Distribution report on height, we can see that the standard deviation is about 3. Suppose we want to detect a difference of 0.5.

- ☞ Click **Two Sample Means**.
- ☞ Enter 3 for **Std Dev** and 0.5 as **Difference to Detect**, as shown on the right in **Figure 8.8**.

**Figure 8.8** Sample Size and Power Window



- ☞ Click **Continue** to see the graph shown on the left in **Figure 8.9**.
- ☞ Use the crosshair tool to find out what sample size is needed to have a power of 90%.

We would need around 1,519 data points to have a probability of 0.90 of detecting a difference of 0.5 with the current standard deviation.

How would this change if we were interested in a difference of 2 rather than a difference of 0.5?

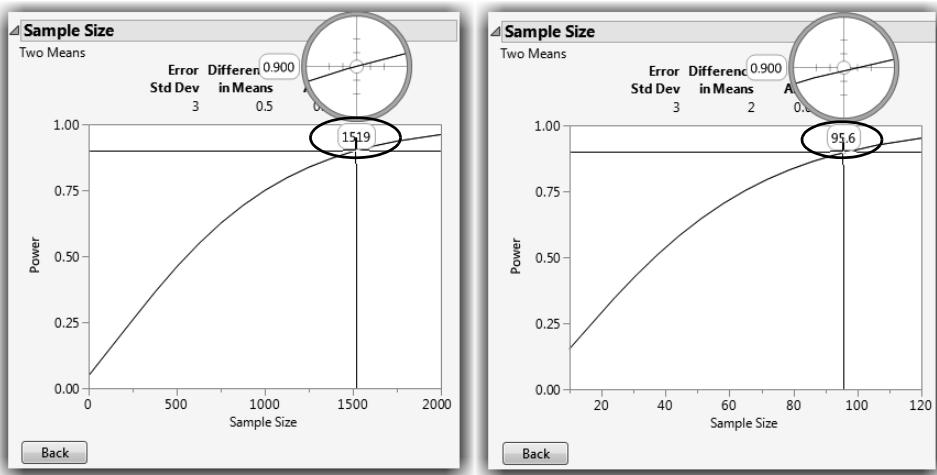
- ☞ Click the **Back** button and change the **Difference to Detect** from 0.5 to 2.
- ☞ Click **Continue**.

- Use the crosshair tool to find the number of data points you need for 90% power.

The results should be similar to the plot on the right in **Figure 8.9**.

We need only about 96 participants if we were interested in detecting a difference of 2.

**Figure 8.9** Finding a Sample Size for 90% Power



## When the Difference Is Significant

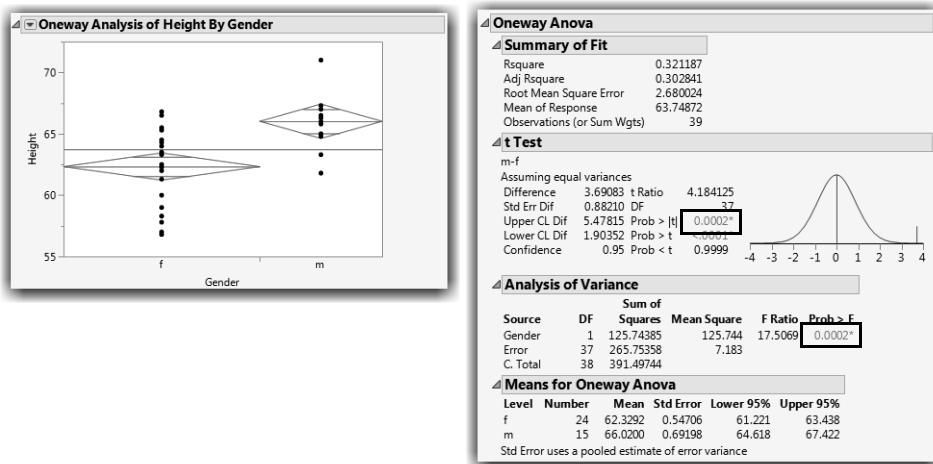
The 12-year-olds in the previous example don't have significantly different average heights, but let's take a look at the 15-year-olds.

- To start, select **Help > Sample Data Library** and open Htwt15.jmp.

Then, proceed as before:

- Select **Analyze > Fit Y by X**, assign Gender to **X, Factor** and Height to **Y, Response**, and then click **OK**.
- Select **Means/Anova/Pooled t** from the red triangle menu next to Oneway Analysis.

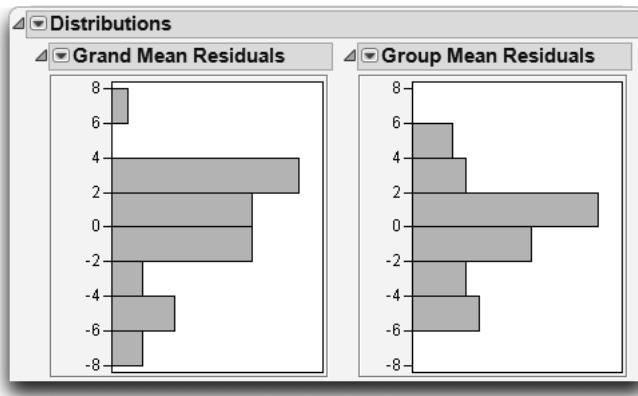
You should see the plot and tables shown in **Figure 8.10**.

**Figure 8.10** Analysis for Mean Heights of 15-year-olds

**Note:** As we discussed earlier, we normally recommend the unpooled (**t Test** command) version of the test. We're using the pooled version here as a basis for comparison between the results of the pooled *t*-test and the *F*-test.

The results for the analysis of the 15-year-old heights are completely different than the results for 12-year-olds. Here, the males are significantly taller than the females. You can see this because the confidence intervals shown by the mean diamonds do not overlap. You can also see that the *p*-values for both the two-tailed *t*-test and the *F*-test are 0.0002, which is highly significant.

The *F*-test results say that the variance around the group means is significantly less than the variance around the grand mean. These two variances are shown, using uniform scaling, in the histograms in **Figure 8.11**.

**Figure 8.11** Histograms of Grand Means Variance and Group Mean Variance

## Normality and Normal Quantile Plots

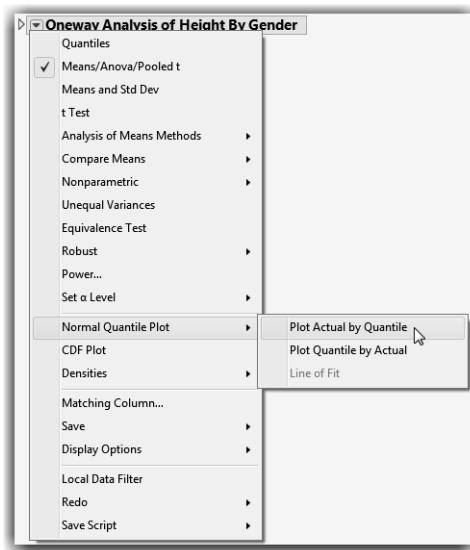
The  $t$ -tests (and  $F$ -tests) used in this chapter assume that the sampling distribution for the group means is the normal distribution. With sample sizes of at least 30 for each group, Normality is probably a safe assumption. The Central Limit Theorem says that means approach a normal distribution as the sample size increases even if the original data are not normal.

If you suspect non-normality (due to small samples, or outliers, or a non-normal distribution), consider using nonparametric methods, covered at the end of this chapter.

To assess normality, use a normal quantile plot. This is particularly useful when overlaid for several groups, because so many attributes of the distributions are visible in one plot.

- Return to the Fit Y by X platform showing Height by Gender for the 12-year-olds and select **Normal Quantile Plot > Plot Actual by Quantile** from the red triangle menu next to Oneway Analysis.
- Do the same for the 15-year-olds.

The resulting plots (**Figure 8.12**) show the data compared to the normal distribution. The normality is judged by how well the points follow a straight line. In addition, the normal quantile plot gives other useful information:

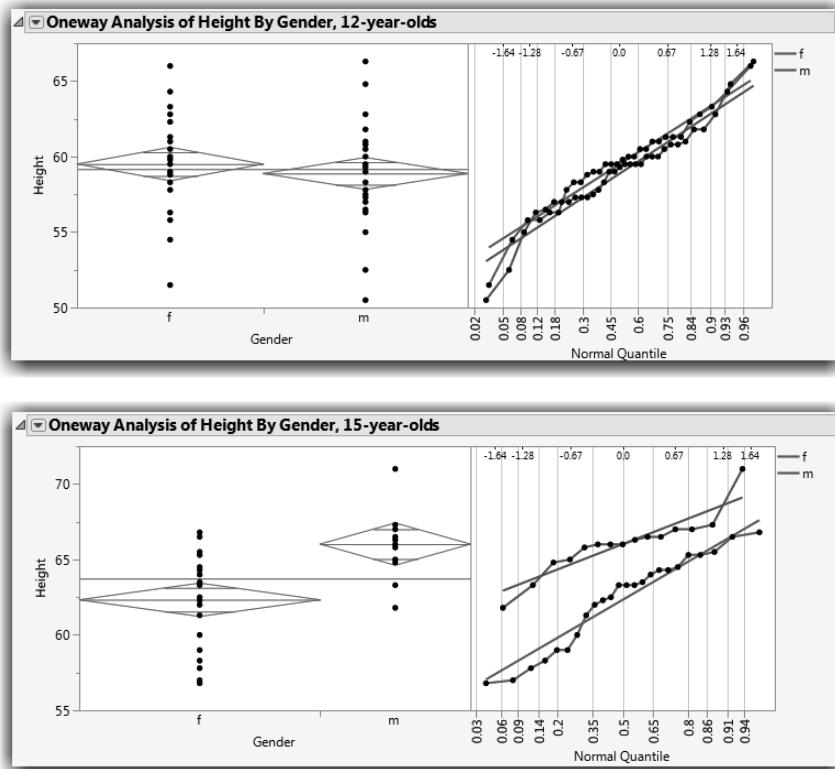


- The standard deviations are the slopes of the straight lines. Lines with steeper slopes represent the distributions with the greater variances.
- The vertical separation of the lines in the middle shows the difference in the means. The separation of other quantiles shows at other points on the  $x$ -axis.

The distributions for all groups look reasonably normal since the points (generally) cluster around their corresponding line.

The first graph in **Figure 8.12** confirms that heights of 12-year-old males and females have nearly the same mean and variance—the slopes (standard deviations) are the same and the positions (means) are only slightly different.

The second graph in **Figure 8.12** shows 15-year-old males and females have different means and different variances—the slope (standard deviation) is higher for the females, but the position (mean) is higher for the males. Recall that we used the pooled  $t$ -test in the analysis in **Figure 8.10**. Since the variances are different, the unpooled  $t$ -test (the **t Test** command) would have been the more appropriate test.

**Figure 8.12** Normal Quantile Plots for 12-year-olds and 15-year-olds

## Testing Means for Matched Pairs

Consider a situation where two responses form a pair of measurements coming from the same experimental unit. A typical situation is a before-and-after measurement on the same subject. The responses are correlated, and if only the group means are compared—ignoring the fact that the groups have a pairing—information is lost. The statistical method called the *paired t-test* enables you to compare the group means, while taking advantage of the information gained from the pairings.

In general, if the responses are positively correlated, the paired *t*-test gives a more significant *p*-value than the *t*-test for independent means (grouped *t*-test) discussed in the previous sections. If responses are negatively correlated, then the paired *t*-test is less significant than the grouped *t*-test. In most cases where the

pair of measurements are taken from the same individual at different times, they are positively correlated, but be aware that it is possible for pairs to have a negative correlation.

## Thermometer Tests

A health care center suspected that temperature readings from a new ear drum probe thermometer were consistently higher than readings from the standard oral mercury thermometer. To test this hypothesis, two temperature readings were taken on 20 patients, one with the ear-drum probe, and the other with the oral thermometer. Of course, there was variability among the readings, so they were not expected to be exactly the same. However, the suspicion was that there was a systematic difference—that the ear probe was reading too high.

For this example, select **Help > Sample Data Library** and open Therm.jmp.

A partial listing of the data table appears in **Figure 8.13**. The Therm.jmp data table has 20 observations and 4 variables. The two responses are the temperatures taken orally and tympanically (by ear) on the same person on the same visit.

**Figure 8.13** Comparing Paired Scores

Pair ID	First paired response	Second paired response	Difference between paired columns
1	John	96.9	98.5
2	Andrew	98.0	98.4
3	Sally	100.5	101.5
4	Joanie	98.3	99.5
5	Kevin	97.7	98.0
6	Katie	101.8	102.6
7	Jennifer	98.4	99.2
8	Bill	98.2	100.5
9	Thor	97.8	98.2

For paired comparisons, the two responses need to be arranged in two columns, each with a continuous modeling type. This is because JMP assumes that each row represents a single experimental unit. Since the two measurements are taken from the same person, they belong in the same row. It is also useful to create a new column with a formula to calculate the difference between the two responses. (If your data table is arranged with the two responses in different rows, use the **Tables > Split** command to rearrange it. For more information, see “Juggling Data Tables” on page 51.)

## Look at the Data

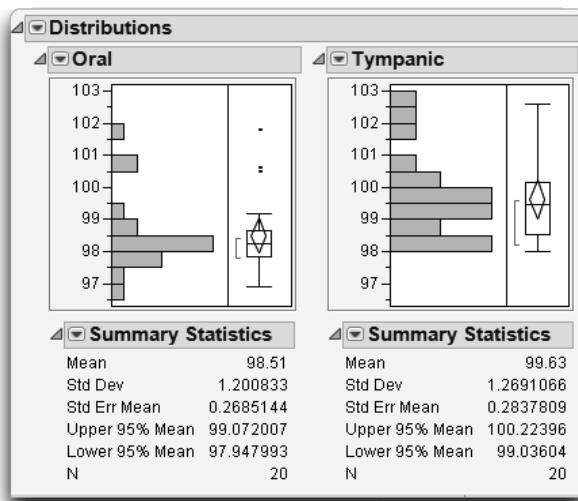
Start by inspecting the distribution of the data. To do this:

- ☛ Select **Analyze > Distribution** and assign Oral and Tympanic to **Y, Columns**.
- ☛ When the results appear, select **Uniform Scaling** from the red triangle menu next to Distribution to display the plots on the same scale.

The histograms (in **Figure 8.14**) show the temperatures to have different distributions. The mean looks higher for the Tympanic temperatures. However, as you will see later, this side-by-side picture of each distribution can be misleading if you try to judge the significance of the difference from this perspective.

What about the outliers at the top end of the Oral temperature distribution? Are they of concern? Can you expect the distribution to be normal? Not really. *It is not the temperatures that are of interest, but the difference in the temperatures.* So there is no concern about the distribution so far. If the plots showed temperature readings of 110 or 90, there would be concern, because that would be suspicious data for human temperatures.

**Figure 8.14** Plots and Summary Statistics for Temperature



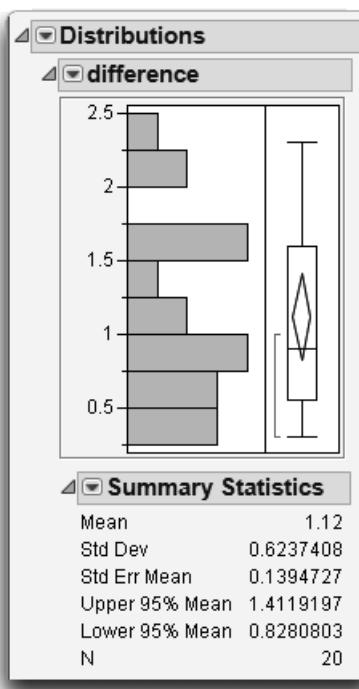
## Look at the Distribution of the Difference

The comparison of the two means is actually a comparison of the difference between them. Inspect the distribution of the differences:

- ❖ Select **Analyze > Distribution** and assign difference to **Y, Columns**.

The results (shown in **Figure 8.15**) show a distribution that seems to be above zero. In the Summary Statistics table, the lower 95% limit for the mean is 0.828—greater than zero.

**Figure 8.15** Histogram and Summary Statistics of the Difference



## Student's *t*-Test

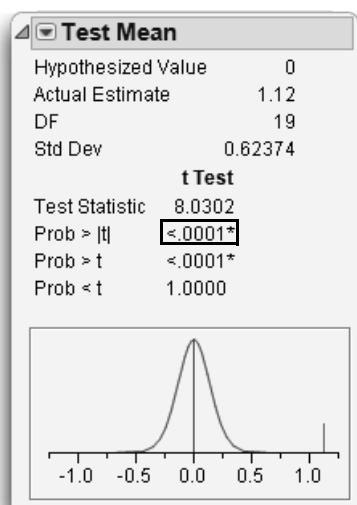
- ❖ Select **Test Mean** from the red triangle menu for the histogram of the difference variable. When prompted for a hypothesized value, accept the default value of zero.
- ❖ Click **OK**.

Now you have the *t*-test for testing that the mean over the matched pairs is the same.

In this case, the results in the Test Mean table, shown here, show a *p*-value of less than 0.0001, which supports our visual guess that there is a significant difference between methods of temperature taking. The tympanic temperatures are significantly higher than the oral temperatures.

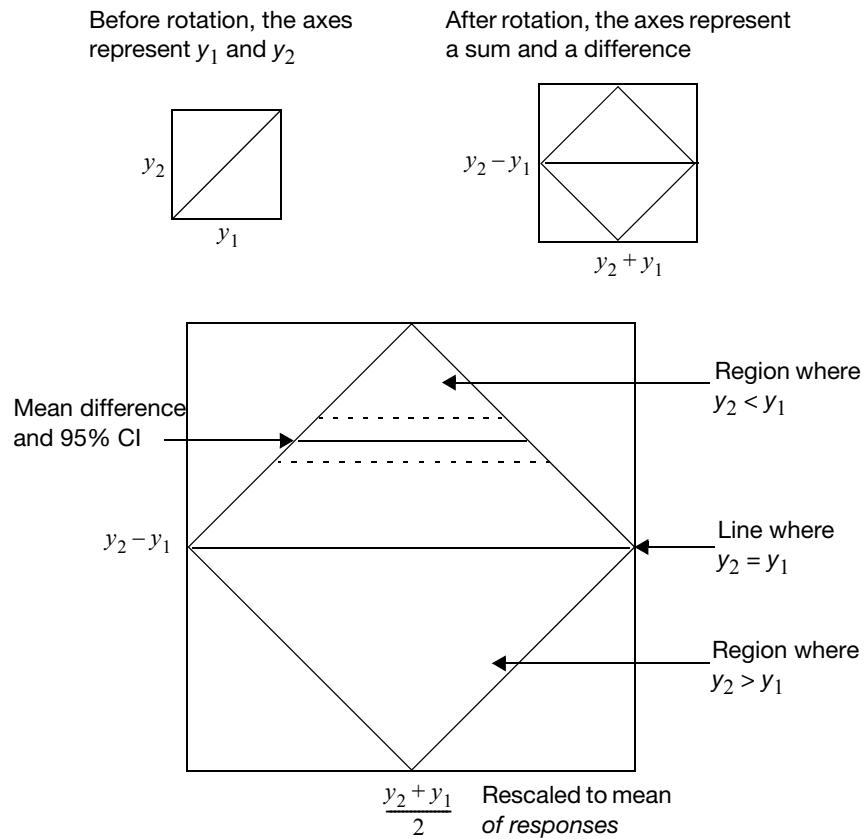
There is also a nonparametric test, the Wilcoxon signed-rank test, described at the end of this chapter, that tests the difference between two means. This test is produced by selecting the appropriate box on the test mean window.

The last section in this chapter discusses the Wilcoxon signed-rank test.



## The Matched Pairs Platform for a Paired *t*-Test

JMP offers a special platform for the analysis of paired data. The Matched Pairs platform compares means between two response columns using a paired *t*-test. The primary plot in the platform is a plot of the difference of the two responses on the *y*-axis, and the mean of the two responses on the *x*-axis. This graph is the same as a scatterplot of the two original variables, but rotated 45° clockwise. A 45° rotation turns the original coordinates into a difference and a sum. By rescaling, this plot can show a difference and a mean, as illustrated in **Figure 8.16**.

**Figure 8.16** Transforming to Difference by Sum Is a Rotation by 45°

- There is a horizontal line at zero, which represents no difference between the group means ( $y_2 - y_1 = 0$  or  $y_2 = y_1$ ).
- There is a line that represents the computed difference between the group means, and dashed lines around it showing a confidence interval.

**Note:** If the confidence interval does not contain the horizontal zero line, the test detects a significant difference.

Seeing this platform in use reveals its usefulness.

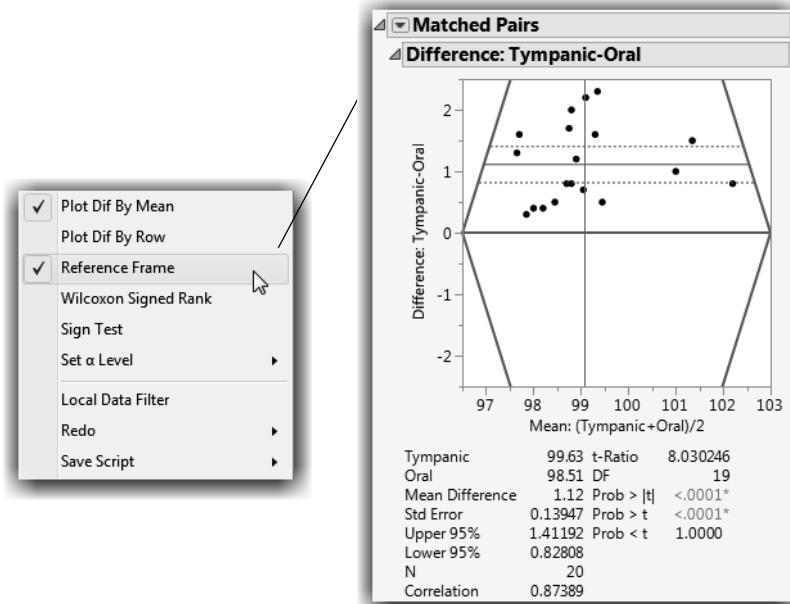
☞ Select **Analyze > Specialized Modeling > Matched Pairs** and assign Oral and Tympanic to **Y, Paired Response**.

☞ Click **OK** to see a scatterplot of Tympanic and Oral as a matched pair.

To see the rotation of the scatterplot in **Figure 8.17** more clearly,

- ✓ Select the **Reference Frame** option from the red triangle menu next to Matched Pairs.

**Figure 8.17** Scatterplot of Matched Pairs Analysis



The analysis first draws a reference line where the difference is equal to zero. This is the line where the means of the two columns are equal. If the means are equal, then the points should be evenly distributed around this line. You should see about as many points above this line as below it. If a point is above the reference line, it means that the difference is greater than zero. In this example, points above the line show the situation where the Tympanic temperature is greater than the Oral temperature.

Parallel to the reference line at zero is a solid red line that is displaced from zero by an amount equal to the difference in means between the two responses. This red line is the line of fit for the sample. The test of the means is equivalent to asking if the red line through the points is significantly separated from the reference line at zero.

The dashed lines around the red line of fit show the 95% confidence interval for the difference in means.

This scatterplot gives you a good idea of each variable's distribution, as well as the distribution of the difference.

**Interpretation Rule for the Paired  $t$ -test Scatterplot:**

If the confidence interval (represented by the dashed lines around the red line) contains the reference line at zero, then the two means are not significantly different.

Another feature of the scatterplot is that you can see the correlation structure. If the two variables are positively correlated, they lie closer to the line of fit, and the variance of the difference is small. If the variables are negatively correlated, then most of the variation is perpendicular to the line of fit, and the variance of the difference is large. It is this variance of the difference that scales the difference in a  $t$ -test and determines whether the difference is significant.

The paired  $t$ -test table beneath the scatterplot of **Figure 8.17** gives the statistical details of the test. The results should be identical to those shown earlier in the Distribution platform. The table shows that the observed difference in temperature readings of 1.12 degrees is significantly different from zero.

### Optional Topic: An Equivalent Test for Stacked Data

There is a third approach to the paired  $t$ -test. Sometimes, you receive grouped data with the response values stacked into a single column instead of having a column for each group.

Suppose the temperature data is arranged as shown here. Both the oral and tympanic temperatures are in the single column called Temperature. They are identified by the values of the Type and the Name columns.

**Note:** You can create this table yourself by using the **Tables > Stack** command to stack the Oral and Tympanic columns in the Therm.jmp table used in the previous examples.

	Name	Type	Temperature
1	John	Oral	96.9
2	John	Tympanic	98.5
3	Andrew	Oral	98.0
4	Andrew	Tympanic	98.4
5	Sally	Oral	100.5
6	Sally	Tympanic	101.5
7	Joanie	Oral	98.3
8	Joanie	Tympanic	99.5
9	Kevin	Oral	97.7
10	Kevin	Tympanic	98.0
11	Katie	Oral	101.8

If you select **Analyze > Fit Y by X** with Temperature (the response of both temperatures) as Y and Type (the classification) as X and select **t Test** from the red triangle menu, you get the *t*-test designed for independent groups, which is inappropriate for paired data.

However, fitting a model that includes an adjustment for each person fixes the independence problem because the correlation is due to temperature differences from person to person. To do this, you need to use the **Fit Model** command, covered in Chapter 14, “Fitting Linear Models.” The response is modeled as a function of both the category of interest (Type—Oral or Tympanic) and the Name category that identifies the person.

- ~ Select **Analyze > Fit Model**.
- ~ When the Fit Model window appears, assign Temperature to **Y**, and both Type and Name as model effects.
- ~ Click **Run Model**.

The resulting *p*-value for the category effect is identical to the *p*-value from the paired *t*-test shown previously. In fact, the *F*-ratio in the effect test is exactly the square of the *t*-test value in the paired *t*-test. In this case the formula is

$$(Paired t\text{-test statistic})^2 = 8.0302^2 = 64.4848 = (\text{stacked } F\text{-test statistic})$$

The Fit Model platform gives you a plethora of information, but for this example you need only the Effect Test table (Figure 8.18). It shows an *F*-ratio of 64.48, which is exactly the square of the *t*-ratio of 8.03 found with the previous approach. It’s just another way of doing the same test.

**Figure 8.18** Equivalent *F*-test on Stacked Data

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Type	1	1	12.544000	64.4848	<.0001*
Name	19	19	54.304000	14.6926	<.0001*

*t*-ratio from previous analysis = 8.03

*F*-ratio = square of *t*-ratio = 64.48

The alternative formulation for the paired means covered in this section is important for cases in which there are more than two related responses. Having many related responses is a *repeated-measures* or *longitudinal* situation. The generalization of the paired *t*-test is called the *multivariate* or  $T^2$  approach, whereas the generalization of the stacked formulation is called the *mixed-model* or *split-plot* approach.

## Two Extremes of Neglecting the Pairing Situation: A Dramatization

What happens if you do the wrong test? What happens if you do a *t*-test for independent groups on highly correlated paired data?

Consider the following two data tables:

- ☞ Select **Help > Sample Data Library** and open Blood Pressure by Time.jmp to see the left-hand table in **Figure 8.19**.

This table represents blood pressure measured for ten people in the morning and again in the afternoon. The hypothesis is that, on average, the blood pressure in the morning is the same as it is in the afternoon.

- ☞ Open the sample data table called BabySleep.jmp to see the right-hand table in **Figure 8.19**.

In this table, a researcher monitored ten two-month-old infants at 10 minute intervals over a day and counted the intervals in which a baby was asleep or awake. The hypothesis is that at two months old, the asleep time is equal to the awake time.

**Figure 8.19** The Blood Pressure by Time and BabySleep Sample Data Tables

**Blood Pressure by ...**

	BP AM	BP PM	Dif
x	1	70	94
x	2	85	100
x	3	92	106
x	4	97	113
x	5	110	130
x	6	110	131
x	7	126	142
x	8	137	149
x	9	140	156
x	10	148	170

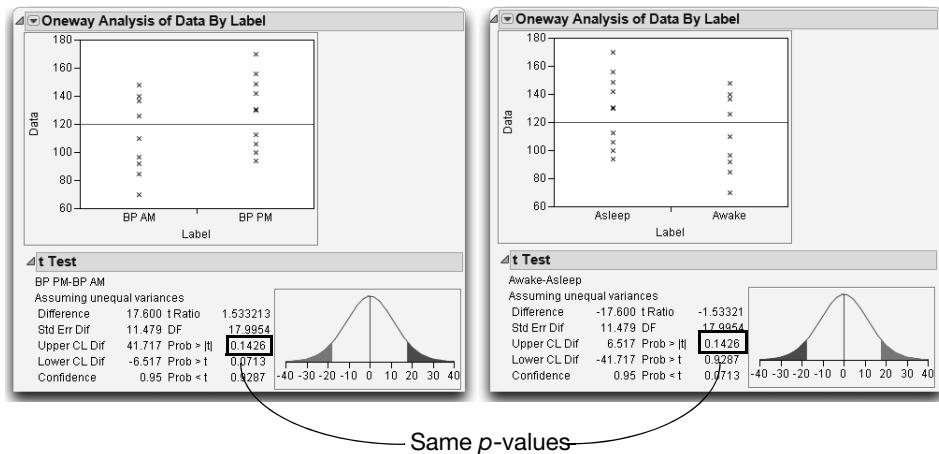
**BabySleep**

	Awake	Asleep	Dif
x	1	110	131
x	2	126	113
x	3	85	156
x	4	140	100
x	5	92	149
x	6	70	170
x	7	148	94
x	8	97	142
x	9	137	106
x	10	110	130

Let's do the incorrect *t*-test (the *t*-test for independent groups). Before conducting the test, we need to reorganize the data using the **Stack** command.

- ⓐ Select **Tables > Stack** to create two new tables (a stacked version of Baby Sleep.jmp and a stacked version of Blood Pressure by Time.jmp). Stack Awake and Asleep to form a single column in one table, and BP AM and BP PM to form a single column in a second table.
- ⓑ Select **Analyze > Fit Y by X** on both new tables, using the Label column to **Y, Response** and the Data column as **X, Factor**.
- ⓒ Select **t Test** from the red triangle menu for each plot.

The results for the two analyses are shown in **Figure 8.20**. The conclusions are that there is no significant difference between Awake and Asleep time, nor is there a difference between time of blood pressure measurement. The summary statistics are the same in both analyses and the probability is the same, showing no significance ( $p = 0.1426$ ).

**Figure 8.20** Results of *t*-test for Independent Means

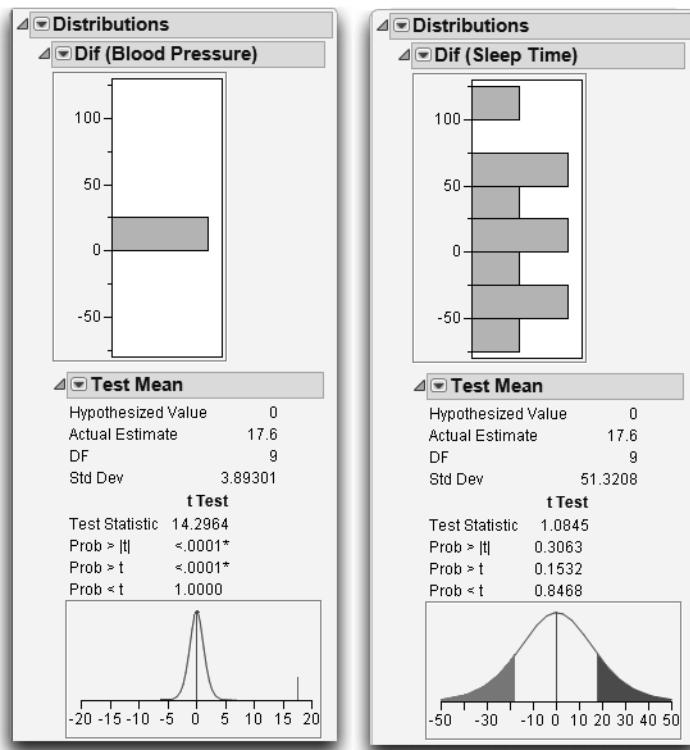
Now do the proper test, the paired *t*-test.

- ⌚ Using the original (unstacked) tables, chose **Analyze > Distribution** and examine a distribution of the Dif variable in each table.
- ⌚ Double-click on the axis of the blood pressure histogram and make its scale match the scale of the baby sleep axis.
- ⌚ Then, test that each mean is zero (see **Figure 8.21**).

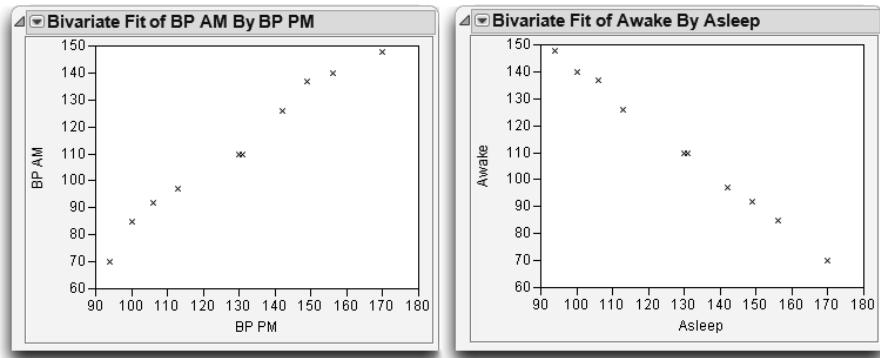
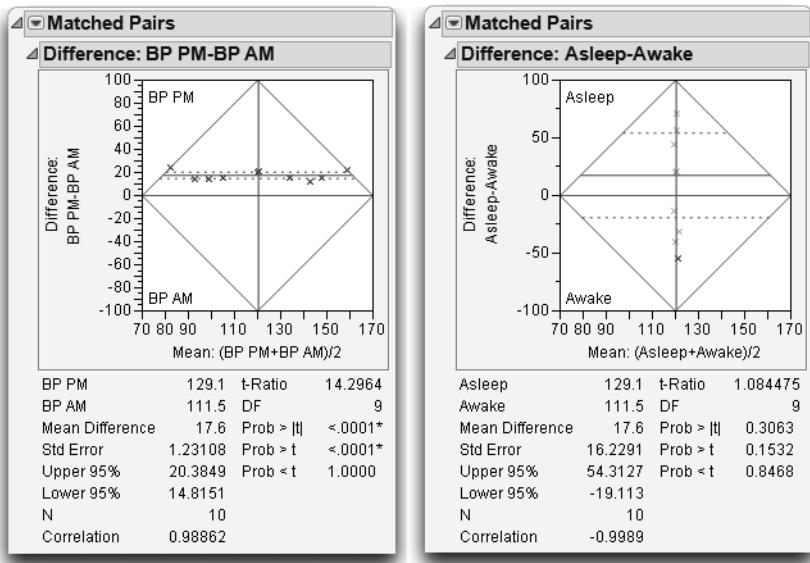
In this case, the analysis of the differences leads to very different conclusions.

- The mean difference between time of blood pressure measurement is highly significant because the variance is small (Std Dev=3.89).
- The mean difference between awake and asleep time is not significant because the variance of this difference is large (Std Dev=51.32).

So don't judge the mean of the difference by the difference in the means without noting that the variance of the difference is the measuring stick, and that the measuring stick depends on the correlation between the two responses.

**Figure 8.21** Histograms and Summary Statistics Show the Problem

The scatterplots produced by the Bivariate platform (**Figure 8.22**) and the Matched Pairs platform (**Figure 8.23**) show what is happening. The first pair is highly positively correlated, leading to a small variance for the difference. The second pair is highly negatively correlated, leading to a large variance for the difference.

**Figure 8.22** Bivariate Scatterplots of Blood Pressure and Baby Sleep Data**Figure 8.23** Paired t-test for Positively and Negatively Correlated Data

To review, make sure you can answer the following question:

What is the reason that you use a different *t*-test for matched pairs?

- a. Because the statistical assumptions for the *t*-test for groups are not satisfied with correlated data.
- b. Because you can detect the difference much better with a paired *t*-test. The paired *t*-test is much more sensitive to a given difference.

c. Because you might be overstating the significance if you used a group  $t$ -test rather than a paired  $t$ -test.

d. Because you are testing a different thing.

Answer: All of the above.

a. The grouped  $t$ -test assumes that the data are uncorrelated and paired data are correlated. So you would violate assumptions using the grouped  $t$ -test.

b. Most of the time the data are positively correlated, so the difference has a smaller variance than you would attribute if they were independent. So the paired  $t$ -test is more powerful—that is, more sensitive.

c. There may be a situation in which the pairs are negatively correlated, and if so, the variance of the difference would be greater than you expect from independent responses. The grouped  $t$ -test would overstate the significance.

d. You are testing the same thing in that the mean of the difference is the same as the difference in the means. But you are testing a different thing in that the variance of the mean difference is different from the variance of the differences in the means (ignoring correlation), and the significance for means is measured with respect to the variance.

### Mouse Mystery

Comparing two means is not always straightforward. Consider this story.

A food additive showed promise as a dieting drug. An experiment was run on mice to see if it helped control their weight gain. If it proved effective, then it could be sold to millions of people trying to control their weight.

After the experiment was over, the average weight gain for the treatment group was significantly less than for the control group, as hoped for. Then someone noticed that the treatment group had fewer observations than the control group. It seems that the food additive caused the obese mice in that group to tend to die young, so the thinner mice had a better survival rate for the final weighing.

These tables are set up such that the values are identical for the two responses, as a marginal distribution, but the values are paired differently so that the Blood Pressure by Time difference is highly significant and the babySleep difference is non-significant. This illustrates that it is the distribution of the difference that is important, not the distribution of the original values. If you don't look at the data correctly, the data can appear the same even when they are dramatically different.

# A Nonparametric Approach

## Introduction to Nonparametric Methods

Nonparametric methods provide ways to analyze and test data that do not depend on assumptions about the distribution of the data. In order to ignore normality assumptions, nonparametric methods disregard some of the information in your data. Typically, instead of using actual response values, you use the *rank ordering* of the response.

Most of the time you don't really throw away much relevant information, but you avoid information that might be misleading. A nonparametric approach creates a statistical test that ignores all the spacing information between response values. This protects the test against distributions that have very non-normal shapes, and can also provide insulation from data contaminated by rogue values.

In many cases, the nonparametric test has almost as much power as the corresponding parametric test and in some cases has more power. For example, if a batch of values is normally distributed, the rank-scored test for the mean has 95% efficiency relative to the most powerful normal-theory test.

The most popular nonparametric techniques are based on functions (scores) of the ranks:

- the rank itself, called a *Wilcoxon score*
- whether the value is greater than the median; whether the rank is more than  $\frac{n+1}{2}$ , called the *Median test*
- a normal quantile, computed as in normal quantile plots, called the *van der Waerden score*

Nonparametric methods are not contained in a single platform in JMP, but are available through many platforms according to the context where that test naturally occurs.

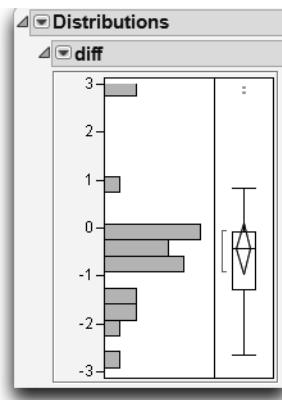
## Paired Means: The Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is the nonparametric analog to the paired *t*-test. You do a signed-rank test by testing the distribution of the difference of matched pairs, as discussed previously. The following example shows the advantage of using the signed-rank test when data are non-normal.

 Select **Help > Sample Data Library** and open Chamber.jmp

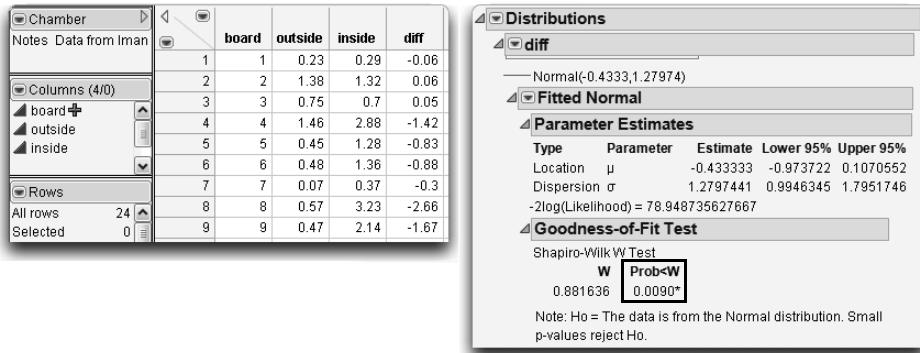
The data represent electrical measurements on 24 wiring boards. Each board is measured first when soldering is complete, and again after three weeks in a chamber with a controlled environment of high temperature and humidity (Iman 1995).

- ⓐ Examine the **diff** variable (difference between the outside and inside chamber measurements) with **Analyze > Distribution**.
- ⓑ Select the **Continuous Fit > Normal** from the red triangle menu next to **diff**.
- ⓒ Select **Goodness of Fit** from the red triangle menu next to **Fitted Normal**.



The Shapiro-Wilk W-test in the report tests the assumption that the data are normal. The probability of 0.0090 given by the normality test indicates that the data are significantly non-normal. In this situation, it might be better to use signed ranks for comparing the mean of **diff** to zero. Since this is a matched pairs situation, use the Matched Pairs platform.

**Figure 8.24** The Chamber Data and Test For Normality



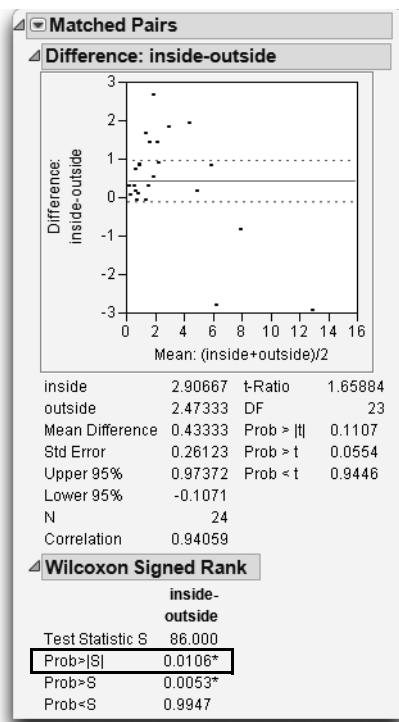
- ☞ Select **Analyze > Specialized Modeling > Matched Pairs.**

- ☞ Assign outside and inside as the paired responses, then click **OK**.

When the report appears,

- ☞ Select **Wilcoxon Signed Rank** from the red triangle menu on the Matched Pairs title bar.

Note that the standard *t*-test probability is insignificant ( $p = 0.1107$ ). However, in this example, the signed-rank test detects a difference between the groups with a *p*-value of 0.0106.



## Independent Means: The Wilcoxon Rank Sum Test

If you want to nonparametrically test the means of two independent groups, as in the *t*-test, then you can rank the responses and analyze the ranks instead of the original data. This is the *Wilcoxon rank sum test*. It is also known as the *Mann-Whitney U-test* because there is a different formulation of it that was not discovered to be equivalent to the Wilcoxon rank sum test until after it had become widely used.

- ☞ Open Hwt15.jmp again, select **Analyze > Fit Y by X** with Height as Y and Gender as X, and then click **OK**.

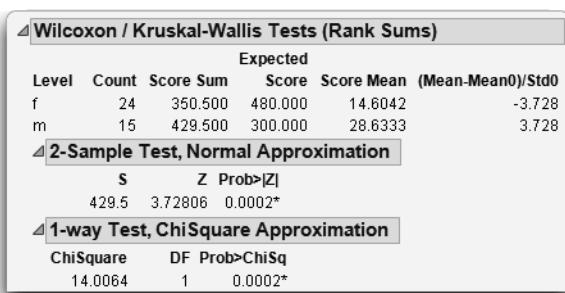
This is the same platform that gave the *t*-test.

- ☞ Select **Nonparametric > Wilcoxon Test** from the red triangle menu next to Matched Pairs.

The result is the report in **Figure 8.25**. This table shows the sum and mean ranks for each group, then the Wilcoxon statistic along with an approximate *p*-value based on the large-sample distribution of the statistic. In this case, the difference in the mean heights is declared significant, with a *p*-value of 0.0002. If you have small samples, you should consider also checking the tables of the Wilcoxon to

obtain a more exact test, because the normal approximation is not very precise in small samples.

**Figure 8.25** Wilcoxon Rank Sum Test for Independent Groups



## Exercises

1. The sample data table On-Time Arrivals.jmp (*Aviation Consumer Home Page*, 1999) contains the percentage of airlines' planes that arrived on time in 29 airports (those that the Department of Transportation designates "reportable"). You are interested in seeing if there are differences between certain months.
  - (a) Suppose you want to examine the differences between March and June. Is this a situation where a grouped test of two means is appropriate or would a matched pairs test be a better choice?
  - (b) Based on your answer in (a), determine if there is a difference in on-time arrivals between the two months.
  - (c) Similarly, determine if there is a significant difference between the months June and August, and also between March and August.
2. William Gosset was a pioneer in statistics. In one famous experiment, he wanted to investigate the yield from corn planted from two different types of seeds. One type of seed was dried in the normal way, while the other was kiln-dried. Gossett planted one seed of each seed type in 11 different plots and measured the yield for each one. The drying methods are represented by the columns Regular or Kiln in the sample data table Gosset's Corn.jmp (Gosset 1908).

- (a) This is a matched-pairs experiment. Explain why it is inappropriate to use the grouped-means method of determining the difference between the two seeds.
  - (b) Using the matched-pairs platform, determine if there is a difference in yield between kiln-dried and regular-dried corn.
3. The sample data table *Companies.jmp* (*Fortune Magazine*, 1990) contains data on sales, profits, and employees for two different industries (Computers and Pharmaceutical). This exercise is interested in detecting differences between the two types of companies.
- (a) Suppose you wanted to test for differences in sales amounts for the two business types. First, examine histograms of the variables *Type* and *Sales \$* and comment on the output.
  - (b) In comparing sales for the two types of companies, should you use grouped means or matched pairs for the test?
  - (c) Using your answer in part (b), determine if there is a difference between the sales amounts of the two types of companies.
  - (d) Should you remove any outliers in your analysis of part (c)? Comment on why this would or would not be appropriate in this situation.
4. The sample data table *Cars.jmp* (Henderson and Velleman, 1981) contains information on several different brands of cars, including number of doors and impact compression for various parts of the body during crash tests.
- (a) Is there a difference between two- and four-door cars when it comes to impact compression on left legs?
  - (b) Is there a difference between two- and four-door cars when it comes to compression on right legs?
  - (c) Is there a difference between two- and four-door cars when it comes to head impact compression?
5. The sample data table *Chamber.jmp* represents electrical measurements on 24 electrical boards. (This is the same data used in “Paired Means: The Wilcoxon Signed-Rank Test” on page 211.) Each measurement was taken when soldering was complete and then again three weeks later after sitting in a temperature- and humidity-controlled chamber. The investigator wants to know if there is a difference between the measurements.
- (a) Why is this a situation that calls for a matched-pairs analysis?

- (b) Using the paired *t*-test, determine if there is a significant difference between the means when the boards were outside versus. inside the chamber.
- (c) Does the analysis in part (b) lead to the same conclusion as the Wilcoxon signed-rank test? Why or why not?
6. A manufacturer of widgets determined the quality of its product by measuring abrasion on samples of finished products. The manufacturer was concerned that there was a difference in the abrasion measurement for the two shifts of workers that were employed at the factory. Use the data stored in the sample data table *Abrasion.jmp* to compute a *t*-test of abrasion comparing the two shifts. Is there statistical evidence for a difference?
7. The manufacturers of a medication were concerned about adverse reactions in patients treated with their drug. Data on adverse reactions is stored in the sample data table *AdverseR.jmp*. The duration of the adverse reaction is stored in the ADR DURATION variable.
- (a) Patients given a placebo are noted with PBO listed in the treatment group variable, while those that received the standard drug regimen are designated with ST\_DRUG. Test whether there is a significant difference in adverse reaction times between the two groups.
- (b) Test whether there is a difference in adverse reaction times based on the gender of the patient.
- (c) Redo the analyses in parts (a) and (b) using a nonparametric test. Do the results differ?
- (d) A critic of the study claims that the weights of the patients in the placebo group are not the same as those of the treatment group. Do the data support the critic's claim?



# 9

## Comparing Many Means: One-Way Analysis of Variance

### Overview

In Chapter 8, “The Difference Between Two Means,” the *t*-test was the tool used to compare the means of two groups. However, if you need to test the means of more than two groups, the *t*-test can’t handle the job because it is defined only for two groups. This chapter shows how to compare more than two means using the one-way *analysis of variance*, or ANOVA for short. The *F*-test, which has made brief appearances in previous chapters, is the key element in an ANOVA. It is the statistical tool necessary to compare many groups, just as the *t*-test compares two groups. This chapter also introduces multiple comparisons, reviews the topic of unequal variances, and extends nonparametric methods to the one-way layout.

## Chapter Contents

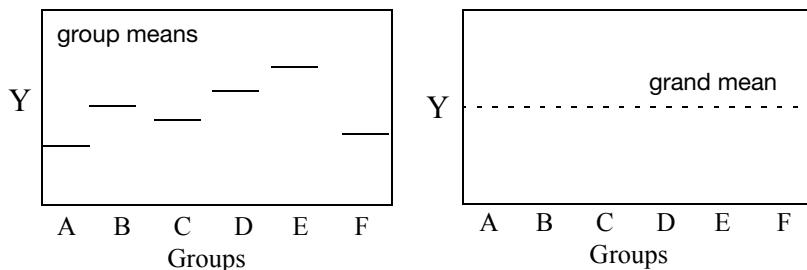
Overview .....	217
What Is a One-Way Layout? .....	219
Comparing and Testing Means .....	221
Means Diamonds: A Graphical Description of Group Means .....	222
Statistical Tests to Compare Means .....	223
Means Comparisons for Balanced Data.....	226
Means Comparisons for Unbalanced Data .....	227
Adjusting for Multiple Comparisons .....	232
Are the Variances Equal across the Groups? .....	235
Testing Means with Unequal Variances .....	238
Nonparametric Methods.....	239
Review of Rank-Based Nonparametric Methods.....	239
The Three Rank Tests in JMP.....	240
Exercises.....	242

## What Is a One-Way Layout?

A one-way layout is the organization of data when a response is measured across a number of groups, and the distribution of the response might be different across the groups. The groups are labeled by a classification variable, which is a column in the JMP data table with the nominal or ordinal modeling type.

Usually, one-way layouts are used to compare group means. **Figure 9.1** shows a schematic that compares two models. The model on the left fits a different mean for each group, and the model on the right indicates a single grand mean (a single-mean model).

**Figure 9.1** Different Mean for Each Group Versus a Single Overall Mean



The previous chapter showed how to use the *t*-test and the *F*-test to compare two means. When there are more than two means, the *t*-test is no longer applicable; the *F*-test must be used.

An *F*-test has the following features:

- An *F*-test compares two models, one constrained and the other unconstrained. The constrained model fits one grand mean. The unconstrained model for the one-way layout fits a mean for each group.
- The measurement of fit is done by accumulating and comparing *sum of the squares* for the constrained and unconstrained models. Note that there can be several types of sums of squares. In this discussion of a one-way layout:

The *total sum of squares*, or *Total SS*, is found by accumulating the squared differences between each point and the grand mean.

The *model sum of squares*, or *Model SS*, is found by accumulating the squared differences between the group means and the grand mean.

The difference between the Total SS and the Mean SS is the *error sum of squares*, or *Error SS*. The Error SS is also called the Residual SS, where a residual is the difference between the actual response and the fitted response (its group mean).

- Degrees of freedom (DF) are numbers, based on the number of parameters and number of data points in the model, that you divide by to get an unbiased estimate of the variance (see Chapter 5, “What Are Statistics?,” for a definition of bias). As with sums of squares, there are different degrees of freedom:

The total degrees of freedom (DF) is the total number of data points minus one (*Total DF*).

The *Model DF* is the number of groups minus one.

The *Error DF* is the total DF minus the model DF.

- A Mean Square is calculated by dividing a sum of squares by its associated degrees of freedom. Mean Squares are estimates of variance, sometimes under the assumption that certain hypotheses are true. As with sums of squares, there are different mean squares, two of which are very important:

The model sum of squares divided by its degrees of freedom is called the model mean square or *Model MS*.

The error sum of squares divided by its DF is called the *Error MS*.

- An *F*-statistic is a ratio of Mean Squares (MS) that are independent and have the same expected value. In our discussion, this ratio is

$$\frac{\text{Model MS}}{\text{Error MS}}$$

- If the null hypothesis that there is no difference between the means is true, this *F*-statistic has an *F distribution*. The Model MS doesn't reflect much more variation than the Error MS. That is, fitting a model doesn't explain any more variation than just fitting the grand mean model.
- If the hypothesis is not true (if there is a difference between the means), the mean square for the model in the numerator of the *F*-ratio includes some effect besides the error variance. This numerator produces a large (and significant) *F* if there is enough data.
- When there is only one comparison (only two groups), the *F*-test is equivalent to the pooled (equal-variance) *t*-test. In fact, when there is only one comparison, the *F*-statistic is the square of the pooled *t*-statistic. This is true

despite the fact that the  $t$ -statistic is derived from the distribution of the estimates, whereas the  $F$ -test is thought of in terms of the comparison of variances of residuals from two different models.

## Comparing and Testing Means

The sample data table Drug.jmp contains the results of a study that measured the response of 30 subjects to treatment by one of three drugs (Snedecor and Cochran, 1967).

- ❖ To begin, select **Help > Sample Data Library** and open Drug.jmp.

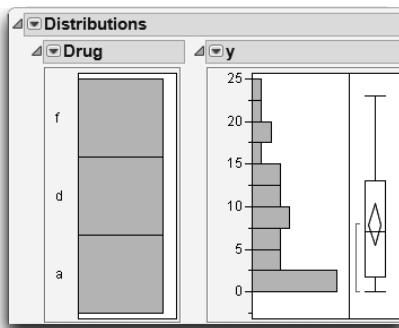
The three drug types are called “a”, “d”, and “f.” The y column is the response measurement. (The x column is used in a more complex model, covered in Chapter 14, “Fitting Linear Models.”)

	Drug	x	y
1	a	11	6
2	a	8	0
3	a	5	2
4	a	14	8
5	a	19	11
6	a	6	4
7	a	10	13
8	a	6	1
9	a	11	8
10	a	3	0
11	d	6	0
12	d	6	2
13	d	7	3

- ❖ For a quick look at the data, select **Analyze > Distribution** and assign Drug and y to **Y, Columns**.

Note in the histogram on the left in **Figure 9.2** that the number of observations is the same in each of the three drug groups; that is what is meant by a *balanced design*.

**Figure 9.2** Distributions of Model Variables

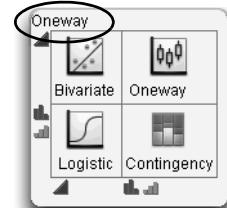


- Next, select **Analyze > Fit Y by X** and assign Drug to **X, Factor** and y to **Y, Response**.

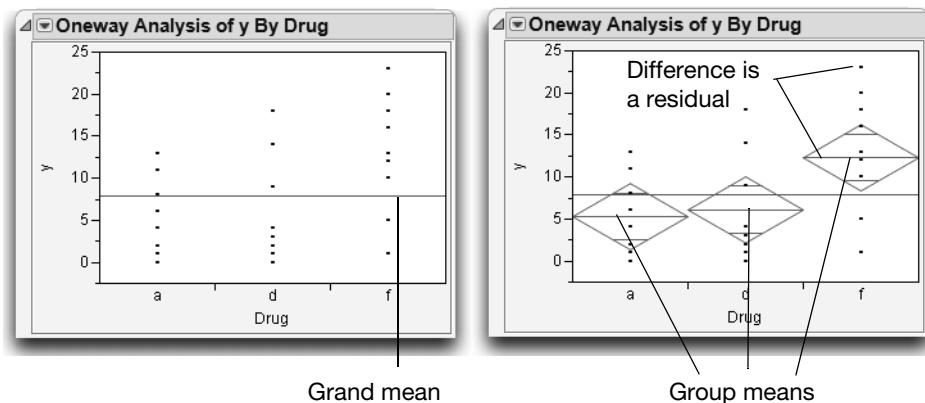
Notice that the launch window displays the message above the analysis type legend that you are requesting a one-way analysis.

- Click **OK**.

The results window on the left in **Figure 9.3** appears. The initial plot on the left shows the distribution of the response in each drug group. The line across the middle is the grand mean. We want to test the null hypothesis that there is no difference in the response among the groups.



**Figure 9.3** Distributions of Drug Groups

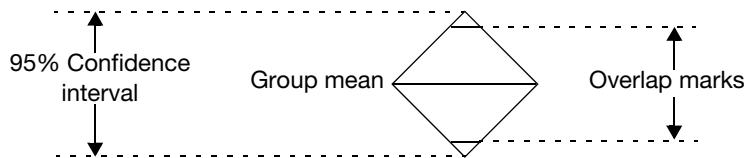


## Means Diamonds: A Graphical Description of Group Means

- Select **Means/Anova** from the red triangle menu next to Oneway Analysis.

This adds means diamonds to the plot and also adds a set of reports. The plot on the right in **Figure 9.3** shows mean diamonds:

- The middle line in the diamond is the response group mean for the group.
- The vertical endpoints form the 95% confidence interval for the mean.
- The x-axis is divided proportionally by group sample size. If we have the same number of observations per group, the x-axis is evenly divided.



If the means are not much different, they are close to the grand mean. If the confidence intervals (the points of the diamonds) of groups don't overlap, the means are significantly different.

Later sections in this chapter show details and interpretation rules for means diamonds.

## Statistical Tests to Compare Means

The **Means/Anova** command produces a report composed of the three tables shown in **Figure 9.4**:

- The **Summary of Fit** table gives an overall summary of how well the model fits.
- The **Analysis of Variance** table gives sums of squares and an *F*-test on the means.
- The **Means for Oneway Anova** table shows the group means, standard error, and upper and lower 95% confidence limits on each mean.

**Figure 9.4** One-Way ANOVA Report

Oneway Analysis of y By Drug					
Oneway Anova					
Summary of Fit					
Rsquare	0.227826				
Adj Rsquare	0.170628				
Root Mean Square Error	6.070878				
Mean of Response	7.9				
Observations (or Sum Wgts)	30				
Analysis of Variance					
Sum of					
Source	DF	Squares	Mean Square	F Ratio	Prob > F
Drug	2	293.6000	146.800	3.9831	0.0305*
Error	27	995.1000	36.856		
C. Total	29	1,288.7000			
Means for Oneway Anova					
Level	Number	Mean	Std Error	Lower 95%	Upper 95%
a	10	5.3000	1.9198	1.3609	9.239
d	10	6.1000	1.9198	2.1609	10.039
f	10	12.3000	1.9198	8.3609	16.239

Std Error uses a pooled estimate of error variance

The Summary of Fit and the Analysis of Variance tables might look like a hodgepodge of numbers, but they are all derived by a few simple rules. **Figure 9.5** illustrates how the statistics relate.

**Figure 9.5** Summary of Fit and ANOVA Tables

<b>Oneway Anova</b>					
<b>Summary of Fit</b>					
Rsquare		0.227826			
Adj Rsquare		0.170628			
Root Mean Square Error		6.070878			
Mean of Response		7.9			
Observations (or Sum Wgts)		30			

<b>Analysis of Variance</b>					
<b>Source</b>	<b>DF</b>	<b>Sum of</b>		<b>F Ratio</b>	<b>Prob &gt; F</b>
		<b>Squares</b>	<b>Mean Square</b>		
Drug	2	293.6000	146.800	3.9831	0.0305*
Error	27	995.1000	36.856		
C. Total	29	1,288.7000			

$$\text{C. Total} = \text{Model} + \text{Error}$$

$$\text{Total DF} = 2 + 27 = 29$$

$$\text{Total SS} = 293.6 + 995.1 = 1288.7$$

$$\text{Mean Square Model} = \frac{\text{SS}}{\text{df}} = \frac{293.6}{2} = 146.8$$

$$\text{Mean Square Error} = \frac{\text{SS}}{\text{df}} = \frac{995.1}{27} = 36.856$$

$$\text{F-Ratio} = \frac{\text{MS (Model)}}{\text{MS (Error)}} = \frac{146.8}{36.856} = 3.9831$$

The Analysis of Variance table (**Figure 9.4** and **Figure 9.5**) describes three source components:

### C. Total

The C. Total Sum of Squares (SS) is the sum of the squares of residuals around the grand mean. C. Total stands for *corrected total* because it is corrected for the mean. The C. Total degrees of freedom is the total number of observations in the sample minus 1.

### Error

After you fit the group means, the remaining variation is described in the Error line. The Sum of Squares is the sum of squared residuals from the individual means. The remaining unexplained variation is C. Total minus Model (labeled Drug in this example). It is called the Error sum of squares. The Error Mean Square estimates the variance.

### Model

The Sum of Squares for the Model line is the difference between C. Total and Error. It is a measure of how much the residuals' sum of squares is accounted for by fitting the model rather than fitting only the grand mean. The degrees of freedom in the drug example is the number of parameters in the model (the number of groups, 3) minus 1.

Everything else in the Analysis of Variance table and the Summary of Fit table is derived from these quantities.

### Mean Square

*Mean Squares* are the sum of squares divided by their respective degrees of freedom.

### F-ratio

The *F-ratio* is the model mean square divided by the error mean square. The *p*-value for this *F*-ratio comes from the *F*-distribution.

### RSquare

The *Rsquare* ( $R^2$ ) is the proportion of variation explained by the model. In other words, it is the model sum of squares divided by the total sum of squares.

### Adjusted RSquare

The *Adjusted Rsquare* is more comparable over models with different numbers of parameters (degrees of freedom). It is the error mean square divided by the total mean square, subtracted from one:

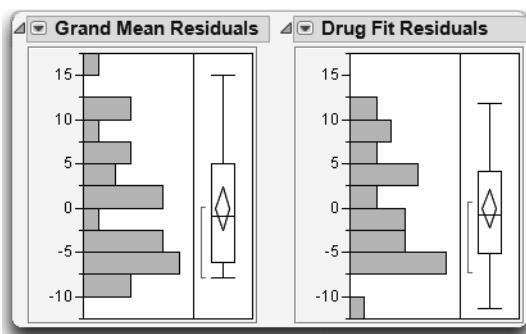
$$1 - \frac{\text{Error MS}}{\text{Total MS}}$$

### Root Mean Square Error

The Root Mean Square Error is the square root of the Mean Square for Error in the Analysis of Variance table. It estimates the standard deviation of the error.

So what's the verdict for the null hypothesis that the group means are the same? The *F-ratio* of 3.98 is significant with a *p*-value of 0.03, which leads us to reject the null hypothesis and confirms that there is a significant statistical difference in the means. The *F*-test does not give any specifics about which means are different, only that at least one pair of means that is statistically different.

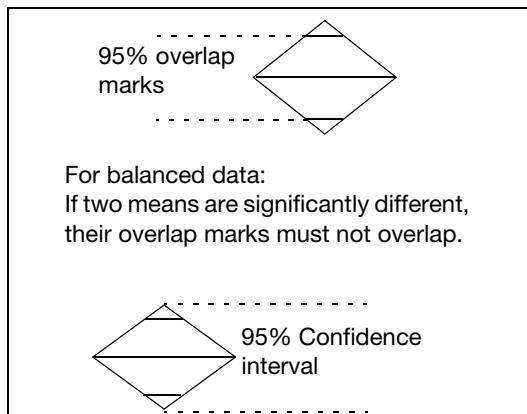
Recall that the  $F$ -test shows whether the variance of residuals from the model is smaller than the variance of the residuals from only fitting a grand mean. In this case, the answer is yes, but just barely. The histograms shown here compare the residuals from the grand means (left) with the group mean residuals (right). Note that these residuals were calculated using the Formula Editor for illustration purposes.



## Means Comparisons for Balanced Data

At this point, we know there is at least one pair of means that are different. Which means are significantly different from which other means? It looks like the mean for the drug “f” (the placebo) is separate from the other two (see **Figure 9.3**). However, since all the confidence intervals for the means intersect, it takes further digging to see significance.

Suppose that two means are from samples with the same number of observations. You could use the overlap marks to get a more precise graphical measure of which means are significantly different. Two means are significantly different when their overlap marks don't overlap. The overlap marks are placed into the confidence interval at a distance of  $1/\sqrt{2}$ , a distance given by the Student's  $t$ -test of separation.



**Note:** Use the crosshairs tool from the toolbar or the **Tools** menu to get a better view of the potential overlap between means diamonds.

When two means do not have the same number of observations, the design is unbalanced and the overlap marks no longer apply. For these cases, JMP provides another technique using *comparison circles* to compare means. The next section describes comparison circles and shows you how to interpret them.

## Means Comparisons for Unbalanced Data

Suppose, for the sake of this example, that the drug data are unbalanced. That is, there is not the same number of observations in each group. The following steps unbalance the Drug.jmp sample data in an extreme way to illustrate an apparent paradox, as well as introduce a new graphical technique.

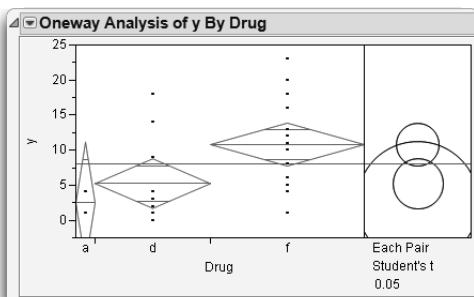
- ☞ Change Drug in rows 1, 4–7, and 9 from “a” to “f”.
- ☞ Change Drug in rows 2 and 3 to “d”.
- ☞ Change y in row 10 to “4.” (Be careful not to save this modified table over the original copy in your sample data.)

Now drug “a” has only two observations, whereas “d” has 12 and “placebo” has 16. The mean for “a” has a very high standard error because it is supported by so few observations compared with the other two levels.

Again, use the **Fit Y by X** command to look at the data:

- ☞ Select **Analyze > Fit Y by X** for the modified data with y as **Y, Response** and Drug as **X, Factor**, and then click **OK**.
- ☞ Select the **Means/Anova** option from the red triangle menu next to Oneway Analysis.
- ☞ Select **Compare Means > Each Pair, Student's t** from the red triangle menu next to Oneway Analysis.

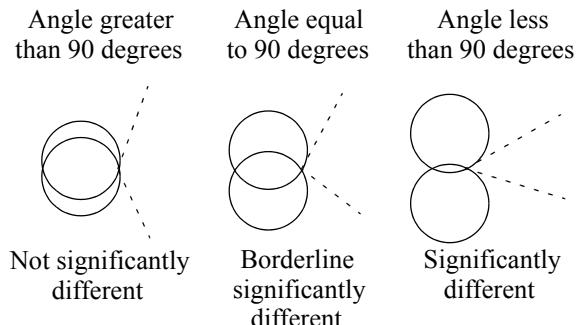
The modified data should give results like those illustrated in **Figure 9.6**. The *x*-axis divisions are proportional to the group sample size, which causes drug “a” to be very thin, because it has fewer observations. The confidence interval on its mean is large compared with the others. Comparison circles for Student's *t*-tests appear to the right of the means diamonds.

**Figure 9.6** Comparison Circles to Compare Group Means

Comparison circles are a graphical technique that lets you see significant separation among means in terms of how the circles intersect. This is the only graphical technique that works in general with both equal and unequal sample sizes. The plot displays a circle for each group, with the centers lined up vertically. The center of each circle is aligned with its corresponding group mean. The radius of a circle is the 95% confidence interval for its group mean, as you can see by comparing a circle with its corresponding means diamond. The non-overlapping confidence intervals shown by the diamonds for groups that are significantly different correspond directly to the case of non-intersecting comparison circles.

When the circles intersect, the angle of intersection is the key to seeing if the means are significantly different. If the angle of intersection is exactly a right angle ( $90^\circ$ ), then the means are on the borderline of being significantly different. For more information about the geometry of comparison circles, select **Help > JMP Help** and refer to the *Basic Analysis* book.

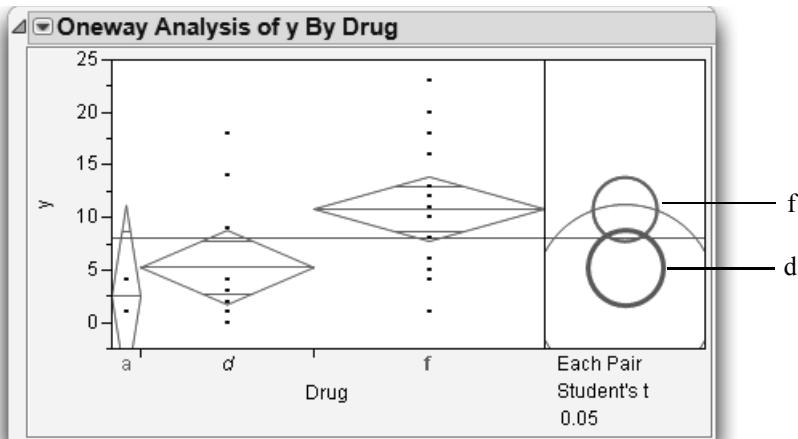
If the circles are farther apart than the right angle case, then the outside angle is more acute and the means are significantly different. If the circles are closer together, the angle is larger than a right angle, and the means are not significantly different. **Figure 9.7** illustrates these angles of intersection.

**Figure 9.7** Diagram of How to Interpret Comparison Circles

So what are the conclusions for the drug example shown in **Figure 9.6**?

You don't need to hunt down a protractor to figure out the size of the angles of intersection. Click on a circle to see what happens (see **Figure 9.8**). The circle highlights and becomes red. Groups that are not different from it also show in red. All groups that are significantly different are gray.

- Click on the “f” circle and use the circles to compare group means.

**Figure 9.8** Clicking on Comparison Circles to See Group Differences

- The “f” and “d” means are represented by the smaller circles, since they are based on more observations. The circles are separated farther than would occur with a right angle. The angle is acute, so these two means are significantly different. This is shown by their different color.

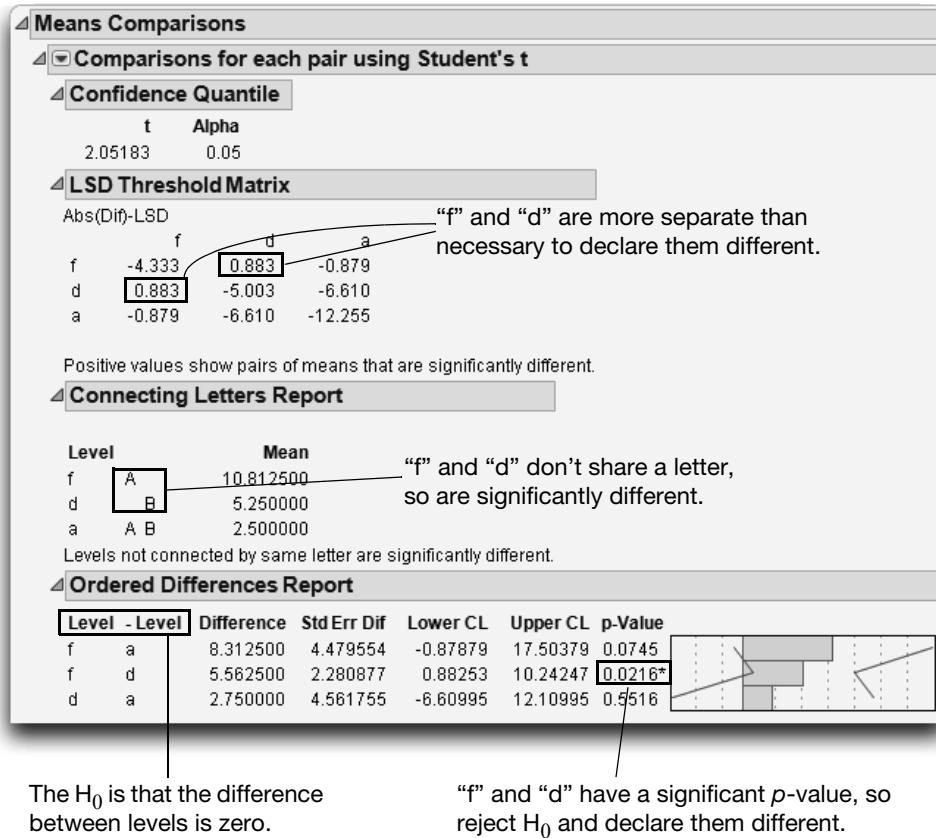
- The circle for the “d” mean is completely nested in the circle for “a”, so they are not significantly different.
- The “a” mean is well below the “d” mean, which is significantly below “f.” By transitivity, you might expect “a” to be significantly different from “f.” The problem with this logic is that the standard error around the “a” mean is so large that it is not significantly different from “f”, even though it is farther away than “d.”

The group differences found with the comparison circles can be verified statistically with the Means Comparisons tables shown in **Figure 9.9**.

The Means Comparisons table uses the concept of Least Significant Difference (LSD). In the balanced case, this is the separation that any two means must have from each other to be significantly different. In the unbalanced case, there is a different LSD for each pair of means.

The Means Comparison report shows all of the comparisons of means ordered from high to low. The elements of the table LSD Threshold Matrix show the absolute value of the difference in two means minus the LSD. If the means are farther apart than the LSD, the element is positive and they are significantly different. For example, the element that compares “f” and “d” is +0.88, which says that the means are 0.88 more separate than needed to be significantly different. If the means are not significantly different, the LSD is greater than the difference. Therefore, the element in the table is negative. The elements for the other two comparisons are negative, showing no significant difference.

In addition, a table shows the classic SAS style means comparison with letters in the Connecting Letters report. Levels that share a letter are not significantly different from each other. For example, both levels “d” and “a” share the letter B, so “d” and “a” are not significantly different from each other.

**Figure 9.9** Statistical Text Reports to Compare Groups

The Ordered Differences report in **Figure 9.9** lists the differences among groups in decreasing order, with confidence limits of the difference. A bar chart displays the differences with blue lines representing the confidence limits. The  $p$ -value tests the  $H_0$  that there is no difference in the means.

The last thing to do in this example is to restore your copy of the Drug.jmp sample data table to its original state so that it can be used in other examples. To do this,

**Important:** Select **File > Revert** or reopen the data table.

## Adjusting for Multiple Comparisons

Making multiple comparisons, such as comparing many pairs of means, increases the possibility of committing a Type I error. Remember, a Type I error is the error of declaring a difference significant (based on statistical test results) that is actually not significant. We are satisfied with a 1 in 20 (5%) chance of committing a Type I error. However, the more tests you do, the more likely you are to happen upon a significant difference occurring by chance alone. If you compare all possible pairs of means in a large one-way layout with many different levels, there are many possible tests, and a Type I error becomes very likely.

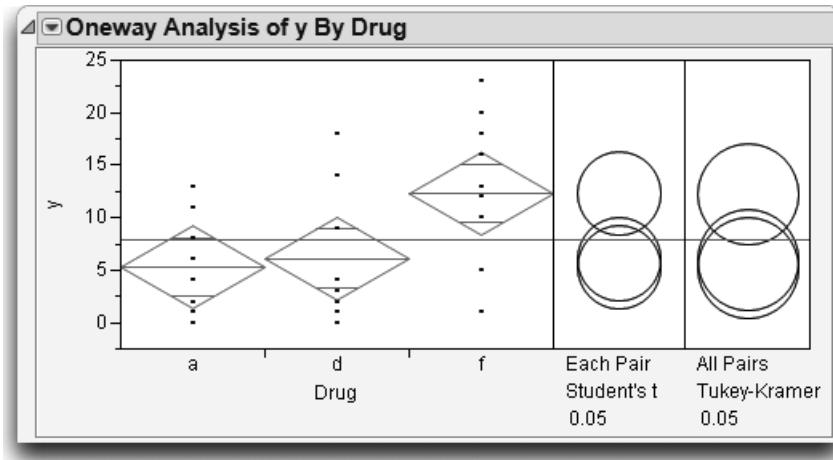
There are many methods that modify tests to control for an overall error rate. This section covers one of the most basic, the *Tukey-Kramer Honestly Significant Difference* (HSD). The Tukey-Kramer HSD uses the distribution of the maximum range among a set of random variables to attempt to control for the multiple comparison problem.

- ☞ Select **Help > Sample Data Library** and open the original copy of Drug.jmp.
- ☞ Select **Analyze > Fit Y by X.**
- ☞ Assign Drug to **X, Factor** and y to **Y, Response**. Because the analysis is the same as before, you can also expedite the launch window entries by clicking the **Recall** button.

Select the following three commands from the red triangle menu on the title bar:

- ☞ **Means/Anova**
- ☞ **Compare Means > Each Pair, Student's t,**
- ☞ **Compare Means > All Pairs, Tukey HSD.**

These commands should give you the results shown in **Figure 9.10**.

**Figure 9.10** *t*-tests and Tukey-Kramer Adjusted *t*-tests for One-Way ANOVA

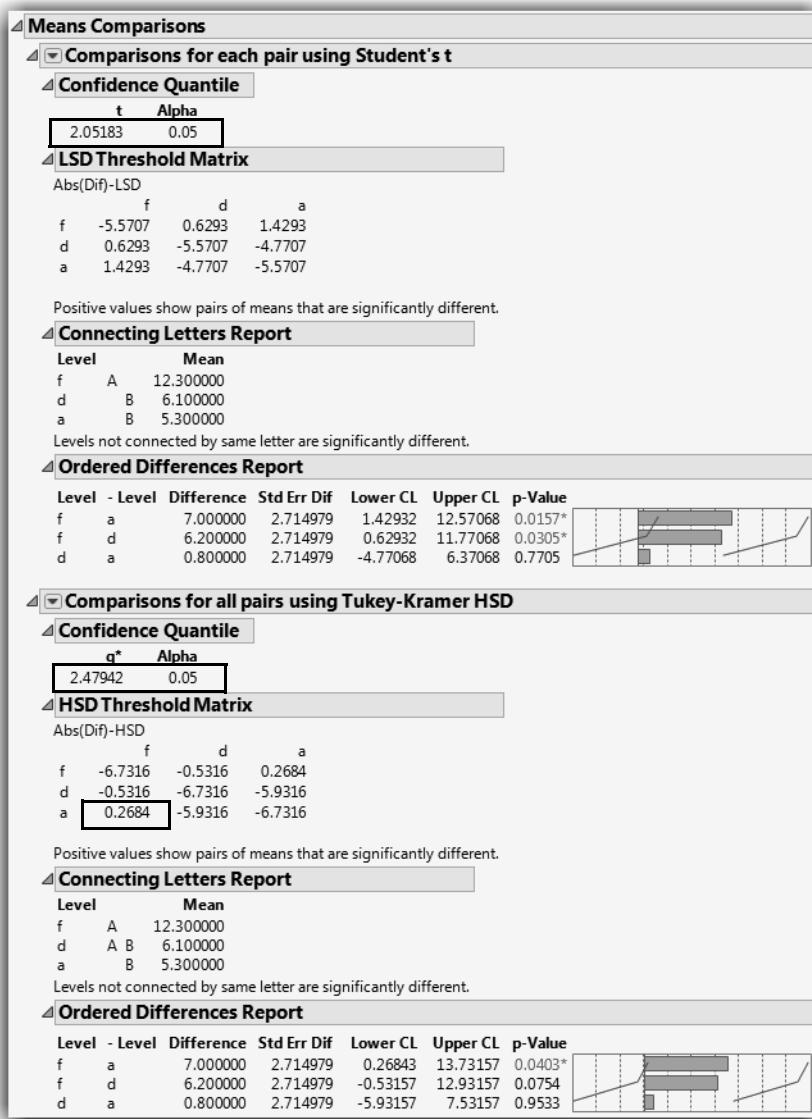
The comparison circles work as before, but have different types of error rates.

The Tukey-Kramer comparison circles are larger than the Student's *t* circles. This protects more tests from falsely declaring significance, but this protection makes it harder to declare two means significantly different.

If you click on the top circle, you see that the conclusion is different between the Student's *t* and Tukey-Kramer's HSD for the comparison of "f" and "d." This comparison is significant for Student's *t*-test but not for Tukey's test.

The difference in significance occurs because the quantile that is multiplied into the standard errors to create a Least Significant Difference has grown from 2.05 to 2.48 between Student's *t*-test and the Tukey-Kramer test (see the Confidence Quantiles in **Figure 9.11**).

The only positive element in the Tukey table is the one for the "a" versus "f" comparison (**Figure 9.11**).

**Figure 9.11** Means Comparisons Table for One-Way ANOVA

## Are the Variances Equal across the Groups?

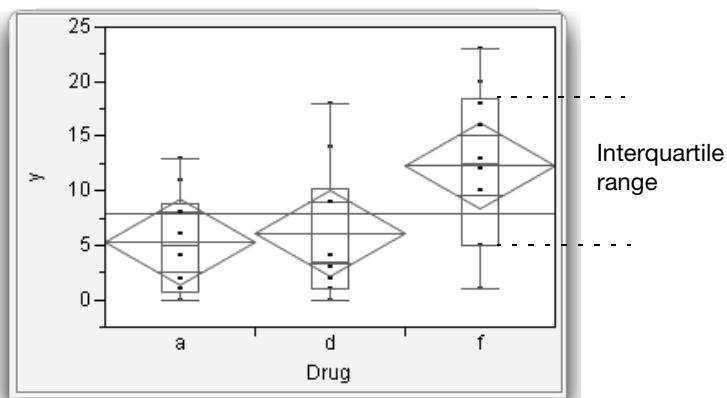
The one-way ANOVA assumes that each group has the same variance. The Analysis of Variance table shows the note “Std Error uses a pooled estimate of error variance.” When testing the difference between two means, as in the previous chapter, JMP provides separate reports for both equal and unequal variance assumptions. This is why, when there are only two groups, the command is **Means/Anova/Pooled t**. ANOVA pools the variances like the pooled *t*-test does.

Before you become too concerned about the equal-variance issue, be aware that there is always a list of issues to worry about; it is not usually useful to be overly concerned about this one.

- ✓ Select **Quantiles** from the red triangle menu next to Oneway Analysis.

This command displays quantile box plots for each group as shown in **Figure 9.12**. Note that the interquartile range (the height of the boxes) is not much different for drugs a and d, but is somewhat different for the placebo (f). The placebo group seems to have a slightly larger interquartile range.

**Figure 9.12** Quantile Box Plots



- ✓ Select **Quantiles** again to turn the box plots off.

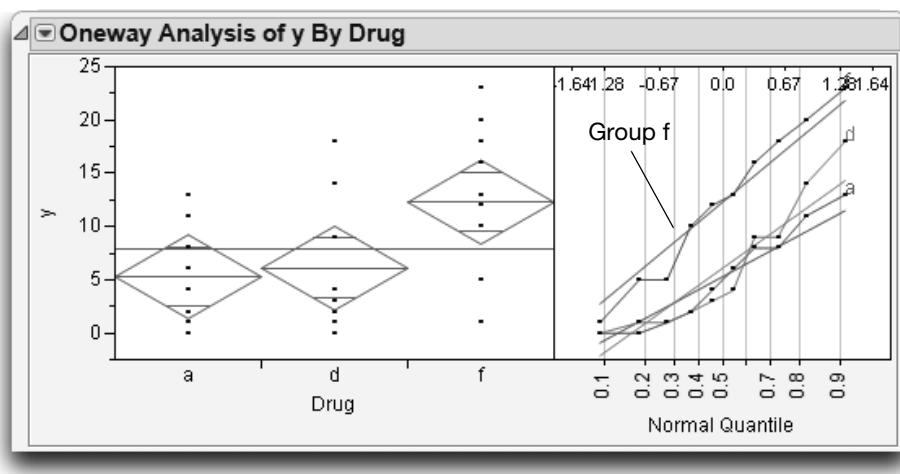
A more effective graphical tool to check the variance assumption is the normal quantile plot.

- ✓ Select **Normal Quantile Plot > Plot Actual By Quantile** from the red triangle menu next to Oneway Analysis.

This option displays a plot next to the Means Diamonds as shown in **Figure 9.13**. The normal quantile plot compares mean, variance, and shape of the group distributions.

There is a line on the normal quantile plot for each group. The height of the line shows the location of the group. The slope of the line shows the group's standard deviation. So, lines that appear to be parallel have similar standard deviations. The straightness of the line segments connecting the points shows how close the shape of the distribution is to the normal distribution. Note that the "f" group is both higher and has a greater slope, which indicates a higher mean and a higher variance, respectively.

**Figure 9.13** Normal Quantile Plot



It's easy to get estimates of the standard deviation within each group:

- ☛ Select **Means and Std Dev** from the red triangle menu next to Oneway Analysis to see the reports in **Figure 9.14**.

**Figure 9.14** Mean and Standard Deviation Report

Means and Std Deviations						
Level	Number	Mean	Std Dev	Std Err		
				Mean	Lower 95%	Upper 95%
a	10	5.3000	4.64399	1.4686	1.9779	8.622
d	10	6.1000	6.15449	1.9462	1.6973	10.503
f	10	12.3000	7.14998	2.2610	7.1852	17.415

You can conduct a statistical test of the equality of the variances as follows:

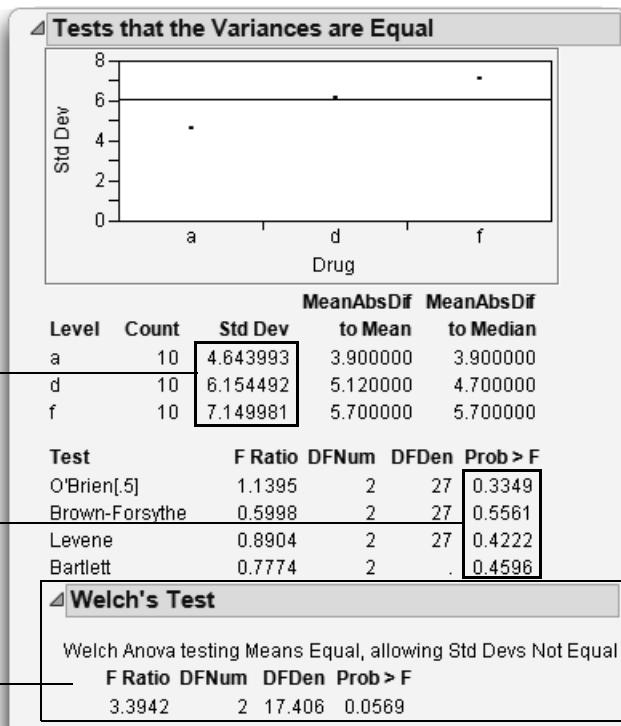
- ✓ Select **Unequal Variances** from the red triangle menu next to Oneway Analysis to see the Tests that the Variances are Equal tables in **Figure 9.15**.

**Figure 9.15** Tests that the Variances are Equal Report

This report shows the standard deviations.

Small *p*-values represent unequal variances.

Use the Welch ANOVA if the variances are not statistically equal.



To interpret these reports, note that the **Std Dev** column lists the estimates that you are testing to be the same. The null hypothesis is that the standard deviations are equal. Then note the results listed under **Prob>F**. As expected, there is no evidence that the variances are unequal. None of the *p*-values are small.

Each of the four tests in **Figure 9.15** (O'Brien, Brown-Forsythe, Levene, and Bartlett) test the null hypothesis that the variances are equal, but each uses a different method for measuring variability.

One way to evaluate dispersion is to take the absolute value of the difference of each response from its group mean. Mathematically, we look at  $|x_i - \bar{x}|$  for each response.

- *Levene's Test* estimates the mean of these absolute differences for each group (shown in the table as **MeanAbsDif to Mean**). The test then does a *t*-test (or equivalently, an *F*-test) on these estimates.
- The *Brown-Forsythe Test* measures the differences from the median instead of the mean and then tests these differences.
- *O'Brien's Test* tricks the *t*-test by telling it that the means were really variances.
- *Bartlett's Test* derives the test mathematically, using an assumption that the data are normal. Though powerful, Bartlett's test is sensitive to departures from the normal distribution.

Statisticians have no apologies for offering different tests with different results. Each test has its advantages and disadvantages.

## Testing Means with Unequal Variances

If you think the variances are different, you should consider avoiding the standard *t*- and *F*-tests that assume the variances are equal. Instead, use the Welch ANOVA *F*-test that appears with the unequal variance tests (**Figure 9.15**). The test can be interpreted as an *F*-test in which the observations are weighted by an amount inversely proportional to the variance estimates. This has the effect of making the variances comparable.

The *p*-values might disagree slightly with those obtained from other software that provide similar unequal-variance tests. These differences arise because some methods round or truncate the denominator degrees of freedom for computational convenience. JMP uses the more accurate fractional degrees of freedom.

In practice, the hope is that there are not conflicting results from different tests of the same hypothesis. However, conflicting results do occasionally occur, and there is an obligation to report the results from all reasonable perspectives.

# Nonparametric Methods

JMP also offers nonparametric methods in the Fit Y by X platform. Nonparametric methods, introduced in the previous chapter, use only the rank order of the data and ignore the spacing information between data points. Nonparametric tests do not assume that the data have a normal distribution. This section first reviews the rank-based methods and then generalizes the Wilcoxon rank-sum method to the  $k$  groups of the one-way layout.

## Review of Rank-Based Nonparametric Methods

Nonparametric tests are useful to test whether means or medians are the same across groups. However, the usual assumption of normality is not made.

Nonparametric tests use functions of the response ranks, called *rank scores* (Hajek 1969).

JMP offers the following nonparametric tests for testing the null hypothesis that distributions across factor levels are centered at the same location. Each is the most powerful rank test for a certain distribution, as indicated in Table 9.1.

- Wilcoxon rank scores are the ranks of the data.
- Median rank scores are either 1 or 0 depending on whether a rank is above or below the median rank.
- Van der Waerden rank scores are the quantiles of the standard normal distribution for the probability argument formed by the rank divided by  $n-1$ . This is the same score that is used in the normal quantile plots.

In addition to the rank-based nonparametric methods, the Kolmogorov-Smirnov Test is available when the X factor has two levels. It uses the empirical distribution function to test whether the distribution of the response is the same across groups.

**Note:** Exact versions of the rank-based methods and the Kolmogorov-Smirnov test are available in JMP Pro when the factor has two levels.

**Table 9.1.** Guide for Using Rank-Based Nonparametric Tests

Fit Y By X Nonparametric Option	Two Levels	Two or More Levels	Most Powerful for Errors Distributed as
<b>Wilcoxon Test</b>	Wilcoxon rank-sum (Mann-Whitney U)	Kruskal-Wallis	Logistic
<b>Median Test</b>	Two-Sample Median	$k$ -Sample Median (Brown-Mood)	Double Exponential
<b>van der Waerden Test</b>	Van der Waerden	$k$ -sample Van der Waerden	Normal

## The Three Rank Tests in JMP

For example, use the Drug.jmp example and request nonparametric tests to compare the Drug group means of  $y$ , the response:

- ☞ Select **Analyze > Fit Y by X** and assign  $y$  to **Y, Response** and Drug to **X, Factor**.

The Oneway analysis of variance platform appears showing the distributions of the three groups, as seen previously in **Figure 9.3**.

- ☞ Select the four tests below from the red triangle menu to compare groups.  
All groups have the null hypothesis that the groups do not differ:

- **Means/Anova**, producing the  $F$ -test from the standard parametric approach
- **Nonparametric > Wilcoxon Test**, also known as the Kruskal-Wallis test when there are more than two groups
- **Nonparametric > Median Test** for the median test
- **Nonparametric > van der Waerden Test** for the Van der Waerden test

**Figure 9.16** shows the results of the four tests that compare groups. In this example, the Wilcoxon and the Van der Waerden agree with the parametric  $F$ -test in the ANOVA. They show borderline significance for a 0.05  $\alpha$ -level, despite a fairly small sample and the possibility that the data are not normal. These tests reject the null and detect a difference in the groups.

The median test is much less powerful than the others and does not detect a difference in this example.

**Note:** Several nonparametric multiple comparison procedures are available. Select **Help > JMP Help** and refer to the *Basic Analysis* book for details.

**Figure 9.16** Parametric and Nonparametric Tests for Drug Example

**Oneway Analysis of y By Drug**

**Oneway Anova**

**Analysis of Variance**

Source	DF	Squares	Mean Square	F Ratio	Prob > F
Drug	2	293.6000	146.800	3.9831	0.03051
Error	27	995.1000	36.856		
C. Total	29	1,288.7000			

Significant

**Means for Oneway Anova**

**Wilcoxon / Kruskal-Wallis Tests (Rank Sums)**

Level	Count	Score Sum	Expected	(Mean-Mean0)/Std0	
			Score		Score Mean
a	10	122.000	155.000	12.2000	-1.433
d	10	132.500	155.000	13.2500	-0.970
f	10	210.500	155.000	21.0500	2.425

**1-way Test, ChiSquare Approximation**

ChiSquare	DF	Prob>ChiSq
6.0612	2	0.04831

Significant

**Median Test (Number of Points Above Median)**

Level	Count	Score Sum	Expected	(Mean-Mean0)/Std0	
			Score		Score Mean
a	10	4.000	5.000	0.400000	-0.762
d	10	4.000	5.000	0.400000	-0.762
f	10	7.000	5.000	0.700000	1.523

**1-way Test, ChiSquare Approximation**

ChiSquare	DF	Prob>ChiSq
2.3200	2	0.3135

Not significant

**Van der Waerden Test (Normal Quantiles)**

Level	Count	Score Sum	Expected	(Mean-Mean0)/Std0	
			Score		Score Mean
a	10	-3.693	0.000	-0.36929	-1.568
d	10	-2.245	0.000	-0.22445	-0.953
f	10	5.937	0.000	0.59374	2.521

**1-way Test, ChiSquare Approximation**

ChiSquare	DF	Prob>ChiSq
6.4804	2	0.03921

Significant

## Exercises

1. This exercise uses the sample data table **Movies.jmp**, which contains the top-grossing movies of all time as of 2003. You are interested in discovering if there is a difference in earnings between the different classifications of movies.
  - (a) Use the Distribution platform to examine the variable **Type**. How many levels are there? Are there approximately equal numbers of movies of each type?
  - (b) Use the Fit Y by X platform to perform an ANOVA, with **Type** as X and **Worldwide \$** as Y. State the null hypothesis that you are testing. Does the test show differences among the different types?
  - (c) Use comparison circles to explore the differences. Does it appear that **Action** and **Drama** differ significantly from all other types?
  - (d) Examine a normal quantile plot of this data and comment on the equality of the variances of the groups. Are they different enough to require a Welch ANOVA? If so, conduct one and comment on its results.
2. The National Institute of Standards and Technology (NIST) references research involving the doping of silicon wafers with phosphorus. Twenty-five wafers were doped with phosphorus by neutron transmutation doping in order to have nominal resistivities of 200 ohm/cm. Each data point is the average of six measurements at the center of each wafer. Measurements of bulk resistivity of silicon wafers were made at NIST with five probing instruments on each of five days. The data were stored in the sample data table **Doped Wafers.jmp**. (See Ehrstein and Croarkin.) The experimenters are interested in testing differences among the probing instruments.
  - (a) Examine a histogram of the resistances. Do the data appear to be normal? Test for normality by stating the null hypothesis, and then conduct a statistical test.
  - (b) State the null hypothesis and conduct an ANOVA to determine whether there is a difference between the probes in measuring resistance.
  - (c) Comment on the sample sizes involved in this investigation. Do you feel confident with your results?

- (d) A retrospective analysis of the power of the test is available under the red triangle for the title bar (under **Power**). Select this option, select the **Solve for Power** check box and click **Done**. What is the power of the test conducted in part b?

**Note:** The command **DOE > Design Diagnostics > Sample Size and Power > k Sample Means** can also be used to explore the power of the test or to determine how many data points would be needed to detect a difference between any two of the means.

3. The sample data table *Michelson.jmp* contains data (as reported by Stigler, 1977) collected to determine the speed of light in air. Five separate collections of data were made in 1879 by Michelson, and the speed of light was recorded in km/sec. The values for velocity in this table have had 299,000 subtracted from them.
  - (a) The true value (accepted today) for the speed of light is 299,792.5 km/sec. What is the mean of Michelson's responses?
  - (b) Is there a significant statistical difference between the trials? Use an ANOVA or a Welch ANOVA (whichever is appropriate) to justify your answer.
  - (c) Using Student's *t* comparison circles, find the group of observations that is statistically different from all the other groups.
  - (d) Does excluding the result in part (c) improve Michelson's prediction?
4. *Run-Up* is a term used in textile manufacturing to denote waste. Manufacturers often use computers to lay out designs on cloth in order to minimize waste, and the percentage difference between human layouts and computer-modeled layouts is the run-up. There are some cases where humans get better results than the computers. Don't be surprised if there are a few negative values for run-up in the sample data table *Levi Strauss Run-Up.jmp* (Koopmans, 1987). The data was gathered from five different supplier plants to determine whether there were differences among the plants.
  - (a) Produce histograms for the values of Run Up for each of the five plants.
  - (b) State a null hypothesis, and then test for differences between supplier plants by using the three nonparametric tests provided by JMP. Do they have similar results? **Note:** Hold down the Alt key (or Option on a Macintosh) before clicking on the red triangle to select all tests at once.
  - (c) Compare these results to results given by an ANOVA and comment on the differences. Would you trust the parametric or nonparametric tests more?

- (d) There are two extreme values. (See the histogram in part a.) Hide and exclude these observations, and re-run the analyses in parts b and c. Do your results, and your ultimate conclusions, change? What does this say about how sensitive ANOVA and the nonparametric tests are to outliers?
5. The sample data table Scores.jmp contains a subset of results from the *Third International Mathematics and Science Study*, conducted in 1995. The data contain information about scores for Calculus and Physics, divided into four regions of the country.
- Is there a difference among regions in Calculus scores?
  - Is there a difference among regions for Physics scores?
  - Do the data fall into easily definable groups? Use comparison circles to explore their groupings.
6. To judge the efficacy of three pain relievers, a consumer group conducted a study of the amount of relief each patient received. The amount of relief is measured by each participant rating the amount of relief on a scale of 0 (no relief) to 20 (complete relief). Results of the study are stored in the sample data table Analgesics.jmp.
- Is there a difference in relief between the males and females in the study? State a null hypothesis, a test statistic, and a conclusion.
  - Conduct an analysis of variance comparing the three types of drug. Is there a significant difference among the three types of drug?
  - Find the mean amount of pain relief for each of the three types of pain reliever.
  - Does the amount of relief differ for males and females? To investigate this question, conduct an analysis of variance on relief versus treatment for each of the two genders. (Conduct a separate analysis for males and females. Assign gender as a By variable.) Is there a significant difference in relief for the female subset? For the male subset?



# 10

## Fitting Curves through Points: Regression

### Overview

Regression is a method of fitting curves through data points. It is a straightforward and useful technique, but people new to statistics often ask about its rather strange name—regression.

Sir Francis Galton, in his 1885 Presidential address before the anthropology section of the British Association for the Advancement of Science (Stigler, 1986), described his study of how tall children are compared with the height of their parents. In this study, Galton defined ranges of parents' heights, and then calculated the mean child height for each range. He drew a straight line that went through the means (as best as he could). He thought he had made a discovery when he found that the child heights tended to be more moderate than the parent heights. For example, if a parent was tall, the children similarly tended to be tall, but not as tall as the parent. If a parent was short, the child tended to be short, but not as short as the parent. This discovery he called *regression to the mean*, with the regression meaning “to come back to.”

Somehow, the term regression became associated with the technique of fitting the line, rather than the process describing inheritance. Galton's data are covered later in this chapter.

This chapter covers the case where there is only one factor and one response—the type of regression situation you can see on a scatterplot.

## Chapter Contents

Overview .....	245
Regression .....	247
Least Squares .....	247
Seeing Least Squares.....	248
Fitting a Line and Testing the Slope .....	250
Testing the Slope By Comparing Models .....	252
The Distribution of the Parameter Estimates .....	255
Confidence Intervals on the Estimates.....	256
Examine Residuals .....	258
Exclusion of Rows.....	258
Time to Clean Up .....	260
Polynomial Models .....	260
Look at the Residuals .....	261
Higher-Order Polynomials .....	261
Distribution of Residuals .....	262
Transformed Fits .....	263
Spline Fit.....	265
Are Graphics Important?.....	266
Why It's Called Regression.....	269
What Happens When X and Y Are Switched?.....	271
Curiosities .....	274
Sometimes It's the Picture That Fools You .....	274
High-Order Polynomial Pitfall .....	275
The Pappus Mystery on the Obliquity of the Ecliptic .....	276
Exercises.....	277

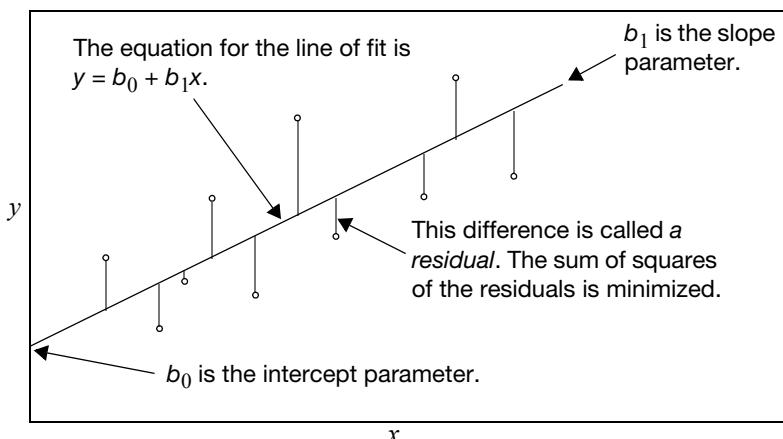
# Regression

Fitting one mean is easy. Fitting several means is not much harder. How do you fit a mean when it changes as a function of some other variable? In essence, how do you fit a line or a curve through data?

## Least Squares

In regression, you select an equation type (linear, polynomial, and so on) and allow the fitting mechanism to determine some of its parameters (coefficients). These parameters are determined by the method of least squares, which finds the parameter values that minimize the sum of squared distances from each point to the line of fit. **Figure 10.1** illustrates a least squares regression line.

**Figure 10.1** Straight-Line Least Squares Regression



For any regression model, the term *residual* denotes the difference between the actual response value and the value predicted by the line of fit. When talking about the true (unknown) model rather than the estimated one, these differences are called the *errors* or *disturbances*.

Least squares regression is the method of fitting of a model to minimize the sum of squared residuals.

The regression line has interesting balancing properties with regard to the residuals. The sum of the residuals is always zero, which was also true for the simple mean fit. You can think of the fitted line as balancing data in the up-and-down direction. If you add the product of the residuals times the  $x$  (regressor)

values, this sum is also zero. This can be interpreted as the line balancing the data in a rotational sense. Chapter 19 shows how these least squares properties can be visualized in terms of the data acting like forces of springs on the line of fit.

An important special case is when the line of fit is constrained to be horizontal (flat). The equation for this fit is a constant. If you constrain the slope of the line to zero, the coefficient of the  $x$  term (regressor) is zero, and the  $x$  term drops out of the model. In this situation, the estimate for the constant is the sample mean. This special case is important because it leads to the statistical test of whether the regressor really affects the response.

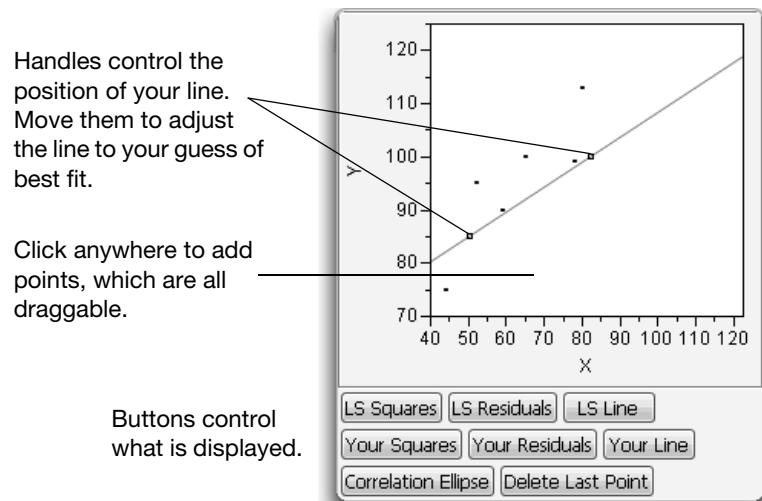
## Seeing Least Squares

The principle of least squares can be seen with one of the sample scripts included in the Sample Scripts folder.

- ☞ Select **Help > Sample Data**, and open the `demoLeastSquares` script from the **Teaching Scripts > Teaching Demonstrations** outline. The script runs automatically.

You should see a plot of points with a line drawn through them (**Figure 10.2**). The line has two small squares (handles) that let you reposition it.

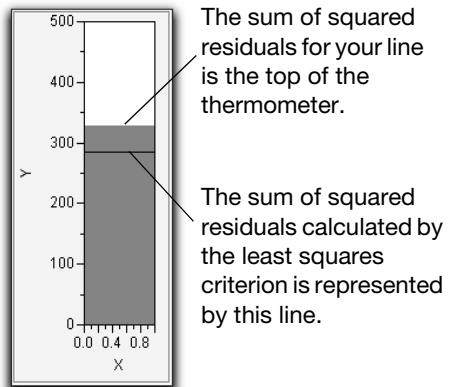
- ☞ Grab the handles and position the line where you think it best runs through (or fits) the data.

**Figure 10.2** demoLeastSquares Display

ⓐ Click the **Your Residuals** button. Again use the handles to move the line around until you think the residuals (in blue) are as small as they can be.

ⓑ Click the **Your Squares** button and try to minimize the total area covered by the blue squares.

To help you to minimize the area of the squares, a second graph is displayed to the right of the scatterplot. Think of it as a “thermometer” that measures the sum of the area of the squares in the scatterplot. The least squares criterion selects the line that minimizes this area. To see the least squares line as calculated by JMP:



ⓐ Click the **LS Line** button to display the least squares line.

ⓑ Click the **LS Residuals** and **LS Squares** buttons to display the residuals and squares for the least squares line.

Notice that a horizontal line has been added in the graph that displays the sum of squares. This represents the sum of the squared residuals from the line calculated by the least squares criterion.

To illustrate that the least squares criterion performs as it claims to:

- ~ Use the handles to drag your line so that it coincides with the least squares line.

The sum of squares is now the same as the sum calculated by the least squares criterion.

- ~ Using one of the handles, move your line off the least squares line, first in one direction, then in the other.

Notice that as your line moves off the line of least squares in any way, the sum of the squares increases. Therefore, the least squares line is truly the line that minimizes the sum of the squared residuals.

**Note:** A more comprehensive version of the demoLeastSquares script, Demonstrate Regression, is available from **Teaching Scripts > Interactive Teaching Modules**.

## Fitting a Line and Testing the Slope

Eppright et al. (1972) as reported in Eubank (1988) measured 72 children from birth to 70 months. You can use regression techniques to examine how the weight-to-height ratio changes as kids grow up.

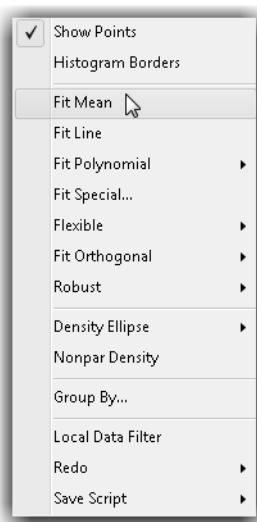
- ~ Select **Help > Sample Data Library** and open Growth.jmp, which holds the Eppright data.
- ~ Select **Analyze > Fit Y by X** and assign ratio to **Y, Response** and age to **X, Factor**. When you click **OK**, the result is a scatterplot of ratio by age.

Click on the red triangle menu next to Bivariate Fit to see the fitting options.

- ☞ Select **Fit Mean** and then **Fit Line** from the red triangle menu next to Bivariate Fit.
- **Fit Mean** draws a horizontal line at the mean of ratio.
- **Fit Line** draws the regression line through the data.

These commands also add statistical tables to the regression report. You should see a scatterplot similar to the one shown in **Figure 10.3**. The statistical tables are actually displayed beneath the scatterplot in the report window, but have been rearranged here to save space.

Each type of fit you select has its own menu icon (found under the scatterplot) that lets you request fitting details.

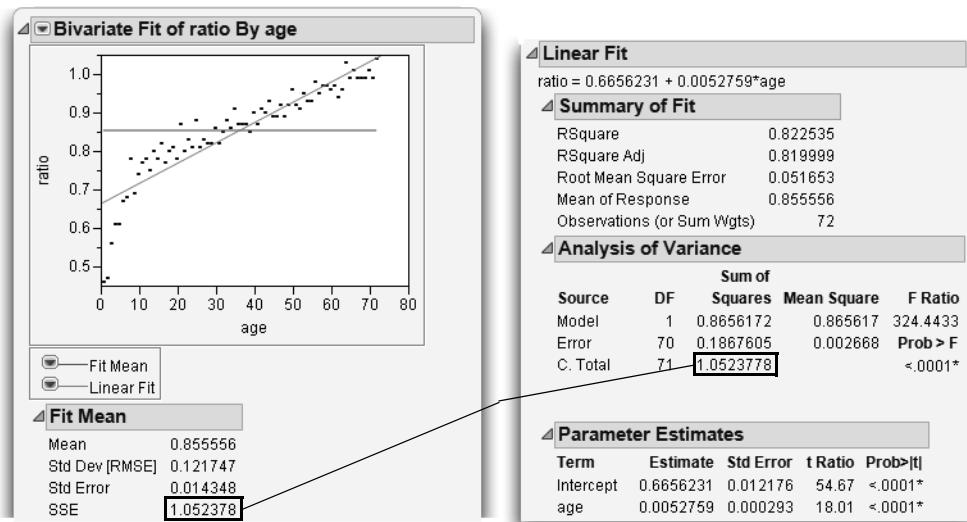


The fitted regression equation is shown under the Linear Fit report (top right in **Figure 10.3**). The equation for the growth data is:

$$\text{ratio} = 0.6656 + 0.005276 \text{ age}$$

The estimated intercept and slope coefficients for this equation are also displayed in the Parameter Estimates report (bottom right in **Figure 10.3**).

**Figure 10.3** Straight-line Least Squares Regression



## Testing the Slope By Comparing Models

If we assume that the linear equation adequately describes the relationship of the weight-to-height ratio with age (which turns out to be incorrect), we have some questions to answer:

- Does the regressor really affect the response?
- Does the ratio of weight to height change as a function of age? Is the true slope of the regression line zero?
- Is the true value for the coefficient of age in the model zero?
- Is the sloped regression line significantly different from the horizontal line at the mean?

Actually, these are all the same question.

Chapter 8, “The Difference Between Two Means,” presented two analysis approaches that turned out to be equivalent. One approach used the distribution of the estimates, which resulted in the  $t$ -test. The other approach compared the sum of squared residuals from two models where one model was a special case of the other. This model comparison approach resulted in an  $F$ -test. In regression, there are the same two equivalent approaches: *distribution of estimates* and *model comparison*.

The model comparison is between the regression line and what the line would be if the slope were constrained to be zero. That is, you compare the fitted regression line with the horizontal line at the mean. This comparison is our null hypothesis (the slope = 0). If the regression line is a better fit than the line at the mean, then the slope of the regression line is significantly different from zero. This is often stated negatively: “If the regression line does not fit much better than the horizontal fit, then the slope of the regression line does not test as significantly different from zero.”

The  $F$ -test in the Analysis of Variance table is the comparison that tests the null hypothesis of the slope of the fitted line. It compares the sum of squared residuals from the regression fit to the sum of squared residuals from the sample mean.

**Figure 10.4** diagrams the relationship between the quantities in the statistical reports and corresponding plot. Here are descriptions of the quantities in the statistical tables:

**C Total**

corresponds to the sum of squares error if you had fit only the mean. You can verify this by looking at the Fit Mean table in the previous example. The C.

Total sum of squares (SSE in the Fit Mean report) is 1.0524 for both the mean fit and the line fit.

**Error**

is the sum of squared residuals after fitting the line, 0.1868. This is sometimes casually referred to as the residual, or residual error. You can think of Error as leftover variation—variation that wasn't explained by fitting a model.

**Model**

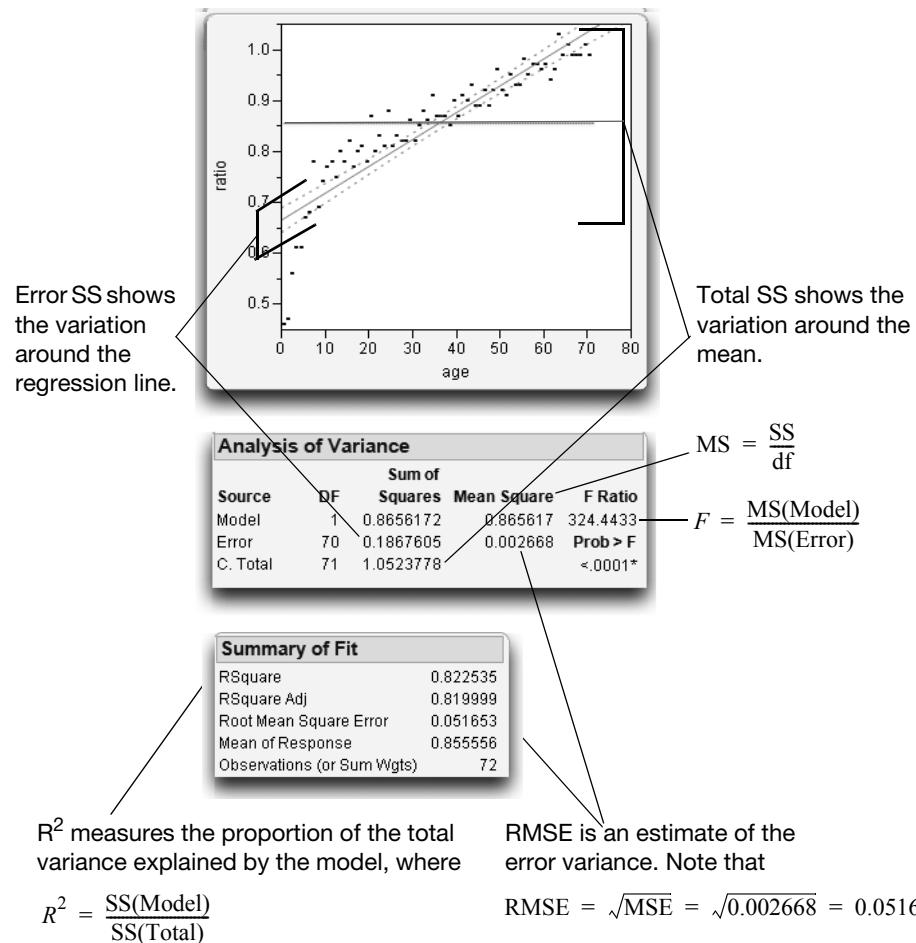
is the difference between the error sum of squares in the two models (the horizontal mean and the sloped regression line). It is the sum of squares resulting from the regression, 0.8656. You can think of Model as a measure of the variation in the data that was explained by fitting a regression line.

**Mean Square (MS)**

is a sum of squares divided by its respective degrees of freedom. The mean square for error (Error Mean Square) is the estimate of the error variance (0.002668 in this example).

**Root Mean Square Error (RSME)**

is found in the Summary of Fit report. It estimates the standard deviation of the error and is calculated as the square root of the Error Mean Square.

**Figure 10.4** Diagram to Compare Models

The  $F$ -statistic is calculated as the model mean square divided by the error mean square. If the model and error both have the same expected value, the  $F$ -statistic is then 1. However, if the model mean square is larger than the error mean square, you suspect that the slope is not zero and that the model is explaining some variation. The  $F$ -ratio has an  $F$ -distribution under the null hypothesis that (in this example) age has no effect on ratio.

If the true regression line has a slope of zero, then the model isn't explaining any of the variation. The model mean square and the error mean square would both estimate the residual error variance and therefore have the same expected value.

## The Distribution of the Parameter Estimates

The formula for a simple straight line has only two parameters, the intercept and the slope. For this example, the model can be written as follows:

$$\text{ratio} = b_0 + b_1 \text{ age} + \text{residual}$$

where  $b_0$  is the intercept and  $b_1$  is the slope. The Parameter Estimates Table also shows these quantities:

### Std Error

is the estimate of the standard deviation attributed to the parameter estimates.

Estimated coefficient in the equation      Estimate of standard deviation of the estimates

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.6656231	0.012176	54.67	<.0001*
age	0.0052759	0.000293	18.01	<.0001*

Ratio of estimates to standard error (Student's  $t$ -test)

$p$ -value for  $t$ -ratio

### t-Ratio

is a test that the true parameter is zero. The  $t$ -ratio is the ratio of the estimate to its standard error. Generally, you are looking for  $t$ -ratios that are greater than in absolute value, which usually corresponds to significance probabilities of less than 0.05.

### Prob>|t|

is the significance probability ( $p$ -value). You can translate this as "the probability of getting an even greater absolute  $t$ -value by chance alone if the true value of the slope is zero."

Note that the  $t$ -ratio for the age parameter, 18.01, is the square root of the  $F$ -ratio in the Analysis of Variance table, 324.44. You can double-click on the  $p$ -values in the tables to show more decimal places, and see that the  $p$ -values are exactly the same. This is not a surprise—the  $t$ -test for simple regression is testing the same null hypothesis as the  $F$ -test.

## Confidence Intervals on the Estimates

There are several ways to look at the significance of the estimates. The *t*-tests for the parameter estimates, discussed previously, test that the parameters are significantly different from zero. A more revealing way to look at the estimates is to obtain confidence limits that show the range of likely values for the true parameter values.

- ☞ Select the **Confid Curves Fit** command from the red triangle menu next to Linear Fit beneath the plot.

This command adds the confidence curves to the graph, as shown in **Figure 10.4**.

The 95% confidence interval is the smallest interval whose range includes the true parameter values with 95% confidence. The upper and lower confidence limits are calculated by adding and subtracting respectively the standard error of the parameter times a quantile value corresponding to a (0.05)/2 Student's *t*-test.

Another way to find the 95% confidence interval is to examine the Parameter Estimates tables. Although the 95% confidence interval values are initially hidden in the report, they can be made visible.

- ☞ Right-click anywhere on the Parameter Estimates report and select **Columns > Lower 95%** and **Columns > Upper 95%**, as shown in **Figure 10.5**.

**Figure 10.5** Add Confidence Intervals to Table

Right-click on the report itself to reveal 95% confidence intervals.

Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%
Intercept	0.61				3397	0.6899065
age	0.01				6917	0.0058601

Table Style

Columns

Sort by Column...

Make into Data Table

Make Combined Data Table

Make Into Matrix

Format Column...

Copy Column

Copy Table

Simulate

Bootstrap

✓ Term

✓ ~Bias

✓ Estimate

✓ Std Error

✓ t Ratio

✓ Prob>|t|

✓ Lower 95%

✓ Upper 95%

Std Beta

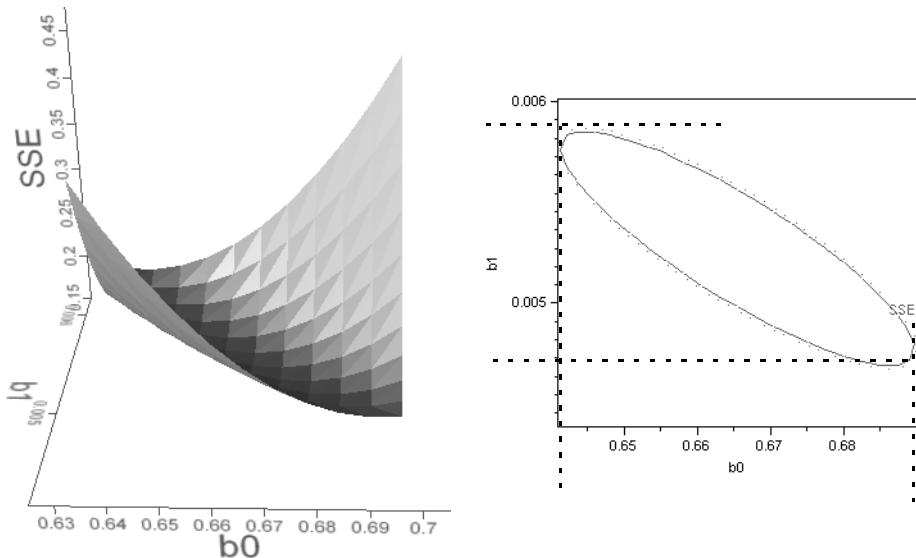
VIF

Design Std Error

An interesting way to see this concept is to look from the point of view of the sum of squared errors. Imagine the sum of squared errors (SSE) as a function of the parameter values, so that as you vary the slope and intercept values, you calculate the corresponding SSE. The least squares estimates are where this surface is at a minimum.

The left plot in **Figure 10.6** shows a three-dimensional view of this interpretation for the growth data regression problem. The 3-D view shows the curvature of the SSE surface as a function of the parameters, with a minimum in the center. The  $x$ - and  $y$ -axes are a grid of parameter values, and the  $z$ -axis is the computed SSE for those values.

**Figure 10.6** Representation of Confidence Limit Regions



One way to form a 95% confidence interval is to turn the  $F$ -test upside down. You take an  $F$  value that would be the criterion for a 0.05 test (3.97), multiply it by the MSE, and add that to the SSE. This gives a higher SSE of 0.19737 and forms a confidence region for the parameters. Anything that produces a smaller SSE is believable because it corresponds to an  $F$ -test with a  $p$ -value greater than 0.05.

The 95% confidence region is the inner elliptical shape in the plot on the right in **Figure 10.6**. The flatness of the ellipse corresponds to the amount of correlation of the estimates. You can look at the plot to see what parameter values correspond to the extremes of the ellipse in each direction.

- The horizontal scale corresponds to the intercept parameter. The confidence limits are the positions of the vertical tangents to the inner contour line, indicating a low point of 0.6413 and high point of 0.6899.
- The vertical scale corresponds to the slope parameter for age. The confidence limits are the positions of the vertical tangents to the inner contour line, indicating a low point of 0.00469 and a high point of 0.00586. These are the lower and upper 95% confidence limits for the parameters.

You can verify these numbers by looking at the confidence limits in **Figure 10.5**.

## Examine Residuals

It is always a good idea to take a close look at the residuals from a regression (the difference between the actual values and the predicted values):

- ☞ Select **Plot Residuals** from the red triangle menu next to Linear Fit beneath the scatterplot (**Figure 10.7**).

This command appends the set of diagnostic plots shown in **Figure 10.7** to the bottom of the regression report.

The picture you usually hope to see is the residuals scattered randomly about a mean of zero and clustered close to the normal quantile line. So, in residual plots like the ones shown in **Figure 10.7**, you are looking for patterns and for points that violate this random scatter. These plots are suspicious because the left side has a pattern of residuals below the reference lines. These points influence the slope of the regression line (**Figure 10.3**), pulling it down on the left.

You can see what the regression would look like without these points by excluding them from the analysis, as described in the next section.

## Exclusion of Rows

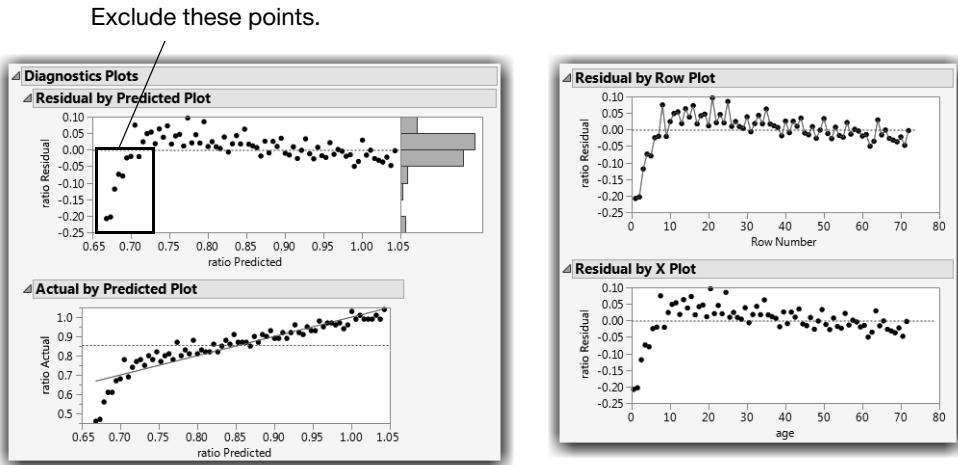
To exclude points (rows) from an analysis, you highlight the rows and assign them the **Exclude** row state characteristic as follows:

- ☞ Drag the cursor to highlight the points at the lower left of the residual plot (below the zero reference line as shown in **Figure 10.7**).
- ☞ Select **Rows > Exclude/Unexclude**.

You then see a *do not use* (🚫) sign in the row areas for the excluded rows (in the data table).

**Note:** You'll still see the excluded points on the scatterplot, since they've been excluded from the analysis but have not been hidden.

**Figure 10.7** Diagnostic Scatterplots to Look at Residuals



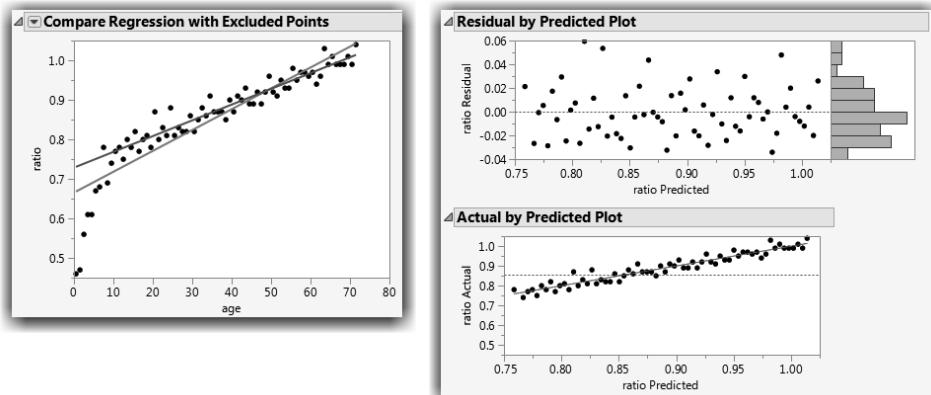
Now we fit a new regression line without these excluded points.

- ⓐ First, clean up the scatterplot by removing the horizontal line for the mean and the confidence curves for the linear fit. To do this, select **Remove Fit** from the red triangle menu next to Fit Mean (beneath the scatterplot). Deselect **Confidence Curves Fit** from the red triangle menu next to Linear Fit. Leave the original regression line on the plot so that you can compare it to a new regression line.
- ⓑ Again select **Fit Line** from the red triangle menu next to Bivariate Fit to overlay a regression line with the lower age points excluded from the analysis.

The plot in **Figure 10.8** shows the two regression lines. Note that the new line of fit seems to go through the bulk of the points better, ignoring the points at the lower left that are excluded.

- ⓒ To see the residuals plot for the new regression line, select **Plot Residuals** from the second Linear Fit red triangle menu.

The first two residual plots are shown in **Figure 10.8**. Notice that the residuals no longer have a pattern.

**Figure 10.8** Regression with Extreme Points Excluded

## Time to Clean Up

The next example uses this same scatterplot, so let's clean it up:

- ❖ Select **Rows > Clear Row States**.

This removes the Excluded row state status from all points so that you can use them in the next steps.

- ❖ To finish the cleanup, select **Remove Fit** from the second Linear Fit red triangle menu to remove the example regression that excluded outlying points.

## Polynomial Models

Rather than refitting a straight line after excluding some points, let's try fitting a different model altogether—a quadratic curve. This is a simple curve. It adds only one term to the linear model that we've been using, a term for the squared value of the regressor, age:

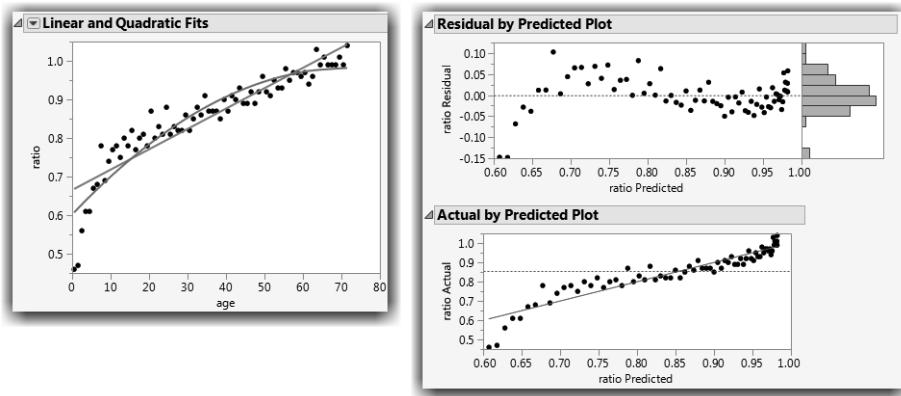
$$\text{ratio} = b_0 + b_1 \text{ age} + b_2 \text{ age}^2 + \text{residual}$$

To fit a quadratic curve to the ratio by age data:

- ❖ Select **Fit Polynomial > 2, quadratic** from the red triangle menu next to Bivariate Fit.

The left plot in **Figure 10.9** shows the linear and quadratic fits overlaid on the scatterplot. You can compare the straight line and curve, and also compare them statistically with the Analysis of Variance reports that show beneath the plot.

**Figure 10.9** Comparison of Linear and Second-Order Polynomial Fits



## Look at the Residuals

The plots on the right in **Figure 10.9** are the first two residual plots for the quadratic fit. To examine the residuals:

- ☞ Select **Plot Residuals** from the red triangle menu next to Polynomial Fit Degree=2.

The Actual by Predicted plot is a type of residual plot that shows the actual ratio values plotted against the predicted ratio values. This plot, on the right in **Figure 10.9**, enables us to see unexplained patterns in the data. There might be some improvement with the quadratic fit, but there still appears to be a pattern in the residuals. You might want to continue to fit a model with higher-order terms.

## Higher-Order Polynomials

To give more flexibility to the curve, specify higher-order polynomials, adding a term to the third power, to the fourth power, and so on.

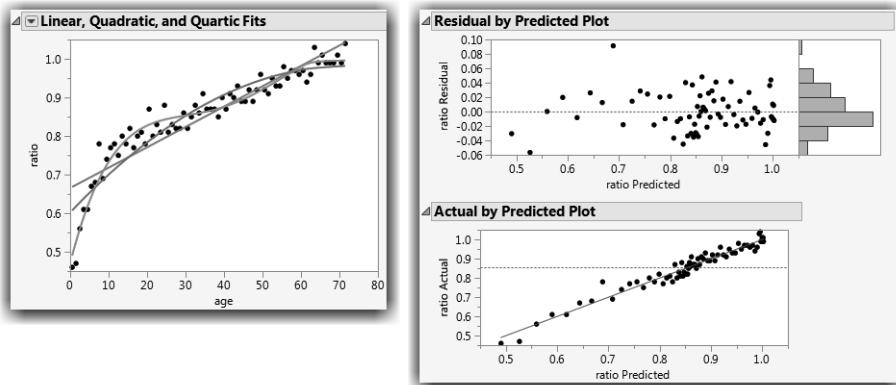
- ☞ With the scatterplot active, request a polynomial of degree 4 from the red triangle menu next to Bivariate Fit.

This plots the curve with linear, quadratic, cubic, and quartic terms, not simply a fourth-power term. Plotting polynomials always includes lower-order terms.

- Then, select **Plot Residuals** from the red triangle menu next to Polynomial Fit Degree=4.

Note that the residuals in **Figure 10.10** no longer appear to have a nonlinear pattern to them. (The points are randomly scattered around the line.)

**Figure 10.10** Comparison of Linear and Fourth-Order Polynomial Fits



## Distribution of Residuals

It is also informative to look at the shape of the distribution of the residuals. If the distribution departs dramatically from the normal, then you might be able to find further phenomena in the data.

Histograms of residuals are automatically provided when residual plots are requested for a model. These histograms are shown next to the residual by predicted plots in **Figure 10.9** and **Figure 10.10**. **Figure 10.11** shows histograms of residuals from the linear fit, the quadratic fit, and quartic (fourth degree) fit side-by-side for comparison.

To generate these histograms:

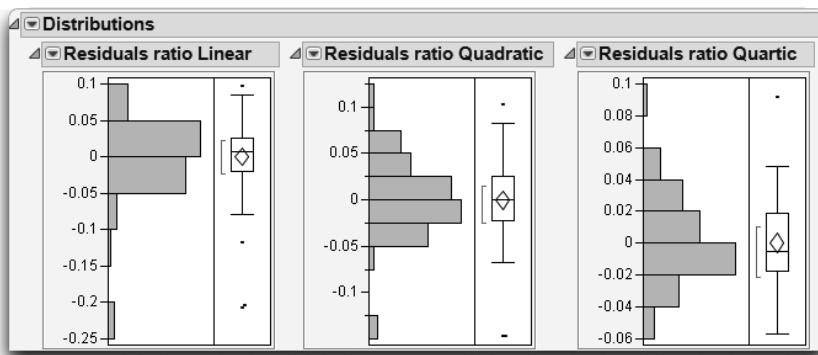
- Select the **Save Residuals** command from the red triangle menu beneath the scatterplot for each regression fit.

This forms three new columns in the data table.

- Select **Analyze > Distribution**, assign each of the three new columns of residual values to **Y, Columns**, and then click **OK**.

You can see in **Figure 10.11** that the distributions evolve toward normality – higher-order models explain more variation, so the residuals become more normal.

**Figure 10.11** Histograms for Distribution of Residuals



## Transformed Fits

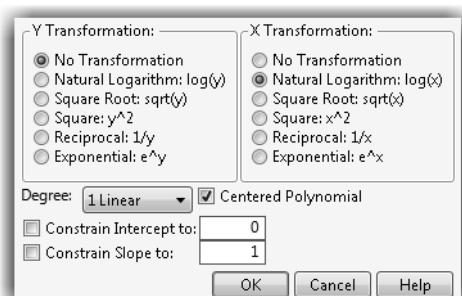
Sometimes, you can find a better fit if you transform either the Y or X variable (or sometimes both). When we say transform, we mean that we apply a mathematical function to a variable and examine these new values. For example, rather than looking at  $x$ , we might examine  $\log x$ . One way of doing this is to create a new column in the data table by right-clicking the column header for age, selecting **New Formula Column**, and then selecting **Transform > Log**. Then, select **Analyze > Fit Y by X** to do a straight-line regression of ratio on the log of age. Results from this method are shown on the right in **Figure 10.12**.

Alternatively, you can use the Fit Y by X platform to do this directly:

- ❖ Select **Analyze > Fit Y by X**, assign ratio to **Y, Response** and age to **X, Factor**, and then click **OK**.
- ❖ Select **Fit Special** from the red triangle menu next to Bivariate Fit.

The **Fit Special** command displays a window that lists natural log, square, square root, exponential, and other transformations as shown here, as selections for both the X and Y variables.

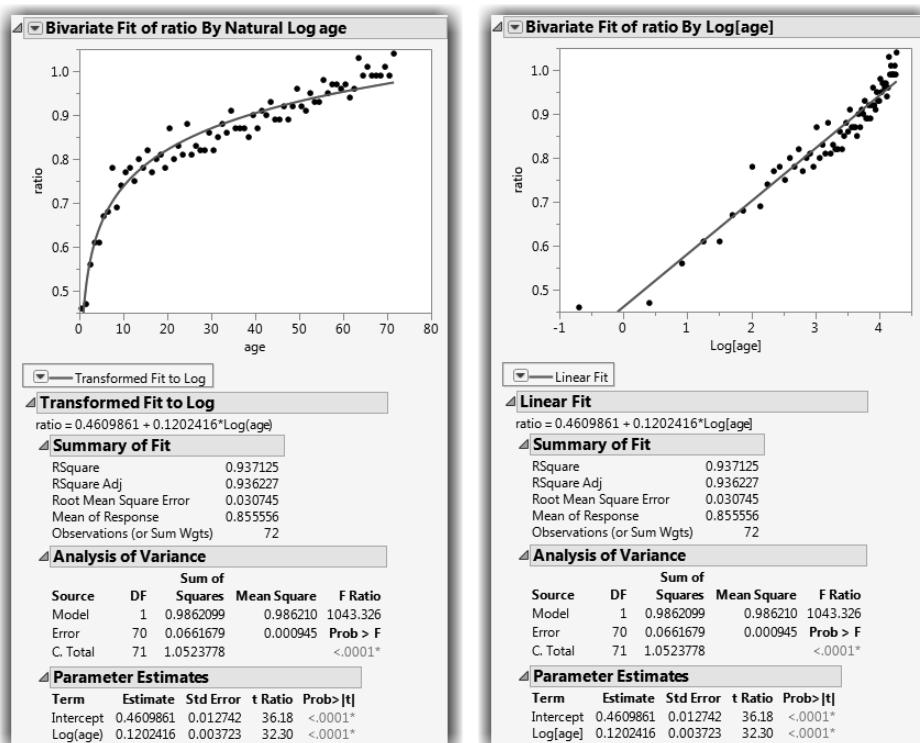
Try fitting ratio to the log of age:



- Ⓐ Click the **Natural Logarithm: (log**
- ✗) radio button for X. When you click **OK**, you see the left plot in **Figure 10.12**.

The regression equation and statistical results are identical to those obtained by fitting a model with the log of age (right, in **Figure 10.12**).

**Figure 10.12** Comparison of Fits



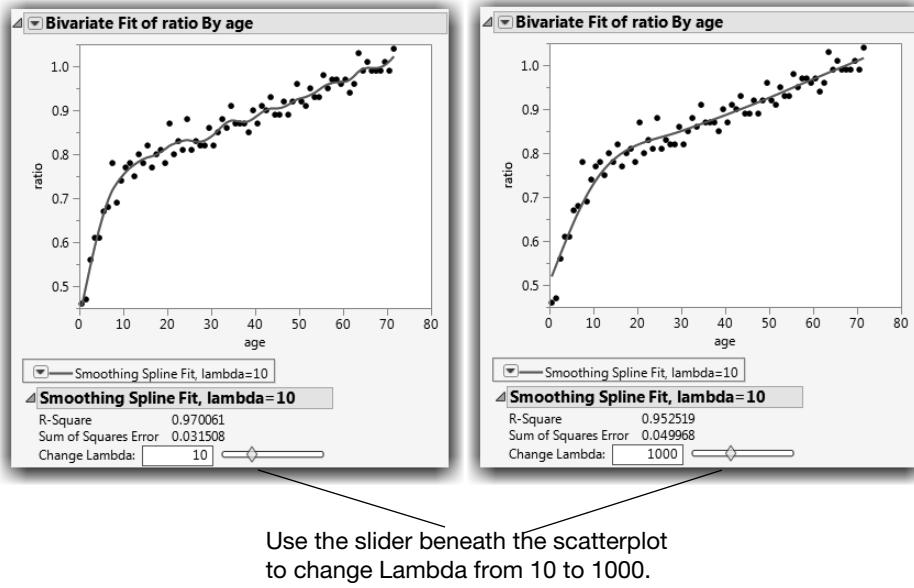
**Note:** If you transform the Y variable, you can't compare the  $R^2$  and error sums of squares of the transformed variable fit with the untransformed variable fit. You are fitting a different variable.

## Spline Fit

It would be nice if you could fit a flexible curve, like a leaf spring, through the points. The leaf spring would resist bending somewhat, but would take gentle bends when it's needed to fit the data better. A smoothing spline is exactly that type of fit. With smoothing splines, you specify how stiff to make the curve. If the spline is too rigid, it looks like a straight line; but if it is too flexible, it curves to try to fit each point. Use these commands to see the spline plots in **Figure 10.13**:

- ☞ Select **Analyze > Fit Y by X**, assign ratio to **Y, Response** and age to **X, Factor**, and then click **OK**.
- ☞ Select **Flexible > Fit Spline > 10** from the menu. This fits a spline with a lambda (the smoothing parameter) of 10.
- ☞ Change lambda from 10 to 1000 using either the slider beneath the scatterplot or the field provided.

Play with the slider and watch the curviness of the spline change from flexible to rigid as you increase the value of lambda.

**Figure 10.13** Comparison of Less Flexible and More Flexible Spline Fits

## Are Graphics Important?

Some statistical packages don't show graphs of the regression. Others require you to make an extra effort to see a regression graph. The following example shows the types of phenomena that you miss if you don't examine a graph.

- ❖ Select **Help > Sample Data Library** and open Anscombe.jmp (Anscombe, 1973).
- ❖ Click the green triangle next to **The Quartet** (in the Table panel on the top left) to run the saved script.

In essence, this stored script is a shortcut for the following actions. By using the script, you don't have to do the following:

- Select **Analyze > Fit Y by X** four times. Each time, fit Y1 by X1, Y2 by X2, Y3 by X3, and Y4 by X4.
- For each pair, select the **Fit Line** command from the red triangle menu next to Bivariate Fit above each scatterplot.

First, look at the text reports for each bivariate analysis, shown in **Figure 10.14**, and compare them. Notice that the reports are nearly identical. The  $R^2$  values, the  $F$ -tests, the parameter estimates and standard errors are all the same. Does this mean that the situations are the same?

Think of it another way. Suppose you are a doctor, and the four data sets represent patients. Their text reports represent the symptoms that they present. Since they are identical, would you give them all the same diagnosis?

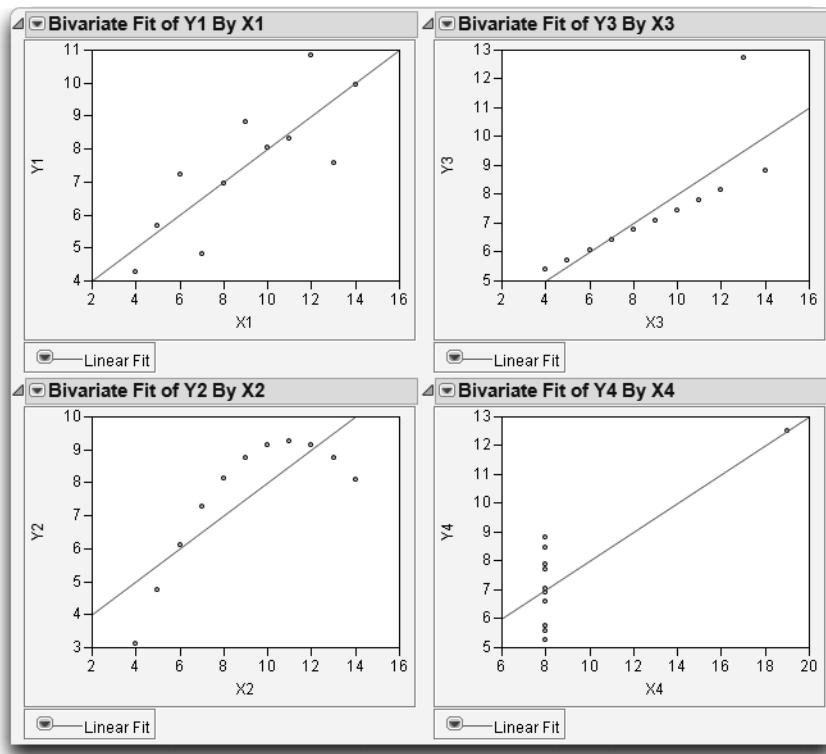
**Figure 10.14** Statistical Reports for Four Analyses

<b>Linear Fit</b> $Y_1 = 3.0000909 + 0.5000909 * X_1$ <b>Summary of Fit</b> RSquare 0.666542 RSquare Adj 0.629492 Root Mean Square Error 1.236603 Mean of Response 7.5000909 Observations (or Sum Wgts) 11	<b>Linear Fit</b> $Y_3 = 3.0024545 + 0.4997273 * X_3$ <b>Summary of Fit</b> RSquare 0.666324 RSquare Adj 0.629249 Root Mean Square Error 1.236311 Mean of Response 7.5 Observations (or Sum Wgts) 11																																																		
<b>Analysis of Variance</b> <table border="1"> <thead> <tr> <th colspan="5">Sum of</th> </tr> <tr> <th>Source</th> <th>DF</th> <th>Squares</th> <th>Mean Square</th> <th>F Ratio</th> </tr> </thead> <tbody> <tr> <td>Model</td> <td>1</td> <td>27.510001</td> <td>27.5100</td> <td>17.9899</td> </tr> <tr> <td>Error</td> <td>9</td> <td>13.762690</td> <td>1.5292</td> <td>Prob &gt; F</td> </tr> <tr> <td>C. Total</td> <td>10</td> <td>41.272691</td> <td></td> <td>0.0022*</td> </tr> </tbody> </table>	Sum of					Source	DF	Squares	Mean Square	F Ratio	Model	1	27.510001	27.5100	17.9899	Error	9	13.762690	1.5292	Prob > F	C. Total	10	41.272691		0.0022*	<b>Analysis of Variance</b> <table border="1"> <thead> <tr> <th colspan="5">Sum of</th> </tr> <tr> <th>Source</th> <th>DF</th> <th>Squares</th> <th>Mean Square</th> <th>F Ratio</th> </tr> </thead> <tbody> <tr> <td>Model</td> <td>1</td> <td>27.470008</td> <td>27.4700</td> <td>17.9723</td> </tr> <tr> <td>Error</td> <td>9</td> <td>13.756192</td> <td>1.5285</td> <td>Prob &gt; F</td> </tr> <tr> <td>C. Total</td> <td>10</td> <td>41.226200</td> <td></td> <td>0.0022*</td> </tr> </tbody> </table>	Sum of					Source	DF	Squares	Mean Square	F Ratio	Model	1	27.470008	27.4700	17.9723	Error	9	13.756192	1.5285	Prob > F	C. Total	10	41.226200		0.0022*
Sum of																																																			
Source	DF	Squares	Mean Square	F Ratio																																															
Model	1	27.510001	27.5100	17.9899																																															
Error	9	13.762690	1.5292	Prob > F																																															
C. Total	10	41.272691		0.0022*																																															
Sum of																																																			
Source	DF	Squares	Mean Square	F Ratio																																															
Model	1	27.470008	27.4700	17.9723																																															
Error	9	13.756192	1.5285	Prob > F																																															
C. Total	10	41.226200		0.0022*																																															
<b>Parameter Estimates</b> <table border="1"> <thead> <tr> <th>Term</th> <th>Estimate</th> <th>Std Error</th> <th>t Ratio</th> <th>Prob&gt; t </th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td>3.0000909</td> <td>1.124747</td> <td>2.67</td> <td>0.0257*</td> </tr> <tr> <td>X1</td> <td>0.5000909</td> <td>0.117906</td> <td>4.24</td> <td>0.0022*</td> </tr> </tbody> </table>	Term	Estimate	Std Error	t Ratio	Prob> t	Intercept	3.0000909	1.124747	2.67	0.0257*	X1	0.5000909	0.117906	4.24	0.0022*	<b>Parameter Estimates</b> <table border="1"> <thead> <tr> <th>Term</th> <th>Estimate</th> <th>Std Error</th> <th>t Ratio</th> <th>Prob&gt; t </th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td>3.0024545</td> <td>1.124481</td> <td>2.67</td> <td>0.0256*</td> </tr> <tr> <td>X3</td> <td>0.4997273</td> <td>0.117878</td> <td>4.24</td> <td>0.0022*</td> </tr> </tbody> </table>	Term	Estimate	Std Error	t Ratio	Prob> t	Intercept	3.0024545	1.124481	2.67	0.0256*	X3	0.4997273	0.117878	4.24	0.0022*																				
Term	Estimate	Std Error	t Ratio	Prob> t																																															
Intercept	3.0000909	1.124747	2.67	0.0257*																																															
X1	0.5000909	0.117906	4.24	0.0022*																																															
Term	Estimate	Std Error	t Ratio	Prob> t																																															
Intercept	3.0024545	1.124481	2.67	0.0256*																																															
X3	0.4997273	0.117878	4.24	0.0022*																																															
<b>Linear Fit</b> $Y_2 = 3.0000909 + 0.5 * X_2$ <b>Summary of Fit</b> RSquare 0.666242 RSquare Adj 0.629158 Root Mean Square Error 1.237214 Mean of Response 7.5000909 Observations (or Sum Wgts) 11	<b>Linear Fit</b> $Y_4 = 3.0017273 + 0.4999091 * X_4$ <b>Summary of Fit</b> RSquare 0.666707 RSquare Adj 0.629675 Root Mean Square Error 1.235695 Mean of Response 7.5000909 Observations (or Sum Wgts) 11																																																		
<b>Analysis of Variance</b> <table border="1"> <thead> <tr> <th colspan="5">Sum of</th> </tr> <tr> <th>Source</th> <th>DF</th> <th>Squares</th> <th>Mean Square</th> <th>F Ratio</th> </tr> </thead> <tbody> <tr> <td>Model</td> <td>1</td> <td>27.500000</td> <td>27.5000</td> <td>17.9856</td> </tr> <tr> <td>Error</td> <td>9</td> <td>13.776291</td> <td>1.5307</td> <td>Prob &gt; F</td> </tr> <tr> <td>C. Total</td> <td>10</td> <td>41.276291</td> <td></td> <td>0.0022*</td> </tr> </tbody> </table>	Sum of					Source	DF	Squares	Mean Square	F Ratio	Model	1	27.500000	27.5000	17.9856	Error	9	13.776291	1.5307	Prob > F	C. Total	10	41.276291		0.0022*	<b>Analysis of Variance</b> <table border="1"> <thead> <tr> <th colspan="5">Sum of</th> </tr> <tr> <th>Source</th> <th>DF</th> <th>Squares</th> <th>Mean Square</th> <th>F Ratio</th> </tr> </thead> <tbody> <tr> <td>Model</td> <td>1</td> <td>27.490001</td> <td>27.4900</td> <td>18.0033</td> </tr> <tr> <td>Error</td> <td>9</td> <td>13.742490</td> <td>1.5269</td> <td>Prob &gt; F</td> </tr> <tr> <td>C. Total</td> <td>10</td> <td>41.232491</td> <td></td> <td>0.0022*</td> </tr> </tbody> </table>	Sum of					Source	DF	Squares	Mean Square	F Ratio	Model	1	27.490001	27.4900	18.0033	Error	9	13.742490	1.5269	Prob > F	C. Total	10	41.232491		0.0022*
Sum of																																																			
Source	DF	Squares	Mean Square	F Ratio																																															
Model	1	27.500000	27.5000	17.9856																																															
Error	9	13.776291	1.5307	Prob > F																																															
C. Total	10	41.276291		0.0022*																																															
Sum of																																																			
Source	DF	Squares	Mean Square	F Ratio																																															
Model	1	27.490001	27.4900	18.0033																																															
Error	9	13.742490	1.5269	Prob > F																																															
C. Total	10	41.232491		0.0022*																																															
<b>Parameter Estimates</b> <table border="1"> <thead> <tr> <th>Term</th> <th>Estimate</th> <th>Std Error</th> <th>t Ratio</th> <th>Prob&gt; t </th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td>3.00009091</td> <td>1.125302</td> <td>2.67</td> <td>0.0258*</td> </tr> <tr> <td>X2</td> <td>0.5</td> <td>0.117964</td> <td>4.24</td> <td>0.0022*</td> </tr> </tbody> </table>	Term	Estimate	Std Error	t Ratio	Prob> t	Intercept	3.00009091	1.125302	2.67	0.0258*	X2	0.5	0.117964	4.24	0.0022*	<b>Parameter Estimates</b> <table border="1"> <thead> <tr> <th>Term</th> <th>Estimate</th> <th>Std Error</th> <th>t Ratio</th> <th>Prob&gt; t </th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td>3.0017273</td> <td>1.123921</td> <td>2.67</td> <td>0.0256*</td> </tr> <tr> <td>X4</td> <td>0.4999091</td> <td>0.117819</td> <td>4.24</td> <td>0.0022*</td> </tr> </tbody> </table>	Term	Estimate	Std Error	t Ratio	Prob> t	Intercept	3.0017273	1.123921	2.67	0.0256*	X4	0.4999091	0.117819	4.24	0.0022*																				
Term	Estimate	Std Error	t Ratio	Prob> t																																															
Intercept	3.00009091	1.125302	2.67	0.0258*																																															
X2	0.5	0.117964	4.24	0.0022*																																															
Term	Estimate	Std Error	t Ratio	Prob> t																																															
Intercept	3.0017273	1.123921	2.67	0.0256*																																															
X4	0.4999091	0.117819	4.24	0.0022*																																															

Now look at the scatterplots shown in **Figure 10.15**, and note the following characteristics:

- $Y_1$  by  $X_1$  shows a typical regression situation. In this case, as  $X_1$  increases  $Y_1$  increases.
- The points in  $Y_2$  by  $X_2$  follow a parabola, so a quadratic model is appropriate, with the square of  $X_2$  as an additional term in the model. As an exercise, fit this quadratic model.
- There is an extreme outlier in  $Y_3$  by  $X_3$ , which increases the slope of the line that would otherwise be a perfect fit. As an exercise, exclude the outlying point and fit another line.
- In  $Y_4$  by  $X_4$ , all the  $x$ -values are the same except for one point, which completely determines the slope of the line. This situation is called *leverage*. It is not necessarily bad, but you ought to know about it.

**Figure 10.15** Regression Lines for Four Analyses



## Why It's Called Regression

Remember the story about the study done by Sir Francis Galton mentioned at the beginning of this chapter? He examined the heights of parents and their grown children to learn how much of height is an inherited characteristic. He concluded that the children's heights tended to be more moderate than the parent's heights, and used the term "regression to the mean" to name this phenomenon. For example, if a parent was tall, the children would be tall, but less tall than the parents. If a parent was short, the child would tend to be short, but less short than the parent.

Galton's case is interesting for two reasons. It was the first use of regression, and Galton failed to notice some properties of regression that would have changed his mind about using regression to draw his conclusions. To investigate Galton's data, follow these steps:

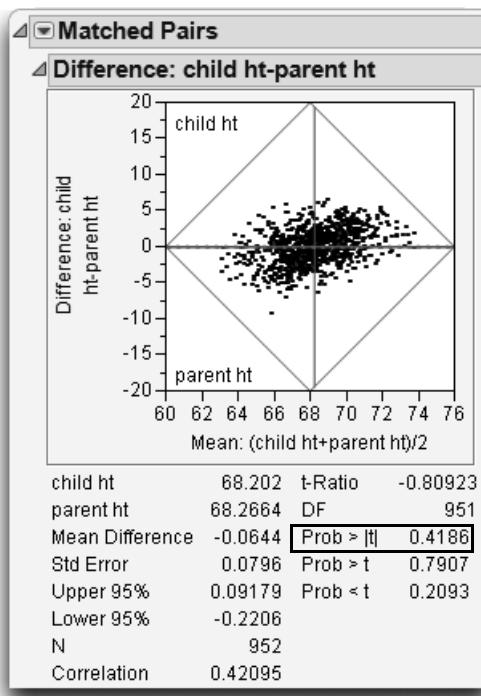
- ☞ Select **Help > Sample Data Library** and open Galton.jmp.
- ☞ Select **Analyze > Specialized Modeling > Matched Pairs** and assign child ht and parent ht to **Y, Paired Responses**.
- ☞ Click **OK**.
- ☞ Select **Reference Frame** from the red triangle menu next to Matched Pairs.
- ☞ Double-click the *y* axis and change the limits to +20 and -20.

The data in the Galton table comes from Galton's published table, but each point is *jittered* by a random amount up to 0.5 in either direction. The jittering is done so that all the points show in the plot instead of overlapping. Also, Galton multiplied the women's heights by 1.08 to make them comparable to men's. The parent ht variable is defined as the average of the two parents' heights.

The scatterplot produced by the Matched Pairs platform is the same that is given by the Fit Y by X platform, but it is rotated by 45°. (See "The Matched Pairs Platform for a Paired t-Test" on page 200 for details about this plot). If the difference between the two variables is zero (our null hypothesis is that the parent and child's heights are the same), the points cluster around a horizontal reference line at zero. The mean difference is shown as a horizontal line, with the 95% confidence interval above and below. If the confidence region includes the horizontal reference line at zero, then the means are not significantly different at

the 0.05 level and we can't reject the null hypothesis. This represents the *t*-test that Galton could have hypothesized to see whether the mean height of the child is the same as the parent.

**Figure 10.16** Matched Pairs Output of Galton's Data



For this test, we do not reject the null hypothesis.

However, this is not the test that Galton considered. He invented regression to fit an arbitrary line on a plot of parent's height by child's height, and then tested to see whether the slope of the line was 1. If the line has a slope of 1, then the predicted height of the child is the same as that of the parent, except for a generational constant. A slope of less than 1 indicates that the children tended to have more moderate heights (closer to the mean) than the parents. To look at this regression:

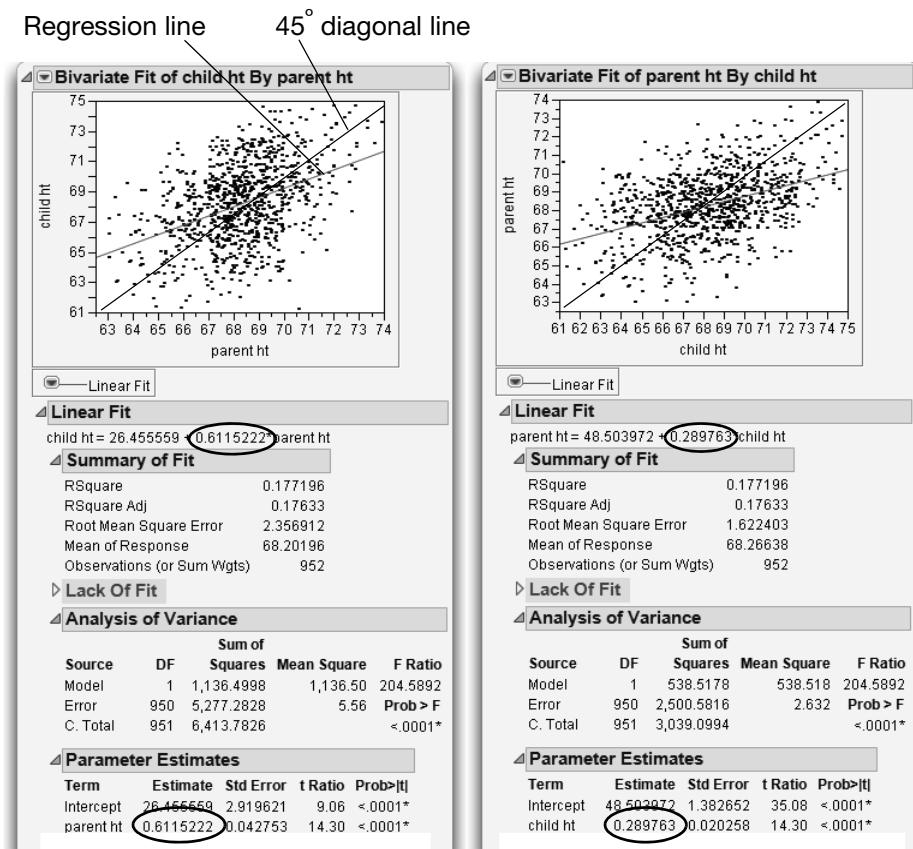
- ⓐ Select **Analyze > Fit Y by X**, with parent ht as **X, Factor** and child ht as **Y, Response**, and then click **OK**.
- ⓑ Select **Fit Line** from the red triangle menu next to Bivariate Fit.

- Fit a line with a slope of 1. To do this, select **Fit Special** from the red triangle menu next to Bivariate Fit, select the **Constrain Intercept to** and **Constrain Slope to** check boxes, and then click **OK**.

When you examine the Parameter Estimates table and the regression line in the left plot of **Figure 10.17**, you see that the least squares regression slope is 0.61. This value is far below 1. This suggests the regression is toward the mean.

## What Happens When X and Y Are Switched?

Is Galton's technique fair to the hypothesis? Think of it in reverse: If the children's heights were more moderate than the parents, shouldn't the parent's heights be more extreme than the children's? To find out, you can reverse the model and try to predict the parent's heights from the children's heights (for example, switch the roles of  $x$  and  $y$ ). The analysis on the right in **Figure 10.17** shows the results when the parent's height is  $y$  and children's height is  $x$ . Because the previous slope was less than 1, you'd think that this analysis would give a slope greater than 1. Surprisingly, the reported slope is 0.28, even less than the first slope!

**Figure 10.17** Child's Height and Parent's Height

**Note:** To fit the diagonal lines, select **Fit Special** from the red triangle menu next to Bivariate Fit. Constrain both the intercept and the slope to 1.

Instead of phrasing the conclusion that children tended to regress to the mean, Galton could have worded his conclusion to say that there is a somewhat weak relationship between the two heights. With regression, there is no symmetry between the Y and X variables. The slope of the regression of Y on X is not the reciprocal of the slope of the regression of X on Y. You cannot calculate the X by Y slope by taking the Y by X equation and solving for the other variable.

Regression is not symmetric because the error that is minimized is only in one direction—that of the Y variable. So if the roles are switched, a completely different problem is solved.

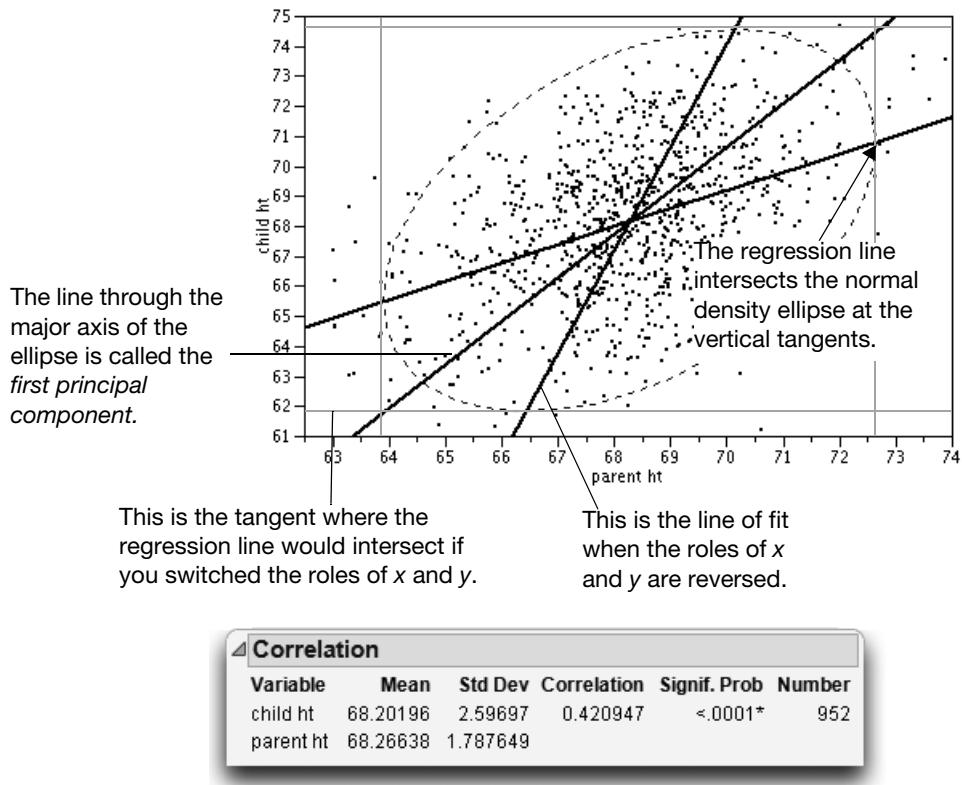
The slope is always smaller than the reciprocal of the inverted variables. However, there is a way to fit the slope symmetrically, so that the role of both variables is the same. This is what you do when you calculate a *correlation*.

Correlation characterizes the bivariate normal continuous density. The contours of the normal density form ellipses like the example illustrated in **Figure 10.18**. If there is a strong relationship, the ellipse becomes elongated along a diagonal axis. The line along this axis even has a name—it's called the *first principal component*.

It turns out that the least squares line is not the same as the first principal component. Instead, the least squares line bisects the contour ellipse at the vertical tangent points (see **Figure 10.18**).

If you reverse the direction of the tangents, you describe what the regression line would be if you reversed the role of the Y and X variables. If you draw the X by Y line fit in the Y by X diagram as shown in **Figure 10.18**, it intersects the ellipse at its horizontal tangents.

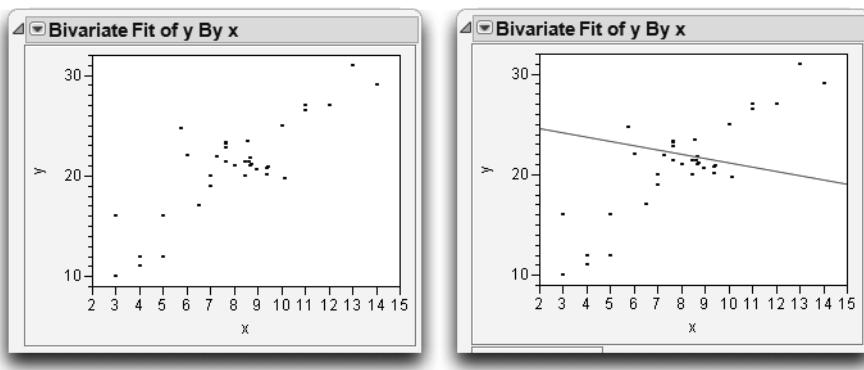
In both cases, the slope of the fitted line was less than 1. Therefore, Galton's phenomenon of regression to the mean was more an artifact of the method, rather than something learned from the data.

**Figure 10.18** Diagram Comparing Regression and Correlation

## Curiosities

### Sometimes It's the Picture That Fools You

An experiment by a molecular biologist generated some graphs similar to the scatterplots in **Figure 10.19**. Looking quickly at the plot on the left, where would you guess the least squares regression line lies? Now look at the graph on the right to see where the least squares fit really lies.

**Figure 10.19** Beware of Hidden Dense Clusters

The biologist was perplexed. How did this unlikely looking regression line happen?

It turns out that there is a very dense cluster of points you can't see. This dense cluster of hundreds of points dominated the slope estimate even though the few points farther out had more individual leverage. There was nothing wrong with the computer, the method, or the computation. It's just that the human eye is sometimes fooled, especially when many points occupy the same position. (This example uses the Slope.jmp sample data table.)

## High-Order Polynomial Pitfall

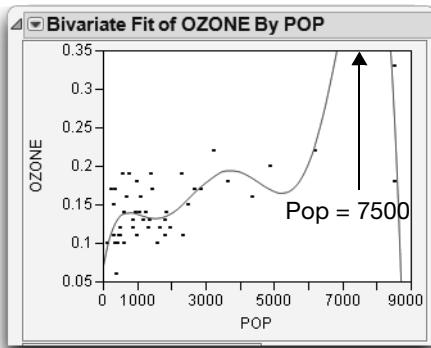
Suppose you want to develop a prediction equation for predicting ozone based on the population of a city. The lower-order polynomials fit fine, but why not take the “more is better” approach and try a higher order one, for example, sixth degree.

- ☞ To see the next example, select **Help > Sample Data Library** and open Polycity.jmp.
- ☞ Select **Fit Y by X** with OZONE as Y and POP as X, and then click **OK**.
- ☞ Select **Fit Polynomial > 6** from the red triangle menu next to Bivariate Fit.

As you can see in the bivariate fit shown here, the curve fits extremely well—too well, in fact. How trustworthy is the ozone prediction for a city with a population of 7500?

This overfitting phenomenon, shown here, occurs in higher-order polynomials when the data are unequally spaced.

More is not always better.



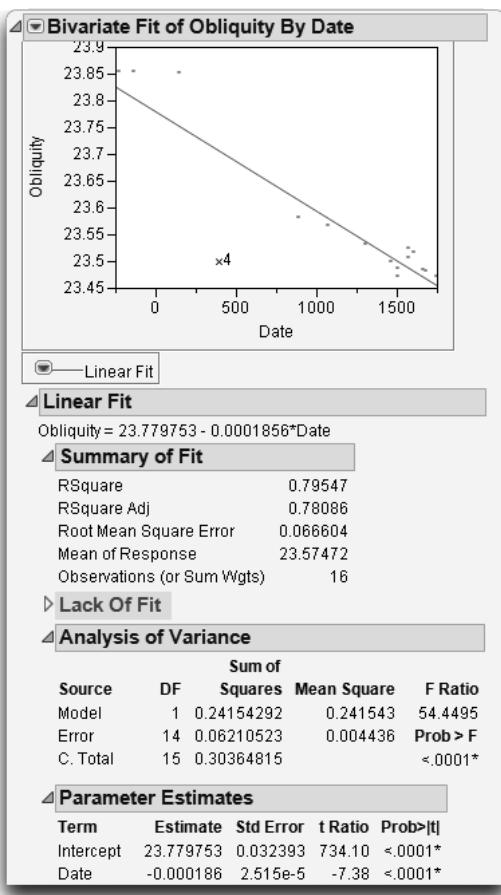
## The Pappus Mystery on the Obliquity of the Ecliptic

Ancient measurements of the angle of the earth's rotation disagree with modern measurements. Is this because modern ones are based on different (and better) technology, or did the angle of rotation actually change?

"Case Study: The Earth's Ecliptic" on page 150 introduced the angle-of-the-ecliptic data. The data that go back to ancient times are in the Cassini.jmp sample data table (Stigler 1986). **Figure 10.20** shows the regression of the obliquity (angle) by time. The regression suggests that the angle has changed over time. The mystery is that the measurement by Pappus is not consistent with the rest of the line. Was Pappus's measurement flawed or did something else happen at that time? We probably will never know.

These types of mysteries sometimes lead to detective work that results in great discoveries. Marie Curie discovered radium because of a small discrepancy in measurements made with pitchblende experiments. If she hadn't noticed this discrepancy, the progress of physics might have been delayed.

Outliers are not to be dismissed casually. Moore and McCabe (1989) point out a situation in the 1980s where a satellite was measuring ozone levels over the poles. It automatically rejected a number of measurements because they were very low. Because of this, the ozone holes were not discovered until years later by experiments run from the earth's surface that confirmed the satellite measurements.

**Figure 10.20** Measurements of the Earth's Angular Rotation

## Exercises

1. This exercise deals with the data on the top-grossing box office movies (as of June 2003) found in the sample data table `movies.jmp`. Executives are interested in predicting the amount that movies gross overseas based on the domestic gross.
  - (a) Examine a scatterplot of Worldwide \$ versus Domestic \$. Do you notice any outliers? Identify them.
  - (b) Fit a line through the mean of the response using **Fit Mean**.

- (c) Perform a linear regression on Worldwide \$ versus Domestic \$. Does this linear model describe the data better than the constrained model with only the mean? Justify your answer.
  - (d) Exclude any outliers that you found in part (a) and re-run the regression. Describe the differences between this model and the model that included all the points. Which would you trust more?
  - (e) Use the Graph Builder (under the **Graph** menu) to further explore the relationship between these two variables. Drag Worldwide \$ into the **Y** zone and Domestic \$ into the **X** zone. Explore the graph icons at the top of the window. Fit a regression line (with confidence curves) using the appropriate icon.
  - (f) In the graph produced in part (e), drag Type into the **Group X**, **Group Y**, or **Overlay** zone. Is there a difference in gross dollars for the different types of movies?
  - (g) Finally, select the red triangle next to **Graph Builder** and then select **Local Data Filter**. In the Local Data Filter, select **Type**, click the **Add** button, and then select both Drama and Comedy to display results for only these two movie types. Comment on what you see, and whether you think a single prediction equation for all movie types will be useful to the executives.
2. How accurate are past elections in predicting future ones? To answer this question, open the sample data table Presidential Elections.jmp. (See Ladd and Carle.) This file contains the percent of votes cast for the Democratic nominee in three previous elections.
    - (a) Produce histograms for the percent of votes in each election, with the three axes having uniform scaling. (The **Uniform Scaling** option is in the red triangle menu next to Distribution.) What do you notice about the three means? If you find a difference, explain why it might be there.
    - (b) Fit regression lines and find the correlations for 1996 versus 1980 and 1984 versus 1980. Comment on the associations that you see.
    - (c) Would the lines generated in these analyses have been useful in predicting the percent votes cast for the Democratic nominee in the next presidential election (2000 and 1988 respectively)? Justify your answer.
  3. Open the sample data table Birth Death.jmp to see data on the birth and death rates of several countries around the world.

- (a) Identify any univariate outliers in the variables birth and death.
  - (b) Fit a line through the mean of the response and a regression line to the data with birth as X and death as Y. Is the linear fit significantly better than the constrained fit using just the mean?
  - (c) Produce residual plots for the linear fit in part (b).
  - (d) The slope of the regression line seems to be highly significant. Why, then, is this model inappropriate for this situation?
  - (e) Use the techniques from this chapter to fit several transformed, polynomial, or spline models to the data and comment on the best model.
4. The sample data table Solubility.jmp (Koehler and Dunn, 1988) contains data from a chemical experiment that tested the solubility characteristics of seven chemicals.
- (a) Produce scatterplots of all the solutions versus ether. Based on the plots, which chemical has the highest correlation with ether?
  - (b) Carbon Tetrachloride has solubility characteristics that are highly correlated with Hexane. Find a 95% confidence interval for the slope of the regression line of Carbon Tetrachloride versus Hexane.
  - (c) Suppose the roles of the variables in part (b) were reversed. What can you say about the slope of the new regression line?





# 11

## Categorical Distributions

### Overview

When a response is categorical, a different set of tools is needed to analyze the data. This chapter focuses on simple categorical responses and introduces these topics:

- There are two ways to approach categorical data. This chapter refers to them as *choosing* and *counting*. They use different tools and conventions for analysis.
- The concept of variability in categorical responses is more difficult than in continuous responses. Monte Carlo simulation helps demonstrate how categorical variability works.
- The *chi-square test* is the fundamental statistical tool for categorical models. There are two types of chi-square tests. They test the same thing in a different way and get similar results.

Fitting models to categorical response data is covered in Chapter 12, “Categorical Models.”

## Chapter Contents

Overview .....	281
Categorical Situations .....	283
Categorical Responses and Count Data: Two Outlooks.....	283
A Simulated Categorical Response .....	286
Simulating Some Categorical Response Data.....	287
Variability in the Estimates .....	288
Larger Sample Sizes .....	290
Monte Carlo Simulations for the Estimators.....	291
Distribution of the Estimates.....	292
The $\chi^2$ Pearson Chi-Square Test Statistic .....	293
The $G^2$ Likelihood-Ratio Chi-Square Test Statistic .....	294
Likelihood Ratio Tests .....	295
The $G^2$ Likelihood Ratio Chi-Square Test .....	296
Univariate Categorical Chi-Square Tests .....	296
Comparing Univariate Distributions .....	297
Charting to Compare Results .....	299
Exercises.....	301

## Categorical Situations

A *categorical response* is one in which the response is from a limited number of choices (called *response categories*). There is a probability associated with each of these choices, and these probabilities sum to 1.

Categorical responses are common:

- Consumer preferences are usually categorical: Which do you like the best—tea, coffee, juice, or soft drinks?
- Medical outcomes are often categorical: Did the patient live or die?
- Biological responses are often categorical: Did the seed germinate?
- Mechanical responses can be categorical: Did the fuse blow at 20 amps?
- Any continuous response can be converted to a categorical response: Did the temperature reach 100 degrees (“yes” or “no”)?

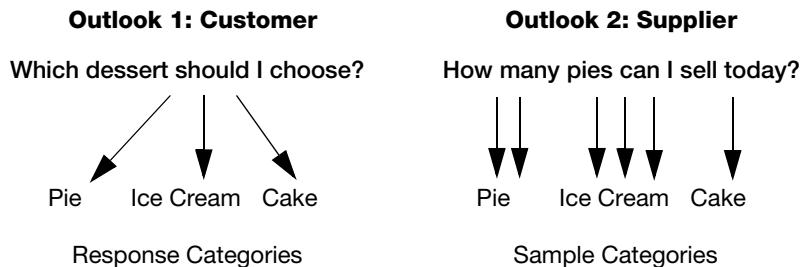
## Categorical Responses and Count Data: Two Outlooks

It is important to understand that there are two approaches to handling categorical responses. The two approaches generally give the same results, but they use different tools and terms.

First, suppose that each observation in a data set represents the response of a chooser. Based on conditions of the observation, the chooser is going to respond with one of the response categories. For example, the chooser might be selecting a dessert from the choices pie, ice cream, or cake. Each response category has some probability of being chosen, and that probability varies depending on other characteristics of the observational unit.

Now reverse the situation and think of yourself as the observation collector for one of the categories. For example, suppose that you sell the pies. The category *Pies* now is a sample category for the vendor, and the response of interest is how many pies can be sold in a day. Given total sales for the day of all desserts, the interest is in the market share of the pies.

**Figure 11.1** diagrams these two ways of looking at categorical distributions.

**Figure 11.1** Customer or Supplier?

- The customer/chooser thinks in terms of *logistic regression*, where the Y variable is which dessert you choose and the X variables affect the probabilities associated with each dessert category.
- The supplier/counter thinks about *log-linear models*, where the Y (responses) is the count, and the X (effect) is the dessert category. There can be other effects interacting with that X.

The modeling traditions for the two outlooks are also different. Customer/chooser-oriented analyses, such as a live/die medical analysis, use continuous Xs (like dose, or how many years you smoked). Supplier/counter-oriented analysts, typified by social scientists, use categorical Xs (like age, race, gender, and personality type) because that keeps the data count-oriented.

The probability distributions for the two approaches are also different. This book won't go into the details of the distributions, but you can be aware of distribution names. Customer/chooser-oriented analysts refer to the *Bernoulli distribution* of the choice. Supplier/counter-oriented analysts refer to the *Poisson distribution* of counts in each category. However, both approaches refer to the *multinomial distribution*.

- To the customer/chooser analysts, the multinomial counts are aggregation statistics.
- To the supplier/counter analysts, the multinomial counts are the count distribution within a fixed total count.

The customer/chooser analyst thinks the basic analysis is fitting response category probabilities. The supplier/counter analyst thinks that basic analysis is a one-way analysis of variance on the counts and uses weights because the distribution is Poisson instead of normal.

Both orientations are right—they just have different outlooks on the same statistical phenomenon.

In this book, the emphasis is on the customer/chooser point of view, also known as the *logistic regression* approach. With logistic regression, it is important to distinguish the responses (Ys), which have the random element, from the factors (Xs), which are fixed from the point of view of the model. The Xs and Ys must be distinguished before the analysis is started.

Let's be clear on what the Xs and Ys are from the chooser point of view:

- Responses (Ys) identify a choice or an outcome. They have a random element because the choice is not determined completely by the X factors. Examples of responses are patient outcome (lived or died), or desert preference (Gobi or Sahara).
- Factors (Xs) identify a sample population, an experimentally controlled condition, or an adjustment factor. They are not regarded as random even if you randomly assigned them. Examples of factors are gender, age, treatment, or block.

**Figure 11.2** illustrates the X and Y variables for both outlooks on categorical models.

The other point of view is the *log-linear model* approach. The log-linear approach regards the count as the Y variable and all the other variables as Xs. After fitting the whole model, the effects that are of interest are identified. Any effect that has no response category variable is discarded, since it is just an artifact of the sampling design. Log-linear modeling uses a technique called *iterative proportional fitting* to obtain test statistics. This process is also called *raking*.

**Figure 11.2** Categories or Counts?**Outlook 1: Categorical Responses**

Dessert Choice is the  $y$ , Count is a frequency, the random response, with a multinomial distribution.

Dessert Choice	Select
Pie	3
Ice Cream	4
Cake	2

**Dessert Choice**

Pie  
Pie  
Pie  
Ice Cream  
Ice Cream  
Ice Cream  
Ice Cream  
Ice Cream  
Cake  
Cake

The frequency count enables you to represent the data compactly. You can expand the data in different ways and the analysis will be the same.

**Outlook 2: Distribution of Counts**

Dessert Choice This is the  $y$ , the random response with a Poisson distribution.

Dessert Choice	Count
Pie	3
Ice Cream	4
Cake	2

**Pie    Ice Cream    Cake**

1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1

## A Simulated Categorical Response

A good way to learn statistical techniques is to simulate data with known properties, and then analyze the simulation to see whether you find the structure that you put into the simulation.

These steps describe the simulation process:

1. Simulate one batch of data and then analyze.
2. Simulate more batches, analyze them, and notice how much they vary.
3. Simulate a larger batch to notice that the estimates have less variance.
4. Do a batch of batches—simulations that for each run obtain sample statistics over a new batch of data.
5. Use this last batch of batches to look at the distribution of the test statistics.

**Note:** The interactive teaching modules under **Help > Sample Data > Teaching Scripts** are simulators that follow this general process for exploring a range of core statistical concepts.

## Simulating Some Categorical Response Data

Let's make a world where there are three soft drinks. The most popular ("Sparkle Cola") has a 50% market share and the other two ("Kool Cola" and "Lemonitz") are tied at 25% each. To simulate a sample from this population, we create a data table that has one variable (call it Drink Choice). The variable is drawn as a random categorical variable using the following formula:

```
p = Random Uniform();
If (p < 0.25 => "Kool Cola"
    If (p < 0.5 => "Lemonitz"
        Else => "Sparkle Cola"));
```

This formula first draws a uniform random number between 0 and 1 using the **Random Uniform** function, and assigns the result to a temporary variable p. Then it compares that random number using **If** conditions and selects the first response where the condition is true. Each case returns the character name of a soft drink as the response value.

- ⓐ Select **Help > Sample Data Library** and open Cola.jmp, which contains the formula shown previously.
- ⓑ Right-click the Drink Choice column header and select **Formula** to display the stored formula.

Note that the two statements in the column formula are delimited with a semicolon. The Formula Editor operations needed to construct this formula to include the following operations:

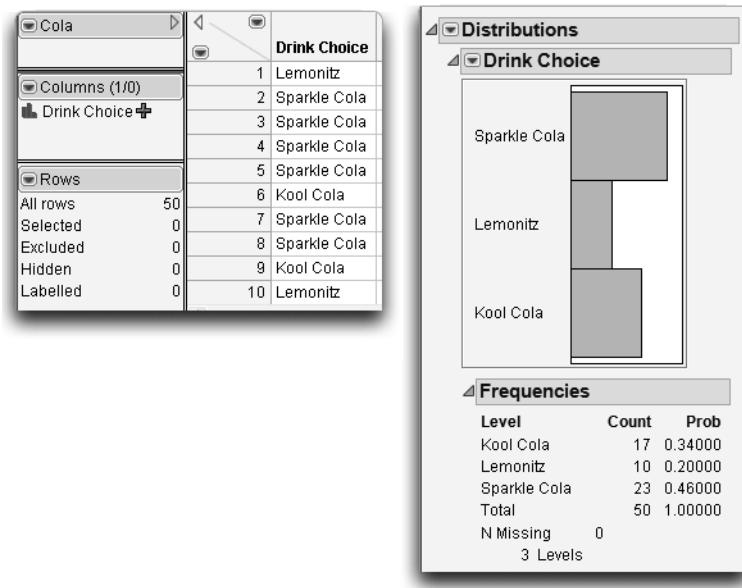
- **Local Variables: New Local** creates a temporary variable named *p*.
- The **If** statement from the **Conditional** functions assigns soft drink names to probability conditions given by the **Random Uniform** function found in the **Random** functions.

The table is stored with no rows. A data table stored with no rows and columns with formulas is called a *table template*.

- ✓ Select **Rows > Add Rows** to add 50 rows to the table.
- ✓ Select **Analyze > Distribution** and assign Drink Choice to **Y, Columns**, which gives an analysis similar to (but not exactly like) that in **Figure 11.3**.

Don't expect to get the same numbers that we show here because the formula generates random data. Each time the computations are performed, a different set of data is produced. Even though the data are based on the true probabilities of 0.25, 0.25, and 0.50, the estimates are different (0.34, 0.20, and 0.46). Your data have random values with somewhat different probabilities.

**Figure 11.3** Histogram and Frequencies of Simulated Data



## Variability in the Estimates

The following sections distinguish between  $\rho$  (Greek rho), the true value of a probability, and  $p$ , its estimate. The true value  $\rho$  is a fixed number, an unknowable "true" value. But its estimate  $p$  is an outcome of a random process, so variability is associated with it.

You cannot compute a standard deviation of the original responses—of Kool Cola, and so on, because they are character values. However, the variability in the probability estimates is well-defined and computable.

Just as with continuous variables, the variability of an estimate is expressed by its variance or its standard deviation, although the quantities are computed with different formulas. The variance of  $p$  is given by the formula

$$\frac{p(1-p)}{n}$$

For Sparkle Cola, having a  $p$  of 0.50, the variance of the probability estimate is  $(0.5 * 0.5) / 50 = 0.005$ . The standard deviation of the estimate is the square root of this variance, 0.07071. Table 11.1 compares the difference between the true  $p$  and its estimate  $p$ . Then, it compares the true standard deviation of the statistic  $p$ , and the standard error of  $p$ , which estimates the standard deviation of  $p$ .

Remember, the term *standard error* is used to label an estimate of the standard deviation of another estimate. Only because this is a simulation with known true values (parameters) can you see both the standard errors and the true standard deviations.

**Table 11.1.** Simulated Probabilities and Estimates

Level	$p$ , the True Probability	$p$ , the Estimate of $p$	True Standard Deviation of the Estimate	Standard Error of the Estimate
Kool Cola	0.25	0.34	0.06124	0.06699
Lemonitz	0.25	0.20	0.06124	0.05657
Sparkle Cola	0.50	0.46	0.07071	0.07048

This simulation shows a lot of variability. As with normally distributed data, you can expect to find estimates that are two standard deviations from the true probability about 5% of the time.

Now let's see how the estimates vary with a new set of random responses.

- ⇨ Right-click the Drink Choice column header and select **Formula**.
- ⇨ Click **Apply** on the calculator to re-evaluate the random formula.
- ⇨ Again perform **Analyze > Distribution** on Drink Choice.
- ⇨ Repeat this evaluate formula/analyze cycle four times.

Each repetition results in a new set of random responses and a new set of estimates of the probabilities. Table 11.2 gives the estimates from our four Monte Carlo runs.

**Table 11.2.** Estimates from Monte Carlo Runs

Level	Probability	Probability	Probability	Probability
Kool Cola	0.32000	0.18000	0.26000	0.40000
Lemonitz	0.26000	0.32000	0.24000	0.18000
Sparkle Cola	0.42000	0.50000	0.50000	0.42000

With only 50 observations, there is a lot of variability in the estimates. The Kool Cola probability estimate varies between 0.18 and 0.40, the Lemonitz estimate varies between 0.18 and 0.32, and the Sparkle Cola estimate varies between 0.42 and 0.50.

## Larger Sample Sizes

What happens if the sample size increases from 50 to 500? Remember that the variance of  $p$  is

$$\frac{p(1-p)}{n}$$

With more data, the denominator is larger, and the probability estimates have a much smaller variance. To see what happens when we add observations:

ⓐ Select **Rows > Add Rows** and enter 450 to get a total of 500 rows.

ⓐ Again perform **Analyze > Distribution** for the response variable Drink Choice.

Five hundred rows give a smaller variance,  $(0.5 * 0.5) / 500 = 0.0005$ , and a standard deviation at about  $\sqrt{0.005} = 0.022$ . The figure here shows the simulation for 500 rows.

Frequencies		
Level	Count	Prob
Kool Cola	134	0.26800
Lemonitz	127	0.25400
Sparkle Cola	239	0.47800
Total	500	1.00000
N Missing	0	
3 Levels		

To see the Std Err Prob column in the report:

ⓐ Right-click in the **Frequencies** table.

ⓐ Select **Columns > StdErr Prob**.

Now we extend the simulation.

ⓐ Repeat the evaluate/analyze cycle four times.

Table 11.3 shows the results of our next four simulations.

**Table 11.3.** Estimates from Four Monte Carlo Runs

Level	Probability	Probability	Probability	Probability
Kool Cola	0.28000	0.25000	0.25600	0.23400
Lemonitz	0.24200	0.28200	0.23400	0.26200
Sparkle Cola	0.47800	0.46800	0.51000	0.50400

Note that the probability estimates are closer to the true values. The standard errors are also much smaller.

## Monte Carlo Simulations for the Estimators

What do the distributions of these counts look like? Variances can be easily calculated, but what is the distribution of the count estimate? Statisticians often use Monte Carlo simulations to investigate the distribution of a statistic.

To simulate estimating a probability (which has a true value of 0.25 in this case) over a given sample size (50 in this case), we'll construct the formula shown below.

The **Random Uniform** function generates a random value distributed uniformly between 0 and 1. The term in the numerator evaluates to 1 or 0 depending on whether the random value is less than 0.25. It is 1 about 25% of the time, and 0 about 75% of the time. This random number is generated 50 times (look at the indices of the summation;  $j = 1$  to 50), and the sum of them is divided by 50.

$$\frac{\sum_{j=1}^{50} \text{Random Uniform}() < 0.25}{50}$$

This formula is a simulation of a Bernoulli event with 50 samplings. The result estimates the probability of getting a 1. In this case, you happen to know the true value of this probability (0.25) because you constructed the formula that generated the data.

Now, it is important to see how well these estimates behave. Theoretically, the mean (expected value) of the estimate,  $p$ , is 0.25 (the true value), and its standard deviation is the square root of

$$\frac{p(1-p)}{n}$$

which is 0.06124.

## Distribution of the Estimates

The sample data has a table template called Simprob.jmp that is a Monte Carlo simulation for the probability estimates of 0.25 and 0.5, based on 50 and 500 trials. You can add rows 1000 to the data to draw 1000 Monte Carlo trials.

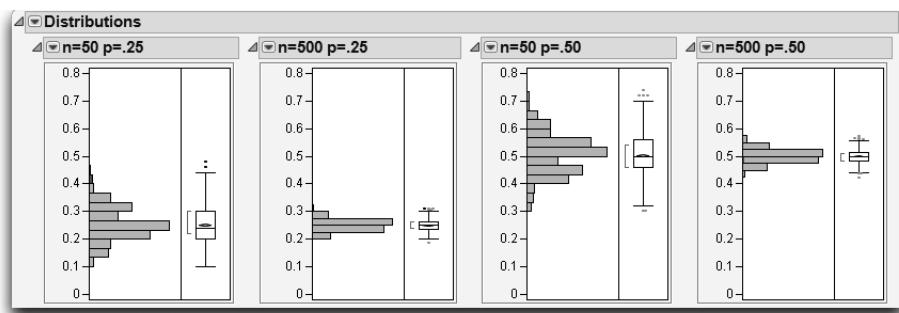
To see how the estimates are distributed:

- ☞ Select **Help > Sample Data** and open Simprob.jmp from the Simulations outline.
- ☞ Select **Rows > Add Rows** and type 1000.
- ☞ Next select **Analyze > Distribution** and use all the columns as Y variables.
- ☞ When the histograms appear, select **Uniform Scaling** from the red triangle menu next to Distributions.
- ☞ Get the grabber (hand) tool from the **Tools** menu and drag the histograms to adjust the bar widths and positions.

**Figure 11.4** and Table 11.4 show the properties as expected:

- The variance decreases as the sample size increases.
- The distribution of the estimates is approximately normally distributed, especially as the sample size gets large.

**Figure 11.4** Histograms for Simulations with Various  $n$  and  $p$  Values



The estimates of the probability  $p$  of getting response indicator values of 0 or 1 behave like sample means. As the sample gets larger, the value of  $p$  gets closer and closer to the true probability (0.25 or 0.50 in our example). Like the mean for continuous data, the standard error of the estimate relates to the sample size by the factor  $1/\sqrt{n}$ .

The Central Limit Theorem applies here. It says that the estimates approach a normal distribution when there are a large number of observations.

**Table 11.4.** Summary of Simulation Results

True Value of $p$	N Used to Estimate $p$	Mean of the Trials of the Estimates of $p$	Standard Deviation of Trials of Estimates of $p$	True Mean of Estimates	True Standard Deviation of the Estimates
0.25	50	0.24774	0.06105	0.25	0.06124
0.25	500	0.24971	0.01936	0.25	0.01936
0.50	50	0.49990	0.07071	0.50	0.07071
0.50	500	0.50058	0.02236	0.50	0.02236

## The $\chi^2$ Pearson Chi-Square Test Statistic

Because of the normality of the estimates, it is reasonable to use normal-theory statistics on categorical response estimates. Remember that the Central Limit Theorem says that the sum of a large number of independent and identically distributed random values have a nearly normal distribution.

However, there is a big difference between having categorical and continuous responses. With categorical responses, the variances of the differences are known. They are solely a function of  $n$  and the probabilities. The hypothesis specifies the probabilities, so calculations can be made under the null hypothesis. Rather than using an  $F$ -statistic, this situation calls for the  $\chi^2$  (*chi-square*) statistic.

The standard chi-square for this model is the following scaled sum of squares:

$$\chi^2 = \frac{\sum_{j=1}^{k-1} (\text{Observed}_j - \text{Expected}_j)^2}{\text{Expected}_j}$$

where Observed and Expected refer to cell counts rather than probabilities.

# The $G^2$ Likelihood-Ratio Chi-Square Test Statistic

The Pearson chi-square assumes normality of the estimates. However, another type of chi-square test is calculated with direct reference to the probability distribution of the response and does not require normality.

Define the *maximum likelihood estimator* (MLE) to be the one that finds the values of the unknown parameters that maximize the probability of the data.

In statistical language, the MLE finds parameters that make the data that actually occurred less improbable than they would be with any other parameter values. The term *likelihood* means that the probability has been evaluated as a function of the parameters with the data fixed.

It would seem that this requires a lot of guesswork in finding the parameters that maximize the likelihood of the observed data. However, just as in the case of least squares, mathematics can provide shortcuts to computing the ideal coefficients. There are two fortunate shortcuts for finding a maximum likelihood estimator:

- Because observations are assumed to be independent, the joint probability across the observations is the product of the probability functions for each observation.
- Because addition is easier than multiplication, instead of multiplying the probabilities to get the joint probability, you add the logarithms of the probabilities, which gives the *log-likelihood*.

This makes for easy computations. Remember that an individual response has a multinomial distribution, so the probability is  $\rho_i$  for the  $i=1$  to  $r$  probabilities over the  $r$  response categories.

Suppose you did the cola simulation, and your first five responses were: Kool Cola, Lemonitz, Sparkle Cola, Sparkle Cola, and Lemonitz. For Kool Cola, Lemonitz, and Sparkle Cola, denote the probabilities as  $\rho_1$ ,  $\rho_2$ , and  $\rho_3$  respectively. The joint log-likelihood is:

$$\log(\rho_1) + \log(\rho_2) + \log(\rho_3) + \log(\rho_3) + \log(\rho_2)$$

It turns out that this likelihood is maximized by setting the probability parameter estimates to the category count divided by the total count, giving

$$p_1 = n_1/n = 1/5$$

$$p_2 = n_2/n = 2/5$$

$$p_3 = n_3/n = 2/5$$

where  $p_1$ ,  $p_2$ , and  $p_3$  estimate  $\rho_1$ ,  $\rho_2$ , and  $\rho_3$ . Substituting this into the log-likelihood gives the maximized log-likelihood of

$$\log(1/5) + \log(2/5) + \log(2/5) + \log(2/5) + \log(2/5)$$

At first it might seem that taking logarithms of probabilities is a mysterious and obscure thing to do, but it is actually very natural. You can think of the negative logarithm of  $p$  as the number of binary questions that you need to determine which of  $1/p$  equally likely outcomes happens. The negative logarithm converts units of probability into units of information. You can think of the negative log likelihood as the *surprise* value of the data because surprise is a good word for unlikeliness.

## Likelihood Ratio Tests

One way to measure the credibility for a hypothesis is to compare how much surprise (-log-likelihood) there would be in the actual data with the hypothesized values compared with the surprise at the maximum likelihood estimates. If there is too much surprise, then you have reason to throw out the hypothesis.

It turns out that the distribution of twice the difference in these two surprises (-log-likelihood) values approximately follows a chi-square distribution.

Here is the setup: Fit a model twice. The first time, fit using maximum likelihood with no constraints on the parameters. The second time, fit using maximum likelihood, but constrain the parameters by the null hypothesis that the outcomes are equally likely. It happens that twice the difference in log-likelihoods has an approximate chi-square distribution (under the null hypothesis). These chi-square tests are called *likelihood ratio chi-squares*, or *LR chi-squares*.

Twice the difference in the log-likelihood is a likelihood ratio chi-square test.

The likelihood ratio tests are very general. They occur not only in categorical responses, but also in a wide variety of situations.

## The G<sup>2</sup> Likelihood Ratio Chi-Square Test

Let's focus on Bernoulli probabilities for categorical responses. The log-likelihood for a whole sample is the sum of natural logarithms of the probabilities attributed to the events that actually occurred.

$$\text{log-likelihood} = \sum \ln(\text{probability the model gives to events that occurred in data})$$

The likelihood ratio chi-square is twice the difference in the two likelihoods, when one is constrained by the null hypothesis and the other is unconstrained.

$$G^2 = 2 (\text{log-likelihood(unconstrained)} - \text{log-likelihood(constrained)})$$

This is formed by the sum over all observations

$$G^2 = 2 \sum [\log(p_{y_i}) - \log(\rho_{y_i})]$$

where  $\rho_{y_i}$  is the hypothesized probability and  $p_{y_i}$  is the estimated rate for the events  $y_i$  that actually occurred.

If you have already collected counts for each of the responses, and bring the subtraction into the log as a division, the formula becomes

$$G^2 = 2 \sum n_i \log \frac{\rho_{y_i}}{p_{y_i}}$$

To compare with the Pearson chi-square, which is written schematically in terms of counts, the LR chi-square statistic can be written

$$G^2 = 2 \sum \text{observed} \left( \log \frac{\text{expected}}{\text{observed}} \right)$$

## Univariate Categorical Chi-Square Tests

A company gave 998 of its employees the Myers-Briggs Type Inventory (MBTI) questionnaire. The test is scored to result in a 4-character personality type for each person. Scores are based on four dichotomies (E/I = Extroversion/Introversion, S/N = Sensing/Intuition, T/F = Thinking/Feeling, and J/P = Judging/Perceiving), giving 16 possible outcomes (see **Figure 11.5**). The company wanted to know

whether its employee force was statistically different in personality types from the general population.

## Comparing Univariate Distributions

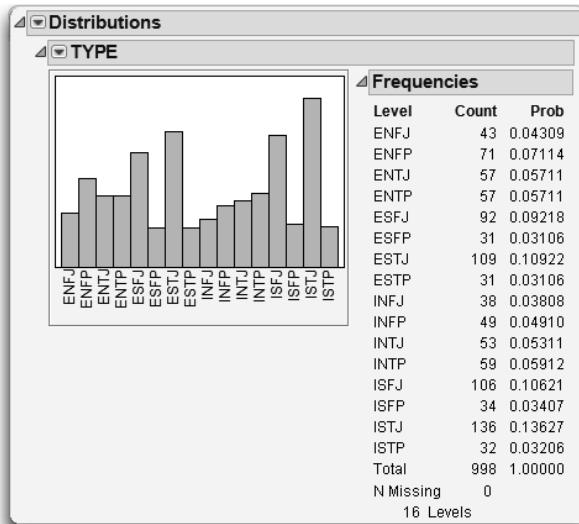
The sample data table Mb-dist.jmp has a column called TYPE to use as a Y response, and a Count column to use as a frequency. To see the company test results:

- ~ Select **Help > Sample Data Library** and open Mb-dist.jmp.
- ~ Select **Analyze > Distribution**.
- ~ Assign Type to **Y, Columns** and Count to Freq. Note that Count is given the Freq role, and is assigned automatically

When the report appears,

- ~ Select **Display Options > Horizontal Layout** from the red triangle menu next to TYPE to see the report in **Figure 11.5**.

**Figure 11.5** Histogram and Frequencies for Myers-Briggs Data



To test the hypothesis that the personality test results at this company occur at the same rates as the general population:

- ~ Select **Test Probabilities** from the red triangle menu next to TYPE.

A window appears with boxes to enter proportions (Hypoth Prob) you want to test against for each category.

- ⇨ Edit the Hypoth Prob (hypothesized probability) values by clicking and then entering the values shown here.

These are the data collected for the general population for each personality type (category).

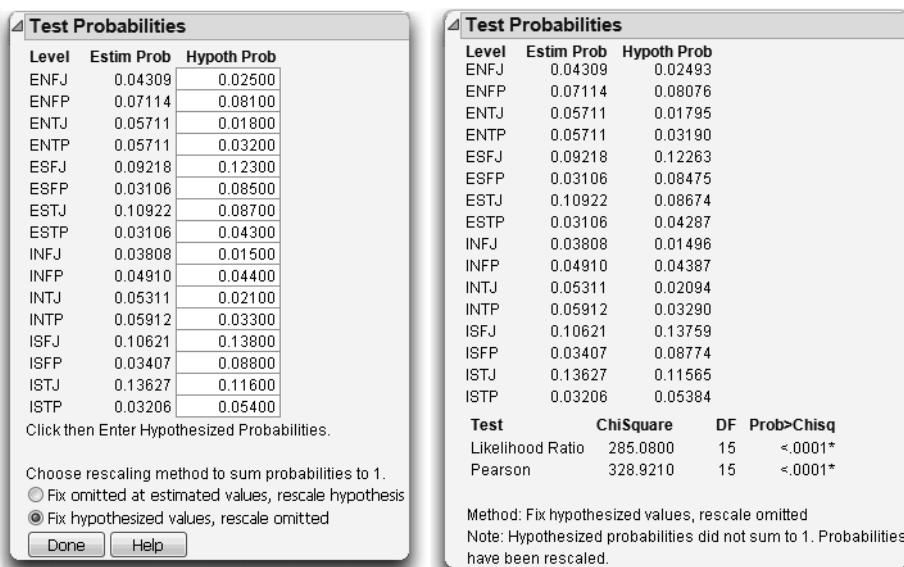
- ⇨ Click the second radio button in the window and then click **Done**.

You now see the test results appended to the Test Probabilities table, as shown in the table on the right in **Figure 11.6**.

ENFJ	0.025
ENFP	0.081
ENTJ	0.018
ENTP	0.032
ESFJ	0.123
ESFP	0.085
ESTJ	0.087
ESTP	0.043
INFJ	0.015
INFP	0.044
INTJ	0.021
INTP	0.033
ISFJ	0.138
ISFP	0.088
ISTJ	0.116
ISTP	0.054

Note that the company does have a significantly different profile than the general population. Both chi-square tests are highly significant.

**Figure 11.6** Test Probabilities Report for the Myers-Briggs Data



By the way, some people find it upsetting that different statistical methods get different results. Actually, the  $G^2$  (likelihood ratio) and  $X^2$  (Pearson) chi-square statistics are usually close.

## Charting to Compare Results

Let's continue with the Myers-Briggs example. We can chart the results and see the contrast between the general population and the company results.

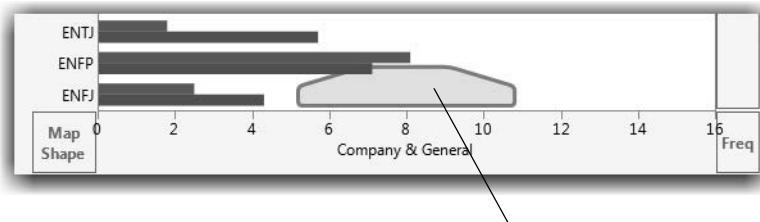
- ❖ Select **Graph > Graph Builder** to see the drag-and-drop palette for graphing.
- ❖ To begin, drag TYPE to the Y drop zone and Company to the X drop zone. The available drop zones are shaded in blue.
- ❖ When the initial points appear, click the bar chart from the selection of graph icons above the plot frame, as shown here. You will see the bar chart for Company scores.



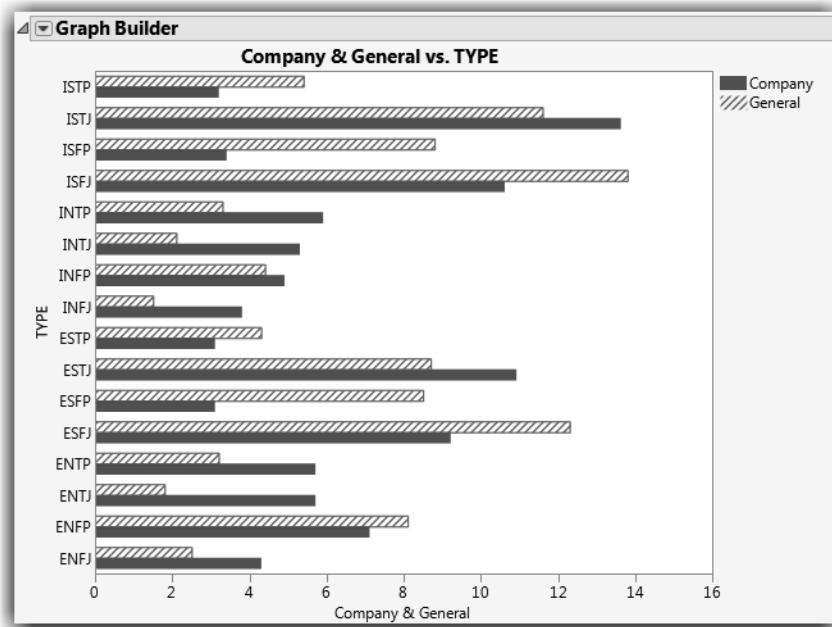
- ❖ Now drag General to the X drop zone just inside the plot frame above Company. The drop zone is shaded in blue and outlined, as shown at the top in **Figure 11.7**.
- ❖ To better distinguish the bars for Company and General, right-click on the legend for General, select **Fill Pattern**, and select a different fill pattern.

The bottom of **Figure 11.7** shows that the company appears to have more ISTJs (introvert-sensing-thinking-judging) and fewer ESFPs (extrovert-sensing-feeling-perceiving) and ESTPs (extrovert-sensing-thinking-perceiving) than the general population.

It's also easy to see that the company has more people who scored high, in general, on IN (Introvert-Intuitive) than the general population.

**Figure 11.7** Mean Personality Scores for Company and General Population

To overlay variables, drop the second X variable, Company, in the plot area defined by blue outline.



## Exercises

1. P&D Candies produces a bag of assorted sugar candies, called Moons, in several colors. Based on extensive market research, they have decided on the following mix of Moons in each bag: Red 20%, Yellow 10%, Brown 30%, Blue 20%, and Green 20%. A consumer advocate suspects that the mix is not what the company claims, so he or she gets a bag containing 100 Moons. The 100 pieces of candy are represented in the sample data table Candy.jmp (fictional data).
  - (a) Can the consumer advocate reasonably claim that the company's mix is not as they say?
  - (b) Do you think a single-bag sample is representative of all the candies produced?
2. One of the ways that public schools can make extra money is to install vending machines for students to access between classes. Suppose a high school installed three drink machines for different manufacturers in a common area of the school. After one week, they collected information about the number of visits to each machine, as shown in the following table:

Machine A	Machine B	Machine C
1546	1982	1221

Is there evidence of the students preferring one machine over another?





# 12

## Categorical Models

### Overview

Chapter 11, “Categorical Distributions,” introduced the distribution of a single categorical response. You were introduced to the Pearson and the likelihood ratio chi-square tests and saw how to compare univariate categorical distributions.

This chapter covers multivariate categorical distributions. In the simplest case, the data can be presented as a two-way contingency table (also called a cross tabulation or cross tab) of frequency counts. The contingency table contains expected cell probabilities and counts formed from products of marginal probabilities and counts. The chi-square test again is used for the contingency table and is the same as testing multiple categorical responses for independence.

Correspondence analysis is shown as a graphical technique useful when the response and factors have many levels or values.

Also, a more general categorical response model is used to introduce nominal and ordinal logistic regression, which allows multiple continuous or categorical factors.

## Chapter Contents

Overview .....	303
Fitting Categorical Responses to Categorical Factors: Contingency Tables .....	305
Testing with $G^2$ and $X^2$ Statistic .....	305
Looking at Survey Data .....	306
Car Brand by Marital Status .....	310
Car Brand by Size of Vehicle .....	311
Two-Way Tables: Entering Count Data.....	312
Expected Values under Independence.....	313
Entering Two-Way Data into JMP .....	314
Testing for Independence.....	314
If You Have a Perfect Fit .....	316
Special Topic: Correspondence Analysis— Looking at Data with Many Levels	318
Continuous Factors with Categorical Responses: Logistic Regression .....	321
Fitting a Logistic Model .....	321
Degrees of Fit.....	325
A Discriminant Alternative .....	326
Inverse Prediction .....	327
Polytomous (Multinomial) Responses: More Than Two Levels .....	330
Ordinal Responses: Cumulative Ordinal Logistic Regression.....	331
Surprise: Simpson's Paradox: Aggregate Data versus Grouped Data .....	334
Generalized Linear Models.....	337
Exercises.....	342

# Fitting Categorical Responses to Categorical Factors: Contingency Tables

When a categorical response is examined in relationship to a categorical factor (in other words, both X and Y are categorical), the question is: do the response probabilities vary across factor-defined subgroups of the population? Comparing a continuous response and a categorical factor in this way was covered in Chapter 9, “Comparing Many Means: One-Way Analysis of Variance.” In that chapter, means were fit for each level of a categorical variable and tested using an ANOVA. When the continuous response is replaced with a categorical response, the equivalent technique is to estimate response probabilities for each subgroup and test that they are the same across the subgroups.

The subgroups are defined by the levels of a categorical factor (X). For each subgroup, the set of response probabilities must add up to 1. For example, consider the following:

- The probability of whether a patient lives or dies (response probabilities) depending on whether the treatment (categorical factor) was drug or placebo
- The probability that type of car purchased (response probabilities) depending on marital status (the categorical factor)

To estimate response probabilities for each subgroup, you divide the count in a given response level by its total count.

## Testing with $G^2$ and $X^2$ Statistic

You want to test whether the factor affects the response. The null hypothesis is that the response probabilities are the same across subgroups. The model compares the fitted probabilities over the subgroups to the fitted probabilities combining all the groups into one population (a constant response model).

As a measure of fit for the models that you want to compare, you can use the negative log-likelihood to compute a likelihood-ratio chi-square test. To do this, subtract the log-likelihoods for the two models and multiply by 2. For each observation, the log-likelihood is the log of the probability attributed to the response level of the observation.

**Warning:** When the table is *sparse*, neither the Pearson or likelihood ratio chi-square is a very good approximation to the true distribution. The Cochran criterion, used to determine whether the tests are appropriate, defines sparse as when more than 20% of the cells have expected counts less than 5. JMP presents a warning when this situation occurs.

The Pearson chi-square tends to be better behaved in sparse situations than the likelihood ratio chi-square. However,  $G^2$  is often preferred over  $X^2$  for other reasons, specifically because it is applicable to general categorical models where  $X^2$  is not.

Chapter 11, “Categorical Distributions,” discussed the  $G^2$  and  $X^2$  test statistics in more detail.

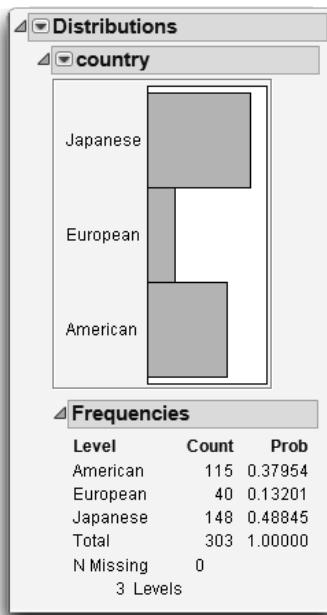
## Looking at Survey Data

Survey data often yield categorical data suitable for contingency table analysis. For example, suppose a company did a survey to find out what factors relate to the brand of automobile people buy. In other words, what type of people buy what type of cars? Cars were classified into three brands: American, European, and Japanese (which included other Asian brands). This survey also contained demographic information (marital status and gender of the purchasers).

The results of the survey are in the sample data table called Car Poll.jmp. A good first step is to examine probabilities for each brand when nothing else is known about the car buyer. Looking at the distribution of car brand gives this information. To see the report on the distribution of brand shown in **Figure 12.1**:

- ~ Select **Help > Sample Data Library** and open Car Poll.jmp.
- ~ Select **Analyze > Distribution** and then assign country to **Y, Columns**.

Overall, the Japanese brands have a 48.8% share.

**Figure 12.1** Histograms and Frequencies for country in Car Poll Data

The next step is to look at the demographic information as it relates to brand of auto.

- ⓐ Select **Analyze > Fit Y by X**, assign country to **Y, Response** and sex, marital status, and size to **X, Factor**.
- ⓐ Click **OK**.

The Fit Y by X platform displays mosaic plots and contingency tables for the combination of country with each of the X variables. By default, JMP displays Count, Total%, Col%, and Row% (listed in the upper left corner of the table) for each cell in the contingency table.

- ⓐ Click the red triangle menu next to Contingency Table or right-click anywhere in the contingency table to see a menu of the optional items to include in the table cells.
- ⓐ Deselect all items except Count and Row% to see the table shown above.

		country			
sex	Count	country			Total
		America	Europe	Japanes	
Female	54	19	65	138	138
	39.13	13.77	47.10		
Male	61	21	83	165	165
	36.97	12.73	50.30		
Total		115	40	148	303

**Note:** Hold down the Ctrl key while deselecting items to broadcast these changes to other analyses.

### Contingency Table: Country by Sex

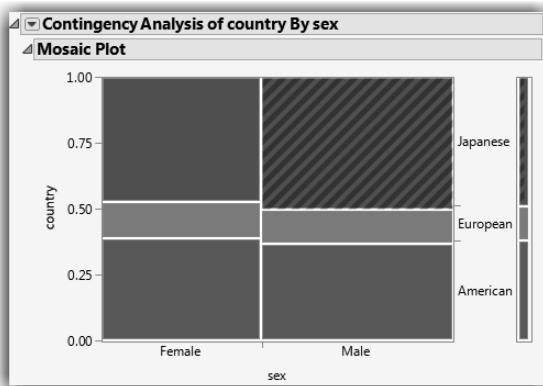
Is the distribution of the response levels different over the levels of other categorical variables? In principle, this is like a one-way analysis of variance, estimating separate means for each sample, but this time they are rates over response categories rather than means.

In the contingency table, you see the response probabilities as the Row% values in the bottom of each cell. The percents for each country are not much different between “Female” and “Male.”

### Mosaic Plot

The Fit Y by X platform for a categorical variable displays information graphically with mosaic plots like the one shown here.

A mosaic plot is a set of side-by-side divided bar plots to compare the subdivision of response probabilities for each sample. The mosaic is formed by first dividing up the horizontal axis according to the sample proportions. Then each of these cells is subdivided vertically by the estimated response probabilities. The area of each rectangle is proportional to the frequency count for that cell. To get a better understanding of the mosaic plot:



- ❖ Hold your mouse over the cells in the mosaic plot to display the number of rows, the corresponding rows in the data table, and the relative frequencies.
- ❖ Right-click anywhere on the plot and select a **Cell Labeling** option, such as **Show Counts** or **Show Percents**.
- ❖ Use your cursor to grab the edge of the country legend and drag to widen it.
- ❖ Click in any cell to highlight, which also selects the corresponding rows in the data table.

### Testing Marginal Homogeneity

Now ask the question, "Are the response probabilities significantly different across the samples (in this example, male and female)?" Specifically, is the proportion of sales by country the same for males and females? The null hypothesis that the distributions are the same across the sample's subgroup is sometimes referred to as the hypothesis of *marginal homogeneity*.

Instead of regarding the categorical X variable as fixed, you can consider it as another Y response variable and look at the relationship between two Y response variables. The test would be the same, but the null hypothesis would be known by a different name, as the *test for independence*.

When the response was continuous, there were two ways to get a test statistic that turned out to be equivalent:

- Look at the distribution of the estimates, usually leading to a *t*-test.
- Compare the fit of a model with a submodel, leading to an *F*-test.

The same two approaches work for categorical models. However, the two approaches to getting a test statistic for a contingency table both result in chi-square tests.

- If the test is derived in terms of the distribution of the estimates, then you are led to the Pearson  $X^2$  form of the  $\chi^2$  test.
- If the test is derived by comparing the fit of a model with a submodel, then you are led to the likelihood-ratio  $G^2$  form of the  $\chi^2$  test.

For the likelihood ratio chi-square ( $G^2$ ), two models are fit by maximum likelihood. One model is constrained by the hypothesis that assumes a single response population, and the other is not constrained. Twice the difference of the log-likelihoods from the two models is a chi-square statistic for testing the hypothesis. The table here has the chi-square tests that test whether country of car purchased is a function of sex.

Tests			
	N	DF	-LogLike R Square (U)
	303	2	0.15593790 0.0005
Test	ChiSquare	Prob>ChiSq	
Likelihood Ratio	0.312	0.8556	
Pearson	0.312	0.8556	

The model constrained by the null hypothesis (fitting only one set of response probabilities) has a negative log-likelihood of 298.45. After you partition the sample by the gender factor, the negative log-likelihood is reduced to 298.30. The difference in log-likelihoods is 0.1559, reported in the -LogLike line. This doesn't

account for much of the variation. The likelihood ratio (LR) chi-square is twice this difference, that is,  $G^2 = 0.312$ , and has a nonsignificant  $p$ -value of 0.8556. These statistics don't support the conclusion that the car country purchase response depends on the gender of the driver.

**Note:** The log-likelihood values of the null and constrained hypotheses are not shown in the Fit Y By X report. If you are interested in seeing them, launch **Analyze > Fit Model** using the same variables: country as Y, sex as an effect. The resulting report has more detail.

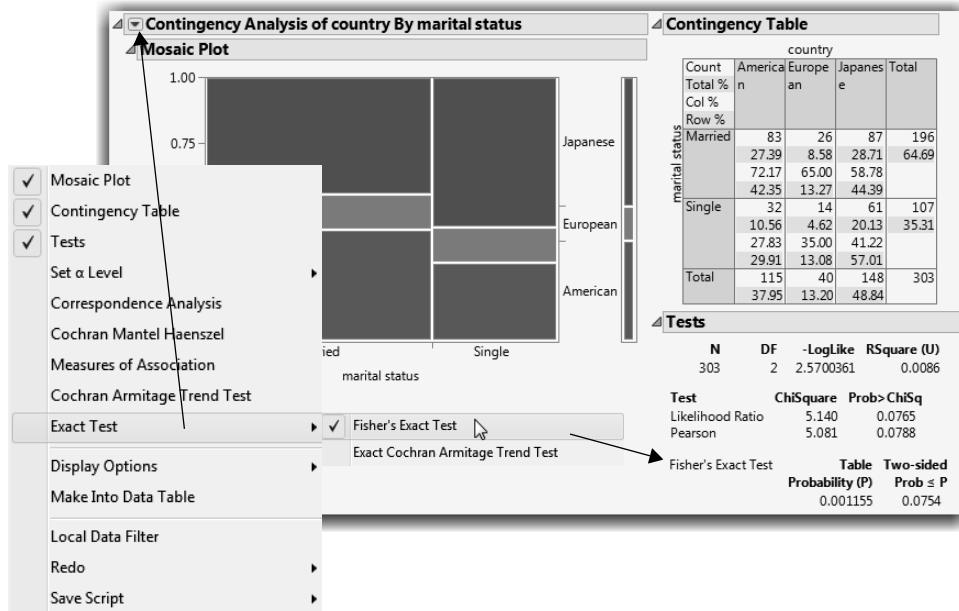
Whole Model Test				
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.15594	2	0.311876	0.8556
Full	298.29531			
Reduced	298.45125			
RSquare (U)	0.0005			
AICc	604.725			
BIC	619.446			
Observations (or Sum Wgts)	303			

If you want to think about the distribution of the estimates, then in each cell, you can compare the actual proportion to the proportion expected under the hypothesis. Square the number and divide by something close to its variance, giving a cell chi-square. The sum of these cell chi-square values is the Pearson chi-square statistic  $X^2$ , here also 0.312, which has a  $p$ -value of 0.8556. In this example, the Pearson chi-square happens to be the same as the likelihood ratio chi-square.

## Car Brand by Marital Status

Let's look at the relationships of country to other categorical variables. In the case of marital status (**Figure 12.2**), there is a more significant result, with the  $p$ -value for the Likelihood Ratio  $G^2$  statistic of 0.0765. Married people are more likely to buy the American brands. Why? Perhaps because the American brands are generally larger vehicles, which make them more comfortable for families.

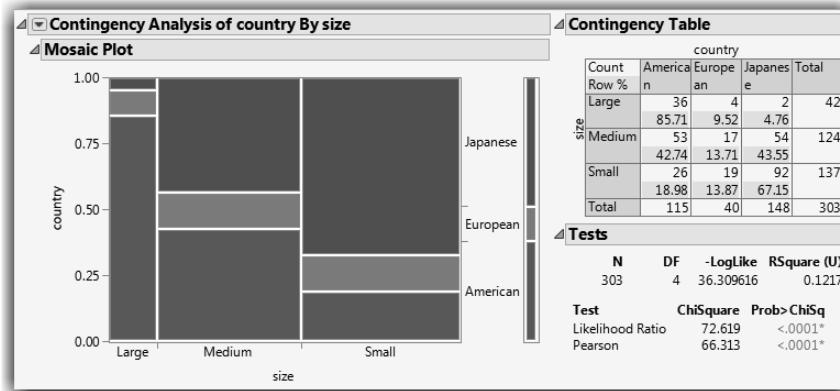
Notice that the red triangle menu next to Contingency Analysis offers many options, including Fisher's Exact Test in JMP Pro. In this example, the result of Fisher's test is 0.0012, with a  $p$ -value of 0.0754.

**Figure 12.2** Mosaic Plot, Crosstabs, and Tests Table for country by marital status

## Car Brand by Size of Vehicle

If marital status is a proxy for size of vehicle, looking at country by size should give more direct information.

The Tests table for country by size (**Figure 12.3**) shows a strong relationship with a very significant chi-square. The Japanese dominate the market for small cars, the Americans dominate the market for large cars, and the European share is about the same in all three markets. The relationship is highly significant, with  $p$ -values less than 0.0001. The null hypothesis that car size and country are independent is easily rejected.

**Figure 12.3** Mosaic Plot, Crosstabs, and Tests Table for country by size

## Two-Way Tables: Entering Count Data

Often, raw categorical data is presented in textbooks in a *two-way table* like the one shown below. The levels of one variable are the rows, the levels of the other variable are the columns, and cells contain frequency counts. For example, data for a study of alcohol and smoking (based on Schiffman, 1982) is arranged in a two-way table, like this:

		Smoking Relapse	
		Yes	No
Alcohol Consumption	Consumed	20	13
	Did Not Consume	48	96

This arrangement shows the two levels of alcohol consumption ("Consumed" or "Did Not Consume") and levels of whether the subject relapsed and returned to smoking (reflected in the "Yes" column) or managed to stay smoke-free (reflected in the "No" column).

In the following discussion, keep the following things in mind:

- The two variables in this table do not fit neatly into independent and dependent classifications. The subjects in the study were not separated into two groups, with one group given alcohol, and the other not. The interpretation of the data, then, needs to be limited to association, and not cause-and-effect. The tests are regarded as tests of independence of two

responses, rather than the marginal homogeneity of probabilities across samples.

- For a  $2 \times 2$  table, JMP Pro automatically produces Fisher's Exact Test in its results. This test, in essence, computes exact probabilities for the data rather than relying on approximations.

Does it appear that alcohol consumption is related to the subject's relapse status? Phrased more statistically, if you assume that these variables are independent, are there surprising entries in the two-way table? To answer this question, we must know what values would be expected in this table, and then determine whether there are observed results that are different from these expected values.

## Expected Values under Independence

To further examine the data, the following table shows the totals for the rows and columns of the two-way table. The row and column totals have been placed along the right and bottom margins of the table and are therefore called *marginal totals*.

		<b>Smoking Relapsed</b>		<b>Total</b>
		<b>Yes</b>	<b>No</b>	
<b>Alcohol Consumption</b>	<b>Consumed</b>	20	13	33
	<b>Did Not Consume</b>	48	96	144
	<b>Total</b>	68	109	177

These totals aid in determining what values would be expected if alcohol consumption and relapse to smoking were not related.

As is usual in statistics, assume at first that there is no relationship between these variables. If this assumption is true, then the proportion of people in the "Yes" and "No" columns should be equal for each level of the alcohol consumption variable. If there was no effect for consumption of alcohol, then we expect these values to be the same except for random variation. To determine the *expected value* for each cell, compute

$$\frac{\text{Row total} \times \text{Column Total}}{\text{Table Total}}$$

for each cell. Instead of computing it by hand, let's enter the data into JMP to perform the calculations.

## Entering Two-Way Data into JMP

Before two-way table data can be analyzed, it needs to be *flattened* or *stacked*. Then it is arranged in two data columns for the variables, and one data column for frequency counts. Follow these steps:

- ✓ Select **File > New > Data Table** to create a new data table.
- ✓ Click on the title of Column 1 and name the column **Alcohol Consumption**.
- ✓ Select **Cols > New Columns** to create a second column in the data table (or, double-click in the column header area next to the first column). Name this column **Relapsed**.
- ✓ Create a third column named **Count** to hold the cell counts from the two-way table.
- ✓ Select the **Count** column, and select **Cols > Preselect Role > Freq** to assign the frequency role.
- ✓ Select **Rows > Add Rows** and add four rows to the table—one for each cell in the two-way table.
- ✓ Enter the data so that the data table looks like the one shown here.

These steps have been completed, and the resulting table is included in the sample data library as **Alcohol.jmp**.

	Alcohol Consumption	Relapsed	Count
1	Consumed	Yes	20
2	Consumed	No	13
3	Didn't Consume	Yes	48
4	Didn't Consume	No	96

## Testing for Independence

One explanatory note is in order at this point. Although the computations in this situation use the counts of the data, the statistical test deals with proportions. The *independence* that we are concerned with is the independence of the probabilities associated with each cell in the table. Specifically, let

$$\rho_i = \frac{n_i}{n} \text{ and } \rho_j = \frac{n_j}{n}$$

where  $\rho_i$  and  $\rho_j$  are, respectively, the probabilities associated with each of the  $i$  rows and  $j$  columns. Now, let  $\rho_{ij}$  be the probability associated with the *cell* located at the  $i$ th row and  $j$ th column. The null hypothesis of independence is that  $\rho_{ij} = \rho_i \rho_j$ .

Although the computations that we present use counts, do not forget that the essence of the null hypothesis is about probabilities.

The test for independence is the  $X^2$  statistic, whose formula is

$$\sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

To compute this statistic in JMP:

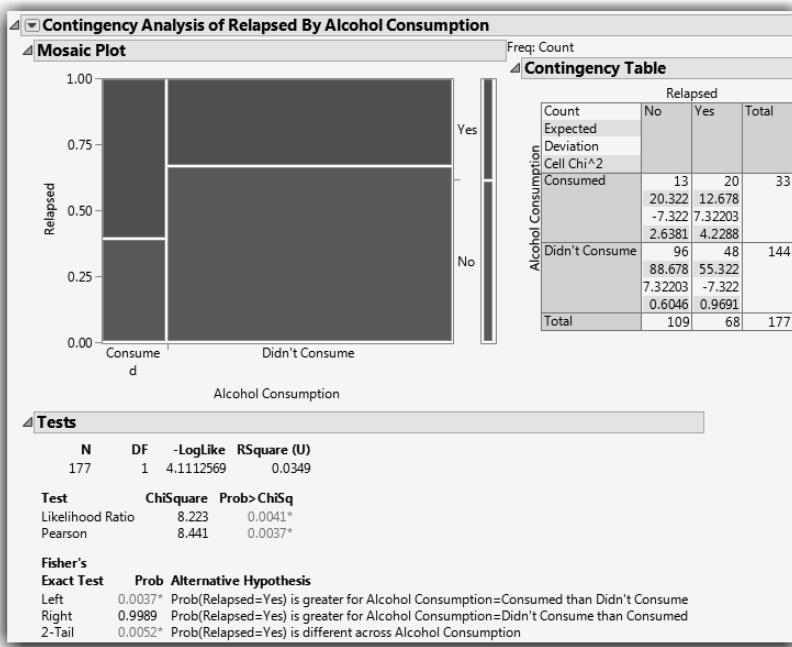
- ❖ Select **Analyze > Fit Y By X**. Assign Alcohol Consumption to **X** and Relapsed to **Y**. Because Count was pre-assigned the Freq role, it automatically is displayed as the **Freq** variable.
- ❖ Click **OK**.

This produces a report that contains the contingency table of counts, which should agree with the two-way table used as the source of the data. To see the information relevant to the computation of the  $X^2$  statistic:

- ❖ Right-click inside the contingency table and deselect **Row%**, **Col%**, and **Total%**.
- ❖ Again, right-click inside the contingency table and make sure that **Count**, **Expected**, **Deviation**, and **Cell Chi Square** are selected.

**Note:** To display all red triangle options for a report, hold down the Alt key (Option on Macintosh) before clicking the red triangle.

The Tests table (**Figure 12.4**) shows the Likelihood Ratio and Pearson Chi-square statistics. The Pearson statistic is the sum of all the cell chi-square values in the contingency table. The  $p$ -values for both chi-square tests are less than 0.05, so the null hypothesis is rejected. Alcohol consumption seems to be associated with whether the patient relapsed into smoking.

**Figure 12.4** Contingency Report

The composition of the Pearson  $\chi^2$  statistic can be seen cell by cell. The cell for “Yes” and “Consumed” in the upper right has an actual count of 20 and an expected count of 12.678. The difference (deviation) between the counts is 7.322. This cell’s contribution to the overall chi-square is

$$\frac{(20 - 12.678)^2}{12.678}$$

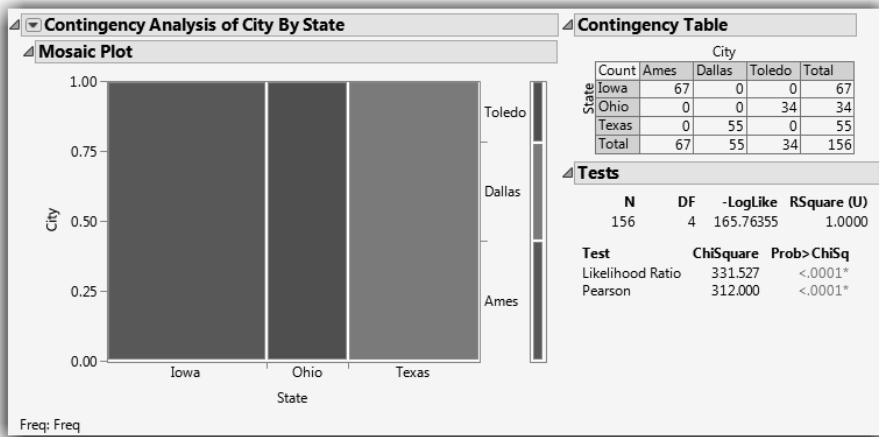
which is 4.22. Repeating this procedure for each cell shows the chi-square as  $2.63 + 0.60 + 4.22 + 0.97 = 8.441$ .

## If You Have a Perfect Fit

If a fit is perfect, every response’s category is predicted with probability 1. The response is completely determined by which sample it is in. In the other extreme, if the fit contributes nothing, then each distribution of the response in each sample subgroup is the same.

For example, consider collecting information for 156 people on what city and state they live in. It's likely that you would think that there is a perfect fit between the city and the state of a person's residence. If the city is known, then the state is almost surely known. **Figure 12.5** shows what this perfect fit looks like.

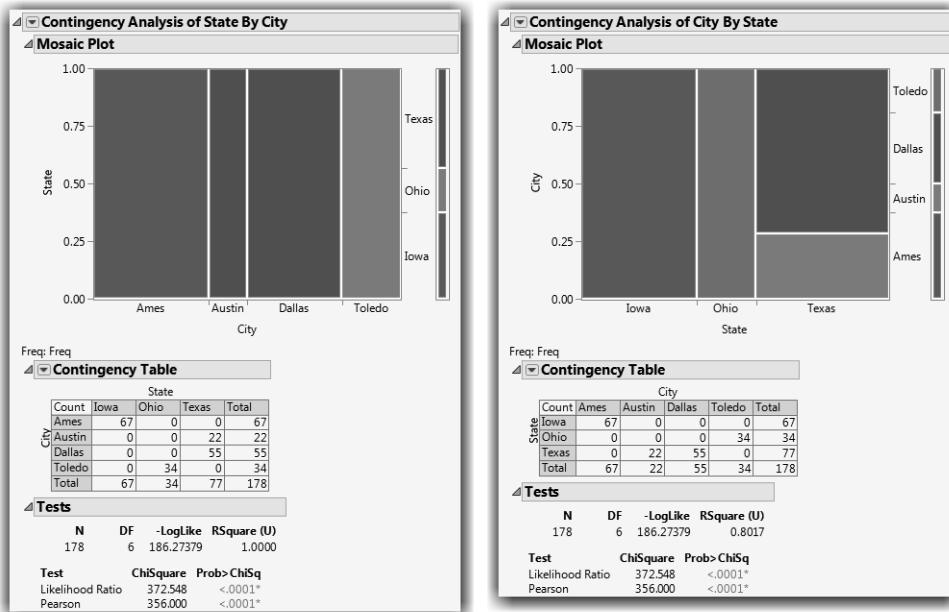
**Figure 12.5** Mosaic Plot, Crosstabs, and Tests Table for City by State



Now suppose the analysis includes people from Austin, a second city in Texas. City still predicts state perfectly, but not the other way around (state does not predict city). Conducting these two analyses shows that the chi-squares are the same. They are invariant if you switch the Y and X variables. However, the mosaic plot and the  $R^2$  are different (**Figure 12.6**).

What happens if the response rates are the same in each cell as in **Figure 12.5**? Examine the artificial data for this situation and notice that the mosaic levels line up perfectly and the chi-squares are zero.

**Figure 12.6** Comparison of Plots, Tables, and Tests When X and Y Are Switched



## Special Topic: Correspondence Analysis – Looking at Data with Many Levels

Correspondence analysis is a graphical technique that shows which rows or columns of a frequency table have similar patterns of counts. Correspondence analysis is particularly valuable when you have many levels, because it is difficult to find patterns in tables or mosaic plots with many levels.

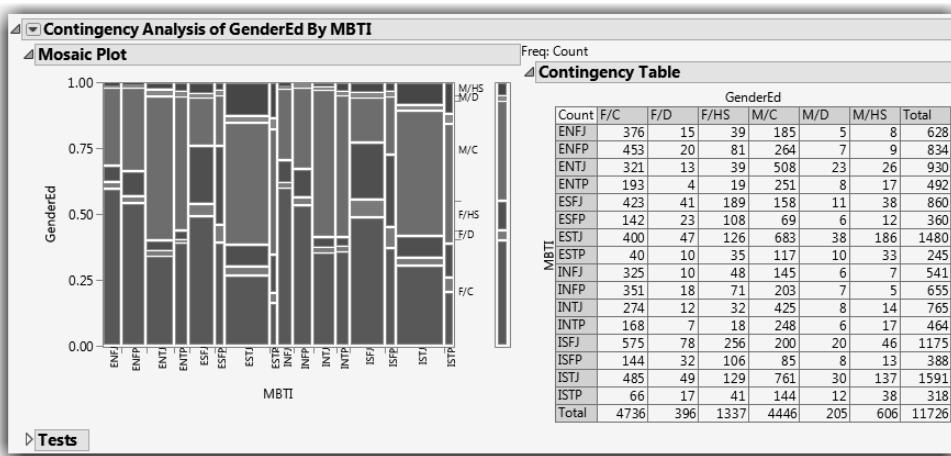
The sample data table Mbtied.jmp has counts of Myers-Briggs personality types by educational level and gender (Myers and McCaulley). The values of educational level (Educ) are D for dropout, HS for high school graduate, and C for college graduate. Gender and Educ are concatenated to form the variable GenderEd. The goal is to determine the relationships between GenderEd and personality type. Remember, there is no implication of any cause-and-effect relationship because there is no way to tell whether personality affects education or education affects personality. The data can, however, show trends. The following example shows how correspondence analysis can help identify trends in categorical data:

- ☞ Select **Help > Sample Data Library** and open Mbti.ed.jmp.
- ☞ Select **Analyze > Fit Y by X** and assign MBTI to **X, Factor** and GenderEd to **Y, Response**. Count is displayed as the Freq variable.
- ☞ Click **OK**.

Now try to make sense out of the resulting mosaic plot and contingency table shown in **Figure 12.7**. It has 96 cells—too big to understand at a glance. A correspondence analysis clarifies some patterns.

**Note:** For purposes of illustration, only the Count column is shown in **Figure 12.7**

**Figure 12.7** Mosaic Plot and Table for MBTI by GenderEd



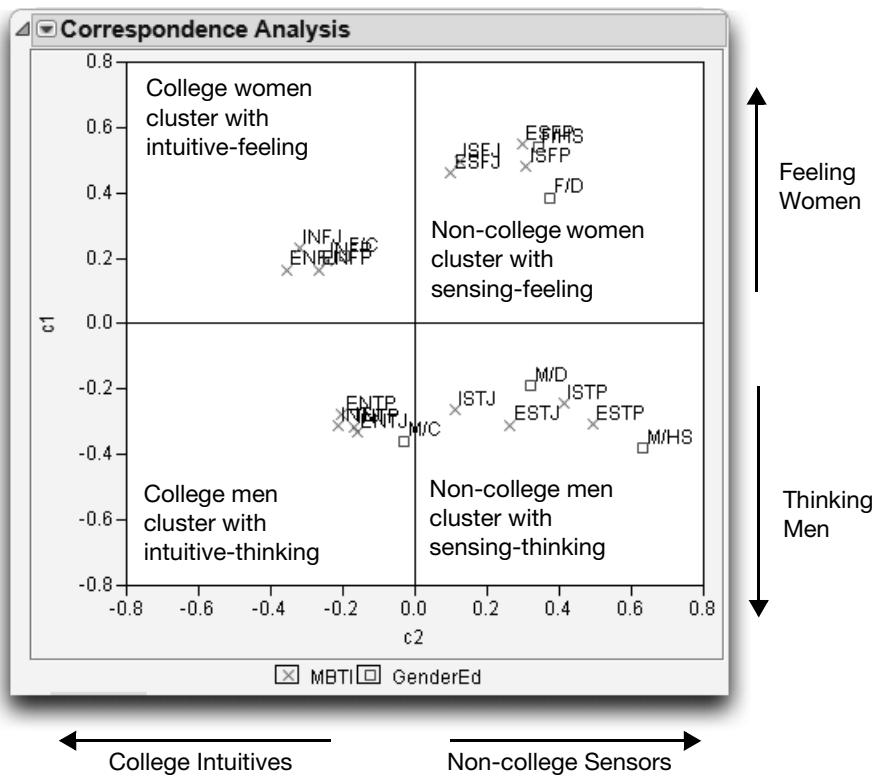
- ☞ Select **Correspondence Analysis** from the red triangle menu next to Contingency Analysis to see the plot in **Figure 12.8**.
- ☞ Resize the graph to better see the labels.

The Correspondence Analysis plot organizes the row and column profiles in a two-dimensional space. The X values that have similar Y profiles tend to cluster together, and the Y values that have similar X profiles tend to cluster together. In this case, you want to see how the GenderEd groups are associated with the personality groups.

This plot shows patterns more clearly. Gender and the Feeling(F)/Thinking(T) component form a cluster, and education clusters with the Intuition(N)/Sensing(S) personality indicator. The Extrovert(E)/Introvert(I) and Judging(J)/

Perceiving(P) types do not separate much. The most separation among these is the Judging(J)/Perceiving(P) separation among the Sensing(S)/Thinking (T) types (mostly non-college men).

**Figure 12.8** Correspondence Analysis Plot



The correspondence analysis indicates that the Extrovert/Introvert and Judging/Perceiving do not separate well for education and gender.

**Note:** Correspondence Analysis is also available from **Analyze > Consumer Research > Multiple Correspondence Analysis**.

# Continuous Factors with Categorical Responses: Logistic Regression

Suppose that a response is categorical, but the probabilities for the response change as a function of a continuous predictor. In other words, you are presented with a problem with a continuous X and a categorical Y. Some situations like this are the following:

- Whether you bought a car this year (categorical) as a function of your disposable income (continuous).
- The type of car that you bought (categorical) as a function of your age (continuous).
- The probability of whether a patient lived or died (categorical) as a function of blood pressure (continuous).

Problems like these call for *logistic regression*. Logistic regression provides a method to estimate the probability of choosing one of the response levels as a smooth function of the factor. It is called logistic regression because the S-shaped curve that it uses to fit the probabilities is called the logistic function.

## Fitting a Logistic Model

The Spring.jmp sample data is a weather record for the month of April. The variable Precip measures rainfall.

ⓐ Select **Help > Sample Data Library** and open Spring.jmp.

ⓑ Add a column named Rained to categorize rainfall using the formula shown here.

$$\text{If} \begin{cases} \text{Precip} > 0.02 \Rightarrow \text{"Rainy"} \\ \text{else} \qquad \qquad \Rightarrow \text{"Dry"} \end{cases}$$

- Select **Analyze > Distribution** to generate a histogram and frequency table of the Rained variable.

Out of the 30 days in April, there were 9 rainy days. Therefore, with no other information, you predict a  $9/30 = 30\%$  chance of rain for every day.

Suppose you want to increase your probability of correct predictions by including other variables. You might use morning temperature or barometric pressure to help make more informed predictions. Let's examine these cases.

In each case, the thing being modeled is the probability of getting one of several responses. The probabilities are constrained to add to 1. In the simplest situation, like this rain example, the response has two levels (a *binary* response). Remember that statisticians like to take logs of probabilities. In this case, what they fit is the difference in logs of the two probabilities as a linear function of the factor variable.

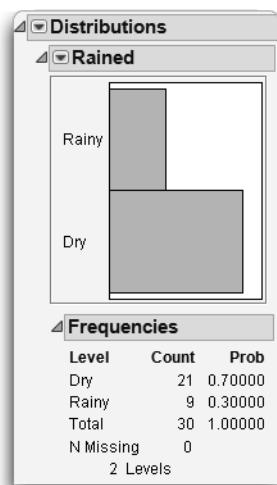
If  $p$  denotes the probability for the first response level, then  $1-p$  is the probability of the second, and the linear model is written

$$\log(p) - \log(1-p) = b_0 + b_1 * X \text{ or } \log(p/(1-p)) = b_0 + b_1 * X$$

where  $\log(p/(1-p))$  is called the *logit* of  $p$  or the *log odds-ratio*.

There is no error term here because the predicted value is not a response level; it is a probability distribution for a response level. For example, if the weather forecast predicts a 90% chance of rain, you don't say there's a mistake if it doesn't rain.

The accounting is done by summing the negative logarithms of the probabilities attributed by the model to the events that actually did occur. So if  $p$  is the precipitation probability from the weather model, then the score is  $-\log(p)$  if it rains, and  $-\log(1-p)$  if it doesn't. A weather forecast that is a perfect prediction comes up with a  $p$  of 1 when it rains ( $-\log(p)$  is zero if  $p$  is 1) and a  $p$  of zero when it doesn't rain ( $-\log(1-p)=0$  if  $p=0$ ). The perfect score is zero. No surprise  $-\log(p)=0$  means perfect predictions. If you attributed a probability of zero to an event that occurred, then the  $-\log$ -likelihood would be infinity, a pretty bad score for a forecaster.



So the inverse logit of the model  $b_0 + b_1 * X$  expresses the probability for each response level, and the estimates are found so as to maximize the likelihood. That is the same as minimizing the negative sum of logs of the probabilities attributed to the response levels that actually occurred for each observation.

You can graph the probability function as shown in **Figure 12.9**. The curve is just solving for  $p$  in the expression

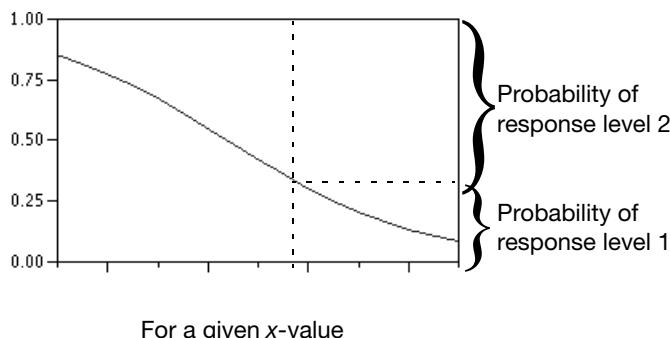
$$\log(p/(1-p)) = b_0 + b_1 * X$$

which is

$$p = 1/(1+\exp(-(b_0+b_1 * X)))$$

For a given value of  $X$ , this expression evaluates the probability of getting the first response. The probability for the second response is the remaining probability,  $1-p$ , because they must sum to 1.

**Figure 12.9** Logistic Regression Fits Probabilities of a Response Level



To fit the rain column by temperature and barometric pressure for the spring rain data:

- ❖ Select **Analyze > Fit Y by X** and assign the nominal column Rained to **Y, Response**, and the continuous columns Temp and Pressure to **X, Factor**.
- ❖ Click **OK**.

**Note:** By default, JMP models the probability that Rainy = Dry, because Dry is the first alphanumeric value. If, instead, we want to model the probability that it will be rainy, we would set **Rainy** as the Target Level in the Fit Y by X launch window.

The Fit Y by X platform produces a separate logistic regression for each predictor variable.

The cumulative probability plot on the left in **Figure 12.10** shows that the relationship with temperature is very weak. As the temperature ranges from 35 to 75, the probability of dry weather changes only from 0.73 to 0.66. The line of fit partitions the whole probability into the response categories. In this case, you read the probability of Dry directly on the vertical axis. The probability of Rained is the distance from the line to the top of the graph, which is 1 minus the axis reading. The weak relationship is evidenced by the very flat line of fit; the precipitation probability doesn't change much over the temperature range.

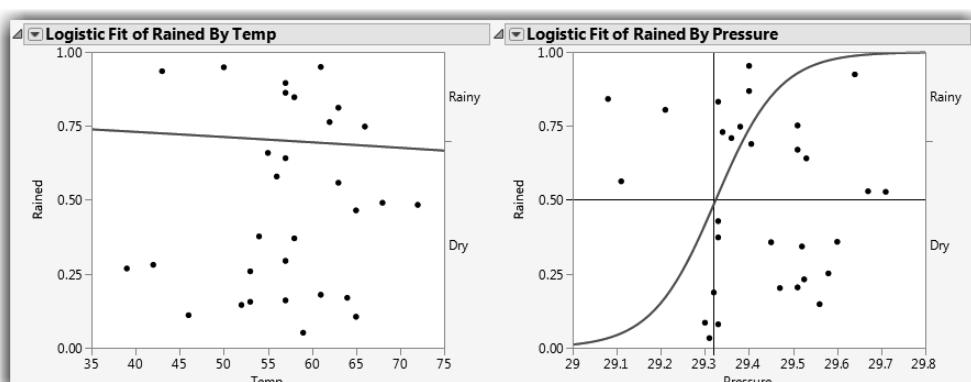
The plot on the right in **Figure 12.10** indicates a much stronger relationship with barometric pressure. When the pressure is 29.0 inches, the fitted probability of rain is near 100% (0 probability for Dry at the left of the graph). The curve crosses the 50% level at 29.32. (You can use the crosshair tool to see this.) At 29.8, the probability of rain drops to nearly zero (therefore, nearly 1.0 for Dry).

You can also add reference lines at the known X and Y values.

- ⓐ Double-click the Rained (Y) axis to open the Axis Settings window. Enter 0.5 as a reference line.
- ⓑ Double-click the Pressure (X) axis and enter 29.32 in the Axis Settings window.

When both reference lines appear, they intersect on the logistic curve as shown in the plot on the right in **Figure 12.10**.

**Figure 12.10** Cumulative Probability Plot for Discrete Rain Data



For the variable Temp, the Whole-Model Test table and the Parameter Estimates table reinforce the plot. The  $R^2$  measure of fit, which can be interpreted on a scale of 0 to 100%, is only 0.07% (shown as 0.0007 in **Figure 12.11**). A 100%  $R^2$  would indicate a model that predicted outcomes with certainty. The likelihood ratio chi-square is not at all significant. The coefficient on temperature is a very small—0.008. The parameter estimates can be unstable because they have high standard errors with respect to the estimates.

In contrast, the overall  $R^2$  measure of fit with barometric pressure is 34%. The likelihood ratio chi-square is highly significant, and the parameter coefficient for Pressure increased to 13.8 (**Figure 12.11**).

The conclusion is that if you want to predict whether it will be rainy, it doesn't help to know the temperature, but it does help to know the barometric pressure.

**Figure 12.11** Logistic Regression for Discrete Rain Data

Iterations				
Whole Model Test				
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.013334	1	0.026668	0.8703
Full	18.312595			
Reduced	18.325929			
RSquare (U)	0.0007			
AICc	41.0696			
BIC	43.4276			
Observations (or Sum Wgts)	30			
Fit Details				
Measure	Training	Definition		
Entropy RSquare	0.0007	1-Loglike(model)/Loglike(0)		
Generalized RSquare	0.0013	(1-(L(0)/L(model)))^(2/n)/(1-L(0)^(2/n))		
Mean -Log p	0.6104	$\sum -\log(p_{ij})/n$		
RMSE	0.4581	$\sqrt{\sum(y_{ij}-p_{ij})^2/n}$		
Mean Abs Dev	0.4196	$\sum  y_{ij}-p_{ij} /n$		
Misclassification Rate	0.3000	$\sum (p_{ij} \neq p_{Max})/n$		
N	30	n		
Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1.34073823	3.0620868	0.19	0.6615
Temp	-0.0086266	0.0529847	0.03	0.8707
Covariance of Estimates				

Iterations				
Whole Model Test				
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	6.250851	1	12.5017	0.0004*
Full	12.075078			
Reduced	18.325929			
RSquare (U)	0.3411			
AICc	28.5946			
BIC	30.9526			
Observations (or Sum Wgts)	30			
Fit Details				
Measure	Training	Definition		
Entropy RSquare	0.3411	1-Loglike(model)/Loglike(0)		
Generalized RSquare	0.4832	(1-(L(0)/L(model)))^(2/n)/(1-L(0)^(2/n))		
Mean -Log p	0.4025	$\sum -\log(p_{ij})/n$		
RMSE	0.3778	$\sqrt{\sum(y_{ij}-p_{ij})^2/n}$		
Mean Abs Dev	0.2739	$\sum  y_{ij}-p_{ij} /n$		
Misclassification Rate	0.3000	$\sum (p_{ij} \neq p_{Max})/n$		
N	30	n		
Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-405.3637	169.29563	5.73	0.0166*
Pressure	13.823423	7.7651473	5.75	0.0165*
Covariance of Estimates				

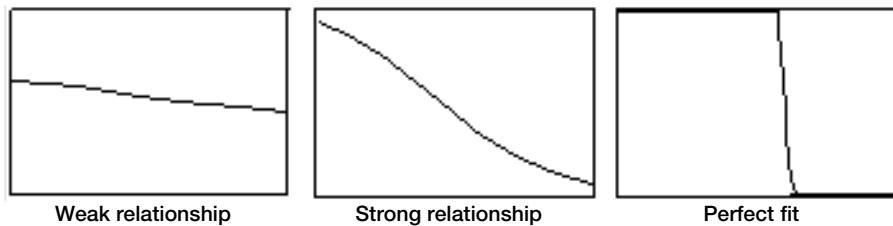
## Degrees of Fit

The illustrations in **Figure 12.12** summarize the degree of fit as shown by the cumulative logistic probability plot.

When the fit is weak, the parameter for the slope term (X factor) in the model is small, which gives a small slope to the line in the range of the data. A perfect fit means that before a certain value of X, all the responses are one level, and after that value of X, all the responses are another level. A strong model can fit almost

all of its probability on one event happening. A weak model has to bet conservatively with the background probability, less affected by the X factor's values.

**Figure 12.12** Strength of Fit in Logistic Regression



Note that when the fit is perfect, as shown on the rightmost graph of **Figure 12.12**, the slope of the logistic line approaches infinity. This means that the parameter estimates are also infinite. In practice, the estimates are allowed to inflate only until the likelihood converges and are marked as unstable by the computer program. You can still test hypotheses, because they are handled through the likelihood, rather than using the estimate's (theoretically infinite) values.

## A Discriminant Alternative

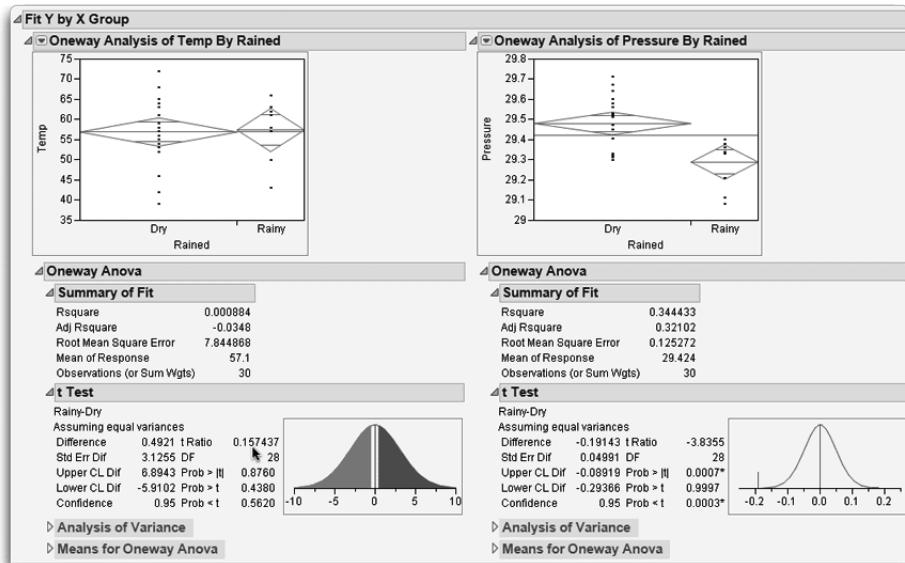
There is another way to think of the situation where the response is categorical and factor is continuous. You can reverse the roles of the Y and X and treat this problem as one of finding the distribution of temperature and pressure on rainy and dry days. Then, work backward to obtain prediction probabilities. This technique is called *discriminant analysis*. (Discriminant analysis is discussed in detail in Chapter 16.)

- ☛ For this example, select **Help > Sample Data Library** and open (or make active) Spring.jmp. (**Note:** If you open it from scratch, you need to add the Rained variable as detailed on page 321.)
- ☛ Select **Analyze > Fit Y by X** and assign Temp and Pressure to **Y, Response**, and Rained to **X, Factor**.
- ☛ Click **OK**.
- ☛ Hold down the Ctrl key and select **Means/Anova/Pooled t** from the red triangle menu next to Oneway to see the results in **Figure 12.13**.

You can quickly see that the difference between the relationships of temperature and pressure to raininess. However, the discriminant approach is a somewhat strange way to go about this example and has some problems:

- The standard analysis of variance assumes that the factor distributions are normal.
- Discriminant analysis works backward: First, in the weather example, you are trying to predict rain. But the ANOVA approach designates Rained as the independent variable, from which you can say something about the predictability of temperature and pressure. Then, you have to reverse-engineer your thinking to infer raininess from temperature and pressure.

**Figure 12.13** Temperature and Pressure as a Function of Raininess



## Inverse Prediction

If you want to know what value of the X regressor yields a certain probability, you can solve the equation,  $\log(p/(1-p)) = b_0 + b_1 * X$ , for X, given p. This is often done for toxicology situations, where the X-value for  $p = 50\%$  is called an *LD<sub>50</sub>* (Lethal Dose for 50%). Confidence intervals for these inverse predictions (called *fiducial confidence intervals*) can be obtained.

The Fit Model platform has an inverse prediction facility. Let's find the LD50 for pressure in the rain data—that is, the value of pressure that gives a 50% chance of rain.

- ~ Select **Analyze > Fit Model**.

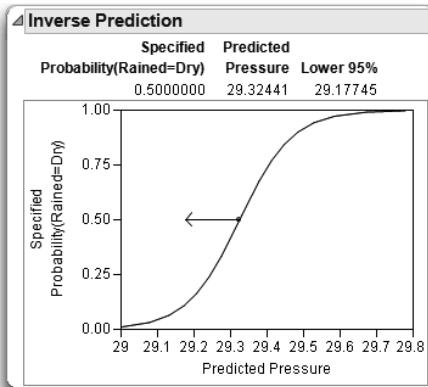
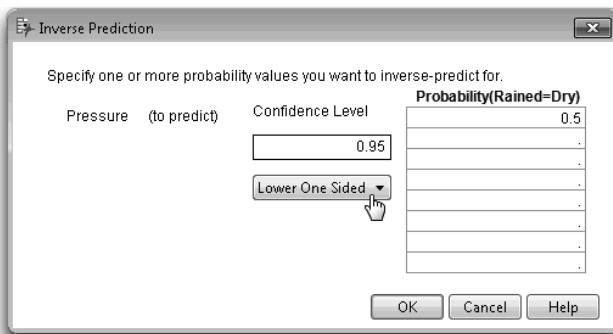
When the Model Specification window appears:

- ~ Assign Rained to **Y**.
- ~ Select Pressure and click **Add** to assign it as a model effect.
- ~ Click **Run**.
- ~ Select **Inverse Prediction** from the red triangle menu next to Nominal Logistic to see the Inverse Prediction window at the top in **Figure 12.14**.

The Probability and Confidence Level fields are editable, and you can select a **Two Sided**, **Lower One Sided**, or **Upper One Sided** prediction from the menu on the window. Enter any values of interest into the window and select the type of test you want. The result is an inverse probability for each probability request value that you entered at the specified alpha (confidence) level.

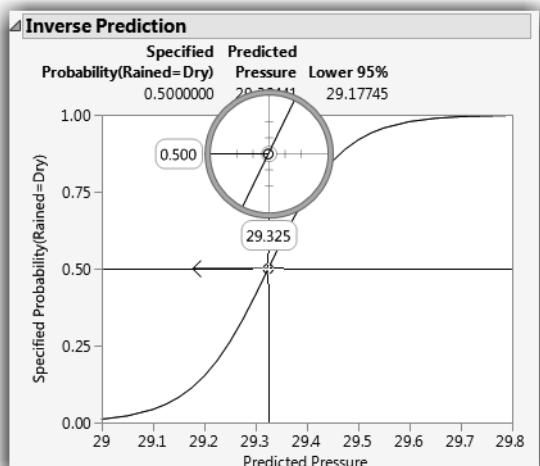
- ~ For this example, enter 0.5 as the first entry in the Probability column, as shown on **Figure 12.14**.
- ~ You know there is a relationship between lower pressure and raininess, so select the **Lower One Sided** test from the menu on the inverse prediction window, and then click **OK**.

The Inverse Prediction table and plot shown at the bottom of **Figure 12.14** are appended to the output. The inverse prediction computations say that there is a 50% chance of rain when the barometric pressure is 29.32.

**Figure 12.14** Inverse Prediction Window

To see this prediction clearly on the graph:

- ❖ Get the crosshair tool from the **Tools** menu or toolbar. Click in the middle of the circle in the crosshairs to see the predicted Pressure when the specified probability of rain is 0.50.
- ❖ Move the crosshair up or down the logistic curve to see other predicted values.

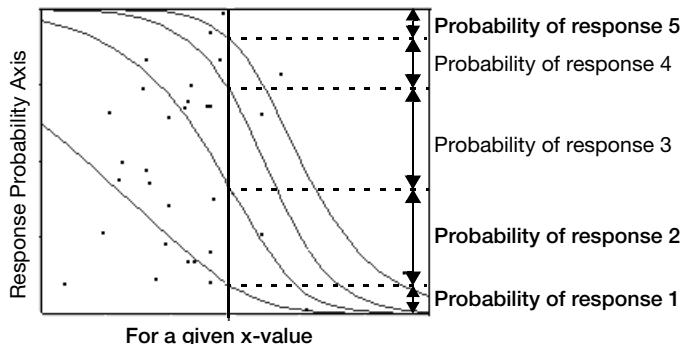


## Polytomous (Multinomial) Responses: More Than Two Levels

If there are more than two response categories, the response is said to be *polytomous*, and a generalized logistic model is used. For the curves to be very flexible, you have to fit a set of linear model parameters for each of  $r - 1$  response levels. The logistic curves are accumulated in such a way as to form a smooth partition of the probability space as a function of the regression model. The probabilities shown in **Figure 12.15** are the distances between the curves, which add up to 1.

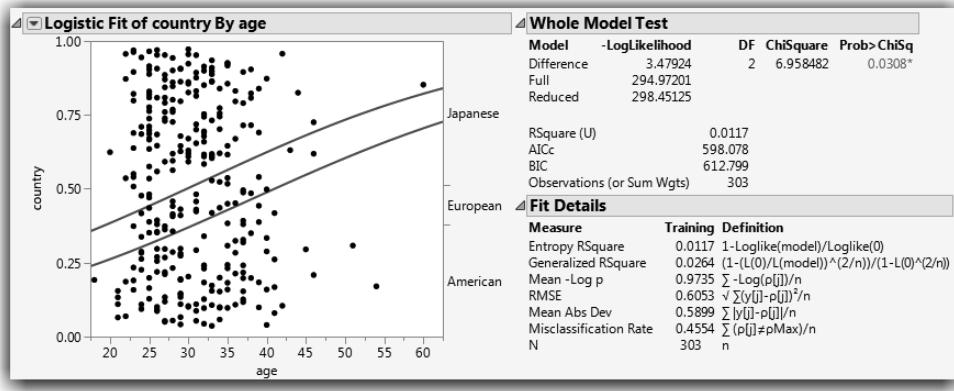
Where the curves are close together, the model is saying that the probability of a response level is very low. Where the curves separate widely, the fitted probabilities are large.

**Figure 12.15** Polytomous Logistic Regression with Five Response Levels



For example, consider fitting the probabilities for country with the Car Poll.jmp sample data as a smooth function of age. The result (**Figure 12.16**) shows the relationship where younger individuals tend to buy more Japanese cars and older individuals tend to buy more American cars. Note the double set of estimates (two curves) needed to describe three responses.

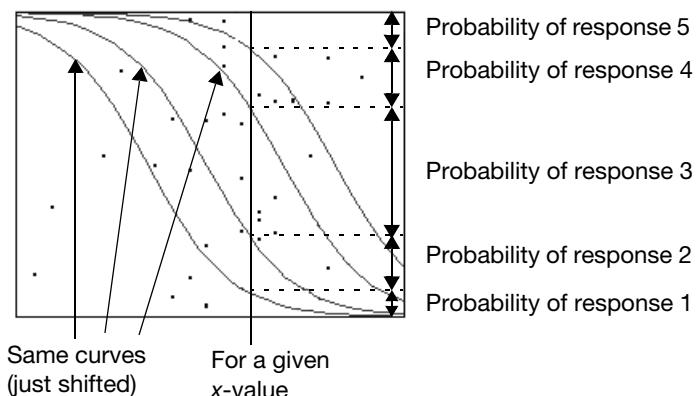
**Figure 12.16** Cumulative Probability Plot and Logistic Regression for Country by Age



## Ordinal Responses: Cumulative Ordinal Logistic Regression

In some cases, you don't need the full generality of multiple linear model parameter fits for the  $r - 1$  cases. However, you can assume that the logistic curves are the same, only shifted by a different amount over the response levels. This means that there is only one set of regression parameters on the factor, but  $r - 1$  intercepts for the  $r$  responses.

**Figure 12.17** Ordinal Logistic Regression Cumulative Probability Plot



The logistic curve is actually fitting the sum of the probabilities for the responses at or below it, so it is called a *cumulative* ordinal logistic regression. In the Spring.jmp sample data table, there is a column called SkyCover with values 0 to 10.

First, note that you don't need to treat the response as nominal because the data have a natural order. Also, in this example, there is not enough data to support the large number of parameters needed by a 10-response level nominal model. Instead, use a logistic model that fits SkyCover as an ordinal variable with the continuous variables Temp and Humid1: PM, the humidity at 1 p.m.

- ⓐ Change the modeling type of the SkyCover column to Ordinal by clicking the icon next to the column name in the Columns panel, located to the left of the data grid.
- ⓑ Select **Analyze > Fit Y by X** and assign SkyCover to **Y, Response**.
- ⓒ Assign Temp and Humid1:PM to **X, Factor** and click **OK**.

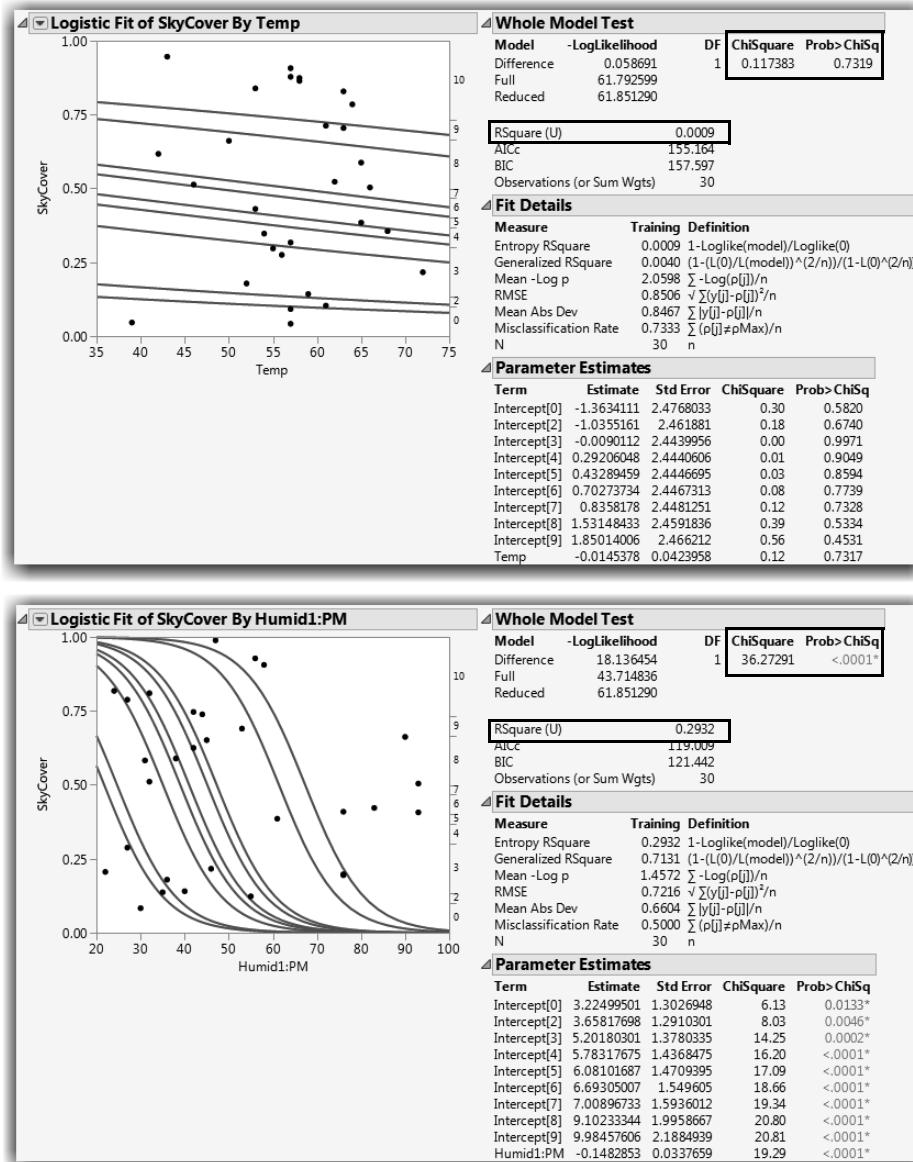


The top analysis in **Figure 12.18** indicates that the relationship of SkyCover to Temp is very weak, with an  $R^2$  of 0.09%, fairly flat lines, and a nonsignificant chi-square. The direction of the relation is that the higher sky covers are more likely with the higher temperatures.

The bottom analysis in **Figure 12.18** indicates that the relationship with humidity is quite strong. As the humidity approaches 70%, it predicts more than a 50% probability of a sky cover of 10. At 100% humidity, the sky cover will most likely be 10. The  $R^2$  is 29%, and the likelihood ratio chi-square is highly significant.

Note that no data occur for SkyCover = 1, so that value is not even in the model.

**Figure 12.18** Ordinal Logistic Regression for Ordinal Sky Cover with Temperature (top) and Humidity (bottom)



There is a useful alternative interpretation to this ordinal model. Suppose you assume that there is some continuous response with a random error component that the linear model is really fitting. But, for some reason, you can't observe the response directly. You are given a number that indicates which of  $r$  ordered

intervals contains the actual response, but you don't know how the intervals are defined. You assume that the error term follows a logistic distribution, which has a shape similar to a normal distribution. This case is identical to the ordinal cumulative logistic model. And the intercept terms are estimating the threshold points that define the intervals corresponding to the response categories.

Unlike the nominal logistic model, the ordinal cumulative logistic model is efficient to fit for hundreds of response levels. It can be used effectively for continuous responses when there are  $n$  unique response levels for  $n$  observations. In such a situation, there are  $n - 1$  intercept parameters constrained to be in order, and there is one parameter for each regressor.

## Surprise: Simpson's Paradox: Aggregate Data versus Grouped Data

Several statisticians have studied the “hot hand” phenomenon in basketball. The idea is that basketball players seem to have hot streaks, when they make the most of their shots, alternating with cold streaks when they shoot poorly. The Hothand.jmp sample data table contains the free throw shooting records for two Boston Celtics players (Larry Bird and Rick Robey) over the 1980-81 and 1981-82 seasons (Tversky and Gilovich, 1989).

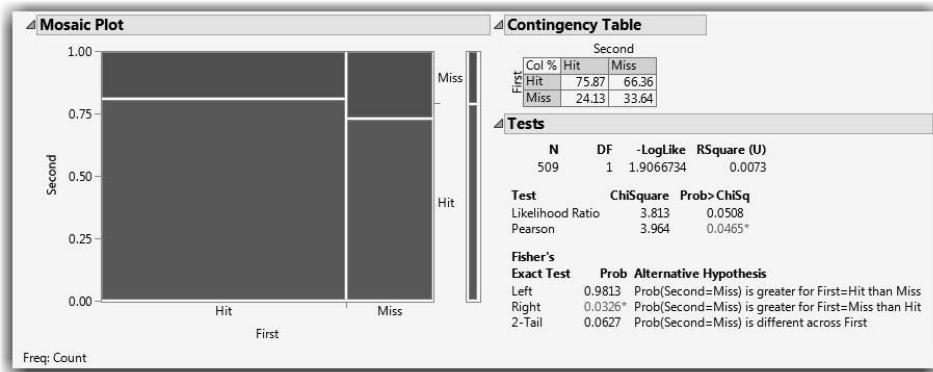
The null hypothesis is that two sequential free throw shots are independent. There are two directions in which they could be non-independent, the positive relationship (hot hand) and a negative relationship (cold hand).

The Hothand.jmp sample data have the columns First and Second (first shot and second shot) for the two players and a count variable. There are four possible shooting combinations: hit-hit, hit-miss, miss-hit, and miss-miss.

- ☛ Select **Help > Sample Data Library** and open Hothand.jmp.
- ☛ Select **Analyze > Fit Y by X**.
- ☛ Assign Second to **Y, Response**, First to **X, Factor**, Count to **Freq**, and then click **OK**.
- ☛ When the report appears, right-click in the contingency table and deselect all displayed numbers except **Col%**.

The results in **Figure 12.19** show that if the first shot is made, then the probability of making the second is 75.8%. If the first shot is missed, the probability of making the second is 24.1%. This tends to support the hot hand hypothesis. The two chi-square statistics are on the border of 0.05 significance.

**Figure 12.19** Crosstabs and Tests for Hot Hand Basketball Data

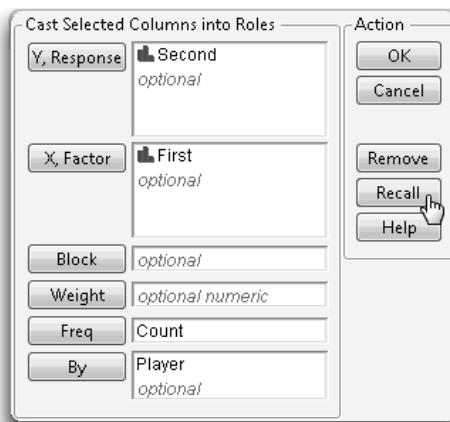


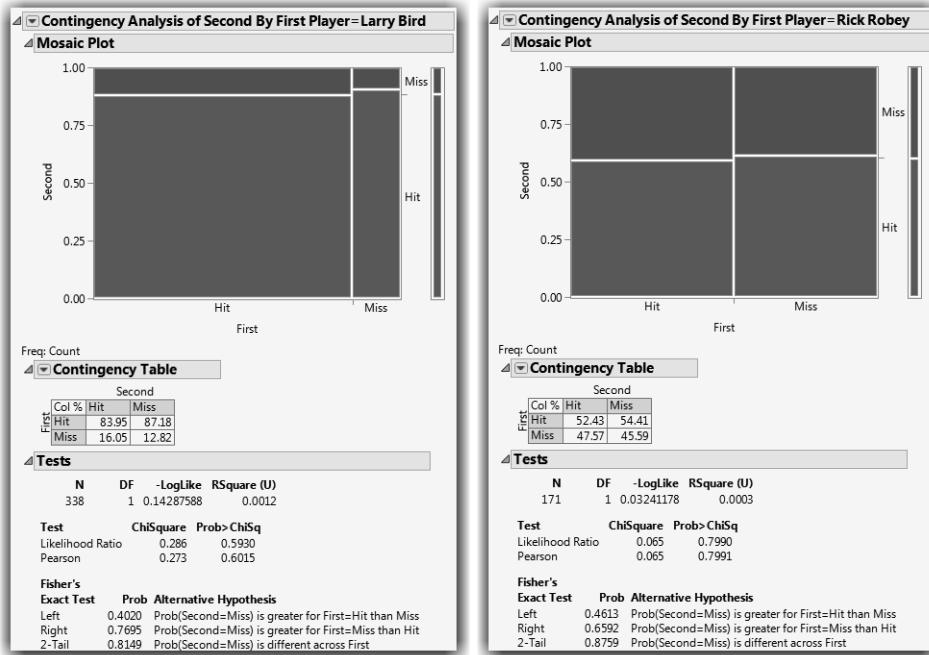
Does this analysis really confirm the hot hand phenomenon? A researcher (Wardrop 1995), looked at contingency tables for each player. You can do this using the **By** grouping variable.

**Note:** To repeat any analysis using the same variables, use the **Recall** button in the launch window.

- Ⓐ Again, select **Analyze > Fit Y by X** and assign Second to **Y, Response**, First as **X, Factor**, and Count to **Freq**.
- Ⓑ This time, assign Player to **By** and click **OK**.

The results for the two players are shown in **Figure 12.20**.



**Figure 12.20** Crosstabs and Tests for Grouped High-End Basketball Data

Contrary to the first result, both players shot better the second time after a miss than after a hit. So how can this be when the aggregate table gives the opposite results from both individual tables? This is an example of a phenomenon called *Simpson's paradox* (Simpson, 1951; Yule, 1903).

In this example, it is not hard to understand what happens if you think how the aggregated table works. If you see a hit on the first throw, the player is probably Larry Bird. Because he is usually more accurate, he will likely hit the second basket. If you see a miss on the first throw, the player is likely Rick Robey. So the second throw will be less likely to hit. The hot hand relationship is an artifact that the players are much different in scoring percentages generally and populate the aggregate unequally.

A better way to summarize the aggregate data, taking into account these background relationships, is to use a blocking technique called the *Cochran-Mantel-Haenszel* test.

- ❖ Click on the report for the analysis shown in **Figure 12.19** (without the *By* variable) and select **Cochran Mantel Haenszel** from the red triangle menu next to Contingency Analysis.

A grouping window appears that lists the variables in the data table.

- Select Player as the grouping variable in this window and click **OK**.

These results are more accurate because they are based on the grouping variable instead of the ungrouped data. Based on these  $p$ -values, the null hypothesis of independence (between the first and second shot) is not rejected.

**Figure 12.21** Crosstabs and Tests for Grouped Hothand Basketball Data

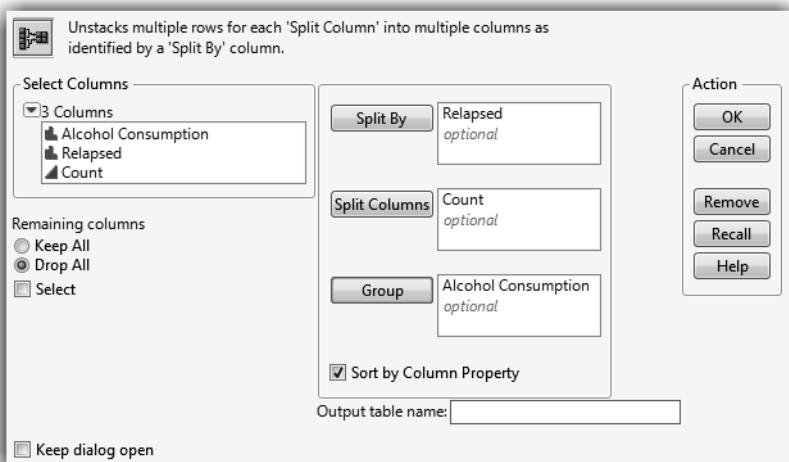
Contingency Analysis of Second By First					
Freq: Count					
Cochran-Mantel-Haenszel Tests					
Stratified by Player					
CMH Test	ChiSquare	DF	Prob>Chisq		
Correlation of Scores	0.2507	1	0.6166		
Row Score by Col Categories	0.2507	1	0.6166		
Col Score by Row Categories	0.2507	1	0.6166		
General Assoc. of Categories	0.2507	1	0.6166		
Frequency Counts					
Player=Larry Bird					
Second					
First	Count	Hit	Miss	Total	
	Hit	251	34	285	
	Miss	48	5	53	
	Total	299	39	338	
Player=Rick Robey					
Second					
First	Count	Hit	Miss	Total	
	Hit	54	37	91	
	Miss	49	31	80	
	Total	103	68	171	

## Generalized Linear Models

In recent years, *generalized linear models* have emerged as an alternative (and usually equivalent) approach to logistic models. There are two different formulations for generalized linear models that apply most often to categorical response situations: the *binomial* model and the *Poisson* model. To demonstrate both formulations, we use the Alcohol.jmp sample data table. **Figure 12.4** on page 316 shows that relapses into smoking are not independent of alcohol consumption, supported by a chi-square test ( $G^2 = 8.22$ ,  $p = 0.0041$ ).

The binomial approach is always applicable when there are only two response categories. To use it, we must first reorganize the data.

- ✓ Select **Help > Sample Data Library** and open Alcohol.jmp if it is not already open.
- ✓ Select **Tables > Split**.
- ✓ Assign variables as shown in **Figure 12.22**.

**Figure 12.22** Split Window for Alcohol Data

- ✓ Click **OK**.
- ✓ In the Alcohol Consumption table, add a new column that computes the total of the columns No + Yes, as shown in **Figure 12.23**.

**Figure 12.23** Final Data Table

	Alcohol Consumption	Relapsed	Count
1	Consumed	Yes	20
2	Consumed	No	13
3	Didn't Consume	Yes	48
4	Didn't Consume	No	96

Original Alcohol Consumption Table

	Alcohol Consumption	No	Yes	Total
1	Consumed	13	20	33
2	Didn't Consume	96	48	144

New table with Count column split into two columns by the “Yes” and “No” values of the Relapsed column

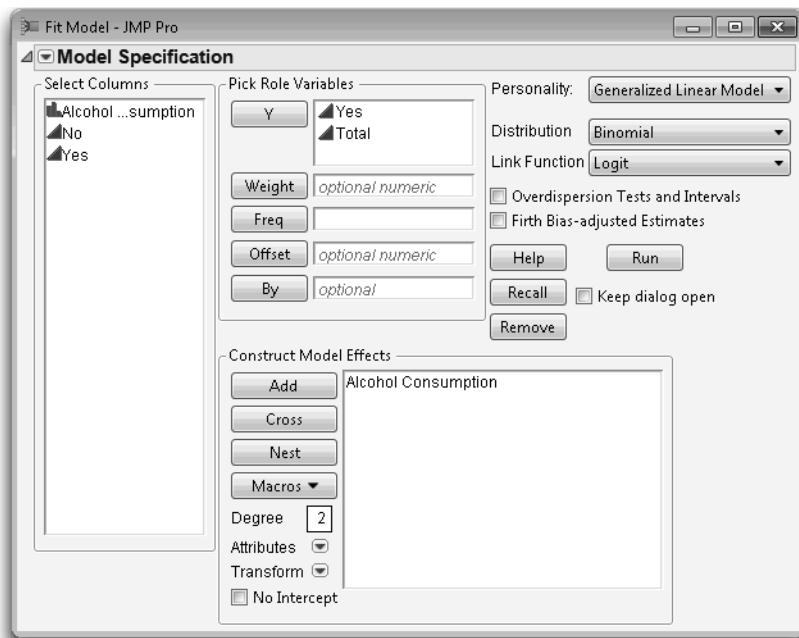
We're now ready to fit a model.

- ✓ Select **Analyze > Fit Model**.
- ✓ Change the **Personality** to **Generalized Linear Model**.

- ✓ Select the **Binomial** distribution.
- ✓ Assign Yes and Total to **Y**.
- ✓ Remove No from the Freq role.
- ✓ Assign Alcohol Consumption as the effect in the model.

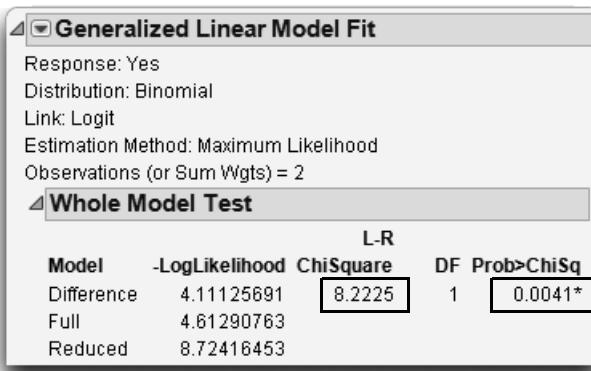
The Model Specification window should look like the one in **Figure 12.24**.

**Figure 12.24** Binomial Model Specification Window



- ✓ Click **Run**.

In the resulting report, notice that the chi-square test and the  $p$ -value agree with our previous findings in **Figure 12.4**.

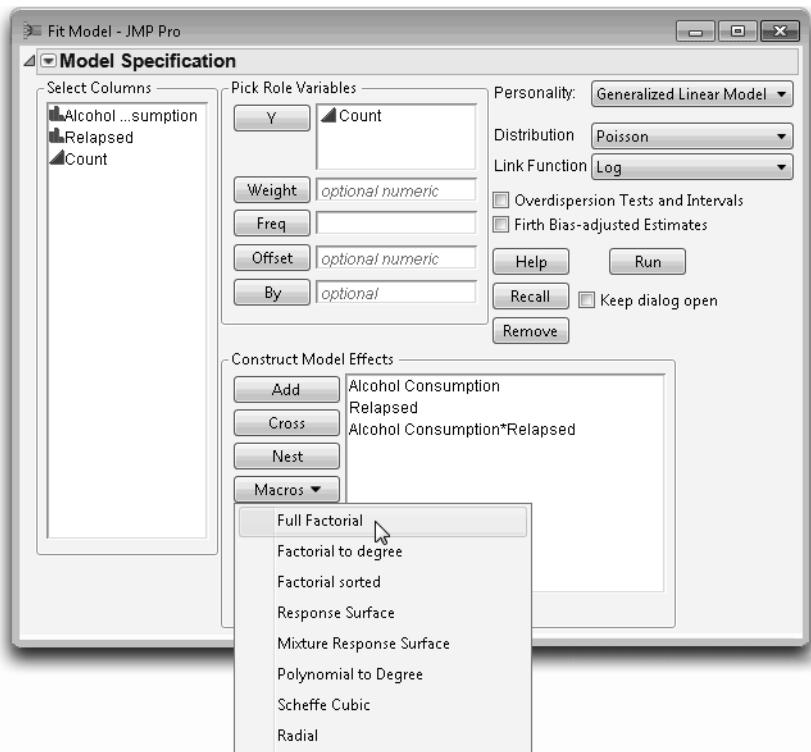
**Figure 12.25** Results of General Linear Model Using Binomial Distribution

The other approach, the Poisson model, uses the original sample data Alcohol.jmp. In this formulation, the count is portrayed as the response.

- ⓐ Select **Analyze > Fit Model**.
- ⓐ Select the **Generalized Linear Model** personality.
- ⓐ Select the **Poisson** distribution.
- ⓐ Assign Count to **Y**.
- ⓐ Remove Count as a Freq variable. (It was automatically assigned since it has a Freq role in the data table, but that's not how we're using it here.)
- ⓐ Select Alcohol Consumption and Relapsed in the Select Columns list on the Model Specification window.
- ⓐ Now select **Macros > Full Factorial** on the Model Specification window.

This adds both main effects and their crossed term.

**Figure 12.26** Model Specification Window for General Linear Model with Poisson Distribution.



☞ Click **Run**.

Examine the Effect Tests section of the report. There are three tests reported, but we ignore the main effect tests because they don't involve a response category. The interaction between the two variables is the one that we care about and, again, it is identical with previous results in **Figure 12.4** and **Figure 12.25**.

Effect Tests			
Source	L-R		
	DF	ChiSquare	Prob>ChiSq
Alcohol Consumption	1	65.938707	<.0001*
Relapsed	1	0.4298643	0.5121
Alcohol Consumption*Relapsed	1	8.2225138	0.0041*

Since all these methods produce equivalent results (in fact, identical test statistics), they are interchangeable. The choice of method can then be a matter of convenience, generalizability, or comfort of the researcher.

## Exercises

1. M.A. Chase and G.M. Dummer conducted a study in 1992 to determine what traits children regarded as important to popularity. Their data is represented in the sample data table Children's Popularity.jmp. Demographic information was recorded, as well as the rating given to four traits assessing their importance to popularity: Grades, Sports, Looks, and Money.
  - (a) Is there a difference based on gender on the importance given to making good grades?
  - (b) Is there a difference based on gender on the importance of excelling in sports?
  - (c) Is there a difference based on gender on the importance of good looks or on having money?
  - (d) Is there a difference between Rural, Suburban, and Urban students on rating these four traits?
2. One of the concerns of textile manufacturers is the absorbency of materials that clothes are made out of. Clothes that can quickly absorb sweat (such as cotton) are often thought of as more comfortable than those that cannot (such as polyester). To increase absorbency, material is often treated with chemicals. In this fictional experiment, several swatches of denim were treated with two acids to increase their absorbency. They were then assessed to determine whether their absorbency had increased or not. The investigator wanted to determine whether there is a difference in absorbency change for the two acids under consideration. The results are presented in the following table:

		<b>Acid</b>	
		<b>A</b>	<b>B</b>
<b>Absorbency</b>	<b>Increased</b>	54	40
	<b>Did Not Increase</b>	25	40

Does the researcher have evidence to say that there is a difference in absorbency between the two acids?

3. The taste of cheese can be affected by the additives that it contains. McCullagh and Nelder (1983) report a study (conducted by Dr. Graeme Newell) to determine the effects of four different additives on the taste of a cheese. The tasters responded by rating the taste of each cheese on a scale of 1 to 9. The results are in the sample data table *Cheese.jmp*.
  - (a) Produce a mosaic plot to examine the difference in taste among the four cheese additives.
  - (b) Do the statistical tests say that the difference amongst the additives is significant?
  - (c) Conduct a correspondence analysis to determine which of the four additives results in the best-tasting cheese.
4. The sample data table *Titanic Passengers.jmp* contains information about the Passengers of the RMS Titanic. Use JMP to answer the following questions:
  - (a) How many passengers were on the ship? How many survived?
  - (b) How many passengers were male? Female?
  - (c) How many passengers were in each class?
  - (d) Test the hypothesis that there is no difference in the survival rate among the passenger classes.
  - (e) Test the hypothesis that there is no difference in the survival rate between males and females.
  - (f) Use logistic regression to determine whether age is related to survival rate.
5. Do dolphins alter their behavior based on the time of day? To study this phenomenon, a marine biologist in Denmark gathered the data presented in the sample data table *Dolphins.jmp* (Rasmussen, 1998). The variables represent different activities observed in groups of dolphins, with the *Groups* variable showing the number of groups observed.
  - (a) Do these data show evidence that dolphins exhibit different behaviors during different times of day?
  - (b) There is a caution displayed with the chi-square statistic. Should you reject the results of this analysis based on the warning?





# 13

## Multiple Regression

### Overview

Multiple regression is the technique of fitting or predicting a response variable from a linear combination of several other variables. The fitting principle is least squares, the same as with simple linear regression.

Many regression concepts were introduced in previous chapters. This chapter concentrates on showing some new concepts not encountered in simple regression: the point-by-point picture of a hypothesis test with the leverage plot; collinearity (the situation in which one regressor variable is closely related to another); and the case of exact linear dependencies.

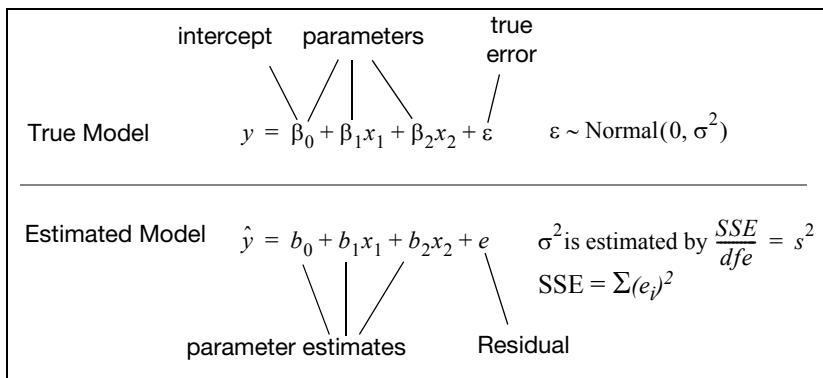
## Chapter Contents

Overview .....	345
Parts of a Regression Model .....	347
Regression Definitions.....	347
A Multiple Regression Example.....	348
Residuals and Predicted Values .....	351
The Analysis of Variance Table .....	354
The Whole Model F-Test .....	354
Whole-Model Leverage Plot .....	355
Details on Effect Tests.....	356
Effect Leverage Plots.....	356
Collinearity .....	358
Exact Collinearity, Singularity, and Linear Dependency .....	362
The Longley Data: An Example of Collinearity .....	364
The Case of the Hidden Leverage Point .....	366
Mining Data with Stepwise Regression .....	369
Exercises.....	373

## Parts of a Regression Model

Linear regression models are the sum of the products of coefficient parameters and factors. In addition, linear models for continuous responses are usually specified with a normally distributed error term. The parameters are chosen such that their values minimize the sum of squared residuals. This technique is called estimation by *least squares*.

**Figure 13.1** Parts of a Linear Model



Note in **Figure 13.1** the differences in notation between the assumed true model with unknown parameters and the estimated model.

## Regression Definitions

### response, Y

The *response* (or *dependent*) *variable* is the one you want to predict. Its estimates are the dependent variable,  $\hat{y}$ , in the regression model.

### regressors, Xs

The *regressors* (*xs*) in the regression model are also called *independent variables*, *predictors*, *factors*, *explanatory variables*, and other discipline-specific terms. The regression model uses a linear combination of these effects to fit the response value.

### coefficients, parameters

The fitting technique produces estimates of the parameters, which are the coefficients for the linear combination that defines the regression model.

**intercept term**

Most models have intercept terms to fit a constant in a linear equation. This is equivalent to having an  $x$ -variable that always has the value 1. The intercept is meaningful by itself only if it is meaningful to know the predicted value where all the regressors are zero. However, the intercept plays a strong role in testing the rest of the model, because it represents the mean if all the other parameters are zero.

**error, residual**

If the fit isn't perfect, then there is error left over. *Error* is the difference between an actual value and its predicted value. When speaking of true parameters, this difference is called *error*. When using estimated parameters, this difference is called a *residual*.

## A Multiple Regression Example

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

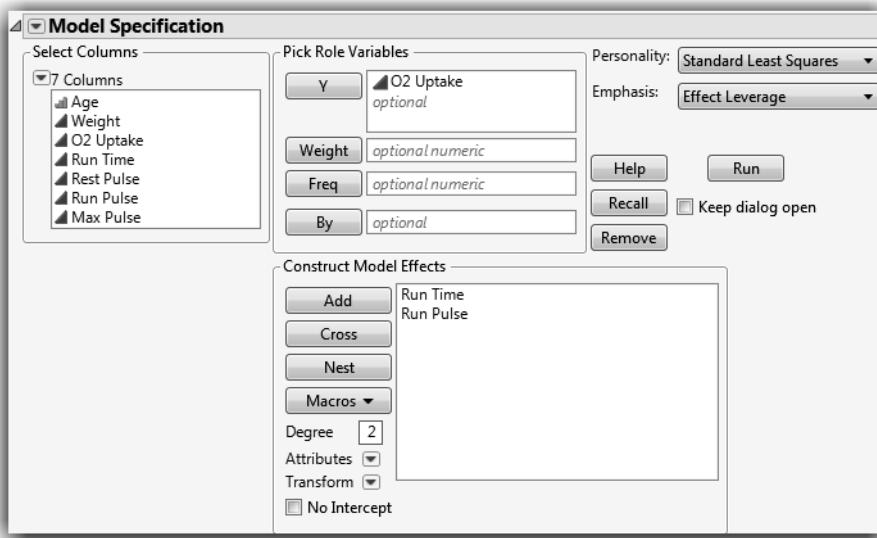
To develop such a formula, running time and pulse measurements were taken for 31 participants who each ran 1.5 miles. Their oxygen uptake, pulses, times, and other descriptive information was recorded. (Rawlings 1988, data courtesy of A.C. Linnerud). **Figure 13.2** shows a partial listing of the data, with variables Age, Weight, O<sub>2</sub> Uptake (the response measure), Run Time, Rest Pulse, Run Pulse, and Max Pulse.

**Figure 13.2** The Oxygen Uptake Data Table

	Age	Weight	O2 Uptake	Run Time	Rest Pulse	Run Pulse	Max Pulse
1	38	81.87	60.055	8.63	48	170	186
2	38	89.02	49.874	9.22	55	178	180
3	40	75.07	45.313	10.07	62	185	185
4	40	75.98	45.681	11.95	70	176	180
5	42	68.15	59.571	8.17	40	166	172
6	44	85.84	54.297	8.65	45	156	184
7	43	81.19	49.091	10.85	64	162	170
8	44	73.03	50.541	10.13	45	168	168
9	44	89.47	44.809	11.37	62	178	182
10	44	81.42	39.442	13.08	63	174	176

Investigate Run Time and Run Pulse as predictors of oxygen uptake (O2 Uptake).

- ☞ Select **Help > Sample Data Library** and open Linnerud.jmp, shown in **Figure 13.2**.
  - ☞ Select **Analyze > Fit Model** to see the Model Specification window.
  - ☞ Assign O2 Uptake to **Y** to make it the response (Y) variable.
  - ☞ Select Run Time and Run Pulse, and click **Add** to make them model effects.
- Your window should look like the one in **Figure 13.3**.
- ☞ Click **Run** to launch the platform.

**Figure 13.3** Model Specification Window for Multiple Regression

This produces a report with several graphs and tables of statistical output. The tables shown in **Figure 13.4** report on the regression fit.

- The Summary of Fit table shows that the model accounted for 76% of the variation around the mean ( $R^2$ , reported as RSquare). The remaining residual error is estimated to have a standard deviation of 2.69 (Root Mean Square Error).
- The Parameter Estimates table shows Run Time to be highly significant ( $p < 0.0001$ ), but Run Pulse is not significant ( $p = 0.1567$ ). Using these parameter estimates, the prediction equation is

$$\text{O2 Uptake} = 93.089 - 3.14 \text{ Run Time} - 0.0735 \text{ Run Pulse}$$

- The Effect Summary and Effect Test tables show details of how each regressor contributes to the fit.

**Figure 13.4** Statistical Tables for Multiple Regression Example

The screenshot displays several statistical tables from a JMP software interface:

- Effect Summary:**

Source	LogWorth	PValue
Run Time	8.424	0.0000
Run Pulse	0.805	0.15673
- Summary of Fit:**

RSquare	0.761424
RSquare Adj	0.744383
Root Mean Square Error	2.693374
Mean of Response	47.37581
Observations (or Sum Wgts)	31
- Analysis of Variance:**

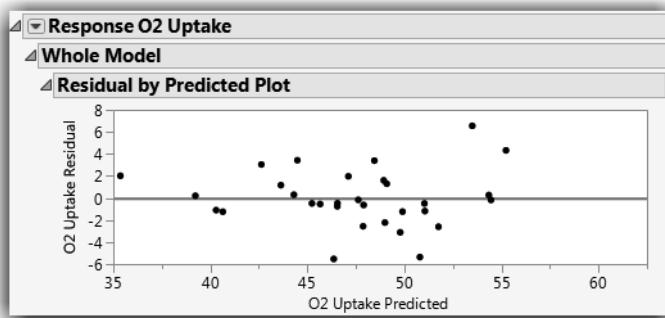
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	648.26218	324.131	44.6815
Error	28	203.11936	7.254	Prob > F
C. Total	30	851.38154		<.0001*
- Parameter Estimates:**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	93.088766	8.248823	11.29	<.0001*
Run Time	-3.140188	0.373265	-8.41	<.0001*
Run Pulse	-0.073509	0.050514	-1.46	0.1567
- Effect Tests:**

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Run Time	1	1	513.41745	70.7746	<.0001*
Run Pulse	1	1	15.36208	2.1177	0.1567

## Residuals and Predicted Values

The residual is the difference between the actual response and the response predicted by the model. The residuals represent the error in the model. Points that don't fit the model have large residuals. It is helpful to look at a plot of the residuals versus the predicted values, so JMP automatically produces a residual plot as shown in **Figure 13.5**.

**Figure 13.5** Residual Plot for Multiple Regression Example

Other residual plots are available from Row Diagnostics in the top red triangle menu.

To further explore model errors, you can save residuals as a column in the data table.

- ☞ Select **Save Columns > Residuals** from the red triangle menu next to Response.

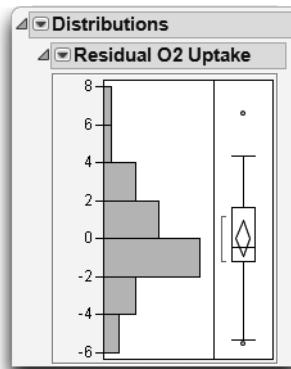
The result is a new column in the data table called Residual O2 Uptake, which lists the residual for each response point. We can now use other built-in JMP tools to examine these residuals,

- ☞ Select **Analyze > Distribution** and select the column of residuals to see the distribution of the residuals, as shown here.

Many researchers do this routinely to verify that the residuals are not so non-normal as to warrant concern about violating normality assumptions.

You might also want to store the prediction formula from the multiple regression.

- ☞ Select **Save Columns > Prediction Formula** from the red triangle menu next to Response O2 Uptake to create a new column in the data table called Pred Formula O2 Uptake. Its values are the calculated predicted values from the model.



- >To see the formula used to generate the values in the column, right-click at the top of the Pred Formula O2 Uptake column in the data table and select **Formula**. You can also click on the plus sign next to the column name in the columns panel of the data table. The Formula Editor window appears and displays the formula shown here.

$$93.08876614 + -3.14018757 \cdot \text{Run Time} + -0.073509488 \cdot \text{Run Pulse}$$

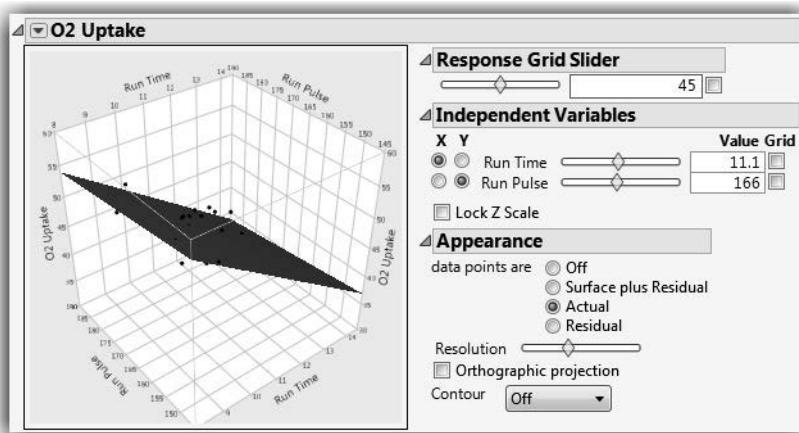
This formula defines a plane for O2 Uptake as a function of Run Time and Run Pulse. The formula stays in the column and is evaluated whenever new rows are added, or when values are changed for variables used in the expression. You can cut-and-paste or drag this formula into other JMP data tables columns.

To visualize the fitted plane, return to the fit least squares report for O2 Uptake.

- Select **Factor Profiling > Surface Profiler** from the red triangle menu next to Response O2 Uptake.
- To see the points on the plane, open the Appearance outline and click the Actual radio button.

You should see the results in **Figure 13.6**. Click and drag in the plot to change its orientation and explore the position of the points relative to the plane.

**Figure 13.6** Fitted Plane for O2 Uptake Multiple Regression Model



**Note:** The Prediction Profiler, which is introduced in “Visualizing the Results with the Prediction Profiler” on page 437, is another tool for exploring the prediction formula. Select **Factor Profiling > Profiler** from the red triangle menu next to Response to access the Prediction Profiler.

## The Analysis of Variance Table

The Analysis of Variance table (shown here) lists the sums of squares and degrees of freedom used to form the whole model test:

Analysis of Variance				
Source	DF	Sum of Squares		
		Mean Square	F Ratio	Prob > F
Model	2	648.26218	324.131	44.6815
Error	28	203.11936	7.254	<.0001*
C. Total	30	851.38154		

- The **Error Sum of Squares (SSE)** is 203.1. It is the sum of squared residuals after fitting the full model.
- The **C. Total Sum of Squares** is 851.4. It is the sum of squared residuals if you removed all the regression effects except for the intercept and therefore fit only the mean.
- The **Model Sum of Squares** is 648.3. It is the sum of squares caused by the regression effects, which measures how much variation is accounted for by the regressors. It is the difference between the Total Sum of Squares and the Error Sum of Squares.

The Error, C. Total, and Model sums of squares are the ingredients needed to test the whole-model null hypothesis that all the parameters in the model are zero except for the intercept (the simple mean model).

## The Whole Model *F*-Test

To conduct the whole model *F*-test:

1. Divide the Model Sum of Squares (648.3 in this example) by the number of parameters in the model excluding the intercept. That divisor (2 in this case) is found in the column labeled DF (Degrees of Freedom). The result is displayed as the *Model Mean Square in the ANOVA table*.
2. Divide the Error Sum of Squares (203.12 in this example) by the associated degrees of freedom, 28, giving the *Error Mean Square*.
3. Compute the *F*-ratio as the Model Mean Square divided by the Error Mean Square.

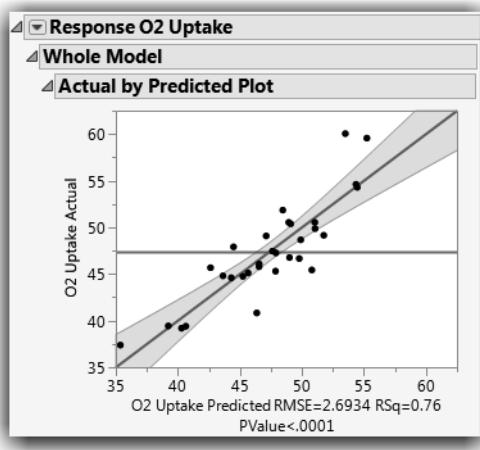
The significance level, or  $p$ -value, for this ratio is then calculated for the proper degrees of freedom (2 used in the numerator and 28 used in the denominator). The  $F$ -ratio, 44.6815, in the analysis of variance table shown above, is highly significant ( $p < 0.0001$ ). This lets us reject the null hypothesis and indicates that the model does fit better than a simple mean fit.

**Note:** To get a better sense of how extreme this  $F$ -ratio is, use the Distribution Calculator. The calculator is in **Help > Sample Data > Teaching Scripts > Interactive Teaching Modules**.

## Whole-Model Leverage Plot

There is a good way to view this whole-model hypothesis graphically using a scatterplot of actual response values against the predicted values. The plot below shows the actual values versus predicted values for this aerobic exercise example.

**Figure 13.7** Leverage Plot for the Whole Model Fit



A  $45^\circ$  line of fit from the origin shows where the actual response and predicted response are equal. The vertical distance from a point to the  $45^\circ$  line of fit is the difference of the actual and the predicted values—the *residual error*. The mean is shown by the horizontal dashed line. The distance from a point to the horizontal line at the mean is what the residual would be if you removed all the effects from the model.

A plot that compares residuals from the two models in this way is called a *leverage plot*. The idea is to get a feel for how much better the sloped line fits than the horizontal line.

Superimposed on the plot are the confidence curves representing the 0.05-level whole-model hypothesis. If the confidence curves do not contain the horizontal line, the whole-model  $F$ -test is significant, which means that the model predicts better than the mean alone.

The leverage plot shown in **Figure 13.7** is for the whole model, which includes both Run Time and Run Pulse.

## Details on Effect Tests

You can explore the significance of an effect in a model by looking at the distribution of the estimate. You can also look at the contribution of the effect to the model.

- To look at the distribution of the estimate, first compute its standard error. The standard error can be used either to construct confidence intervals for the parameter or to perform a  $t$ -test on whether the parameter is equal to some value (usually zero). The  $t$ -tests are given in the Parameter Estimates table.

**Note:** To request confidence intervals, right-click anywhere on the Parameter Estimates Table and select **Columns > Lower 95%** and **Upper 95%**.

- If you take an effect out of the model, then the error sum of squares increases. That difference in sums of squares (with the effect included and excluded) can be used to construct an  $F$ -test on whether the contribution of the effect to the model is significant. The  $F$ -tests are given in the Effect Tests table.

**Note:** You can see the change in sum of squares by selecting an effect in the Effect Summary table and clicking **Remove**. (Click **Undo** to add the term back.)

As we have seen,  $F$ -tests and  $t$ -tests are equivalent. The square of the  $t$ -value in the Parameter Estimates table is the same as the  $F$ -statistic in the Effect Test table. For example, the square of the  $t$ -ratio (-8.41) for Run Time is 70.77, which is the  $F$ -ratio for Run Time.

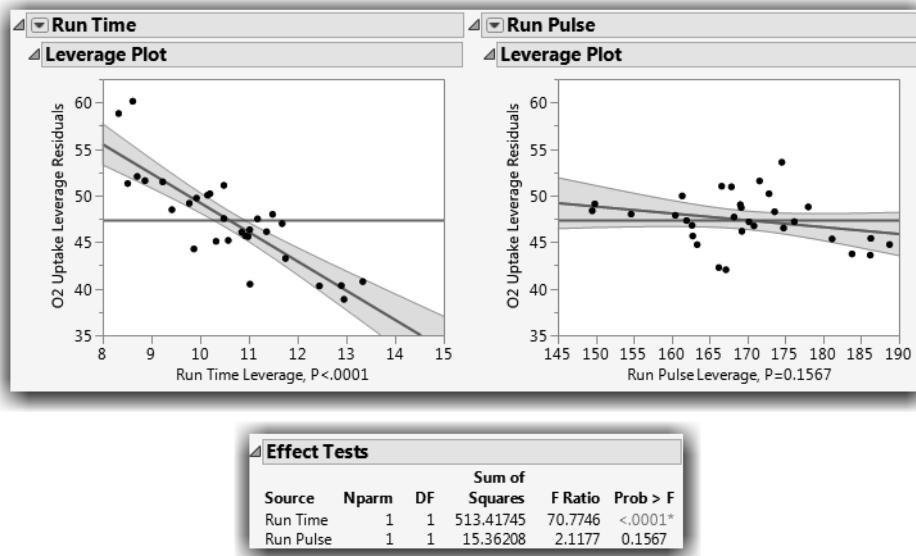
## Effect Leverage Plots

Scroll to the right on the regression report to see plots detailing how each effect contributes to the model fit. The plots for the effect tests are *also* called leverage plots, although they are not the same as the leverage plots encountered in the whole-model test. The *effect leverage* plots (see **Figure 13.8**) show how each effect contributes to the fit after all the other effects have been included in the model. A leverage plot for a hypothesis test (an effect) is any plot with horizontal and sloped reference lines and points laid out having the following two properties:

- The distance from each point to the sloped line measures the residual for the full model. The sums of the squares of these residuals form the error sum of squares (SSE).
- The distance from each point to the horizontal line measures the residual for the restricted model without the effect. The sums of the squares of these residuals form the SSE for the constrained model (the model without the effect).

In this way, it is easy to see point by point how the sum of squares for the effect is formed. The difference in sums of squares of the two residual distances forms the numerator for the *F*-test for the effect.

**Figure 13.8** Leverage Plots for Significant Effect and Nonsignificant Effect



The leverage plot for an effect is interpreted in the same way as a simple regression plot. In fact, JMP superimposes a type of 95% confidence curve on the sloped line that represents the full model. If the line is sloped significantly away from horizontal, then the confidence curves don't surround the horizontal line that represents the constrained model, and the effect is significant. Alternatively, when the confidence curves enclose the horizontal line, the effect is not significant at the 0.05 level.

The leverage plots in **Figure 13.8** show that Run Time is significant and Run Pulse is not. You can see the significance by how the points support (or don't support) the line of fit in the plot and by whether the confidence curves for the line cross the horizontal line.

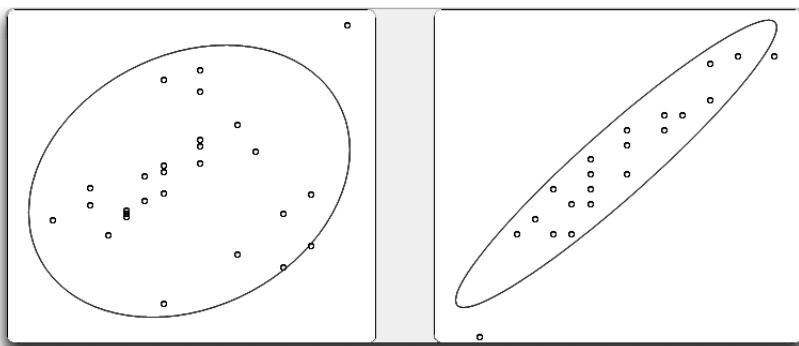
There is a leverage plot for any type of effect or set of effects in a model, or for any linear hypothesis. Leverage plots in the special case of single regressors are also known by the terms *partial plot*, *partial regression leverage plot*, and *added variable plot*.

## Collinearity

Sometimes with a regression analysis, there is a close linear relationship between two or more effects. These two regressors are said to have a *collinearity* problem. It is a problem because the regression points do not occupy all the directions of the regression space very well; the fitting plane is not well supported in certain directions. The fit is weak in those directions, and the estimates become unstable, which means that they are sensitive to small changes in the data.

In the plot on the left in **Figure 13.9**, there is little collinearity, and the points are distributed throughout the region. In the plot on the right, the two regressors are collinear, and the points are constrained to a narrow band.

**Figure 13.9** Illustration of Noncollinearity (Left) and Collinearity (Right)



In the statistical results, this phenomenon translates into large standard errors for the parameter estimates and potentially large values for the parameter estimates themselves. This occurs because a small random change in the more extreme values can have a huge effect on the slope of the corresponding fitting plane.

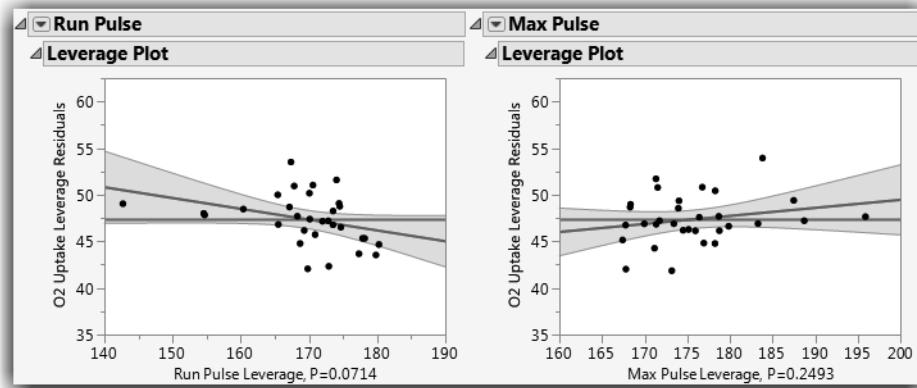
An indication of collinearity in leverage plots is when the points tend to collapse horizontally toward the center of the plot.

To see an example of collinearity, consider the aerobic exercise example with the correlated effects Max Pulse and Run Pulse:

- ❖ With the Linnerud exercise table active, select **Analyze > Fit Model** and click **Recall** (or click on the existing Model Specification window if it is still open).
- ❖ Add Max Pulse as a model effect and then click **Run**.

In the new analysis report, scroll to the Run Pulse and Max Pulse leverage plots. Note in **Figure 13.10** that Run Pulse is very near the boundary of 0.05 significance. Therefore, the confidence curves almost line up along the horizontal line, without actually crossing it.

**Figure 13.10** Leverage Plots for Effects in Model



Now, as an example, let's change the relationship between these two effects by changing a few values to cause collinearity.

- ❖ Select **Analyze > Fit Y by X**, and assign Max Pulse to **Y, Response** and Run Pulse to **X, Factor**.
- ❖ Click **OK**.

This produces a scatterplot showing the bivariate relationship.

- ❖ Select **Density Ellipse > 0.90** from the red triangle menu next to Bivariate Fit.

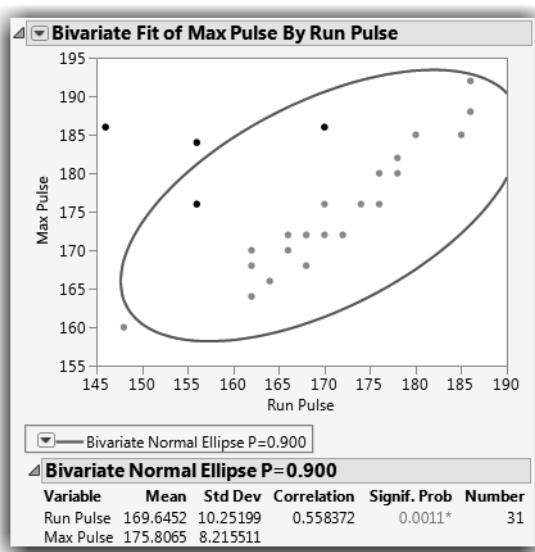
You should now see the scatterplot with the ellipse shown here. The **Density Ellipse** command also generates the Correlation table beneath the plot, which shows the current correlation to be 0.558. (Click on the gray disclosure icon to display.)

The variables don't appear overly collinear, because there is some scatter in the points. However, it appears that if four points are excluded, the correlation would increase dramatically. To increase the correlation, we will exclude the two points outside the ellipse and the two points just inside the ellipse, in the upper left corner of the plot.

To see the result, exclude these points and rerun the analysis.

- ⓐ Use Shift-click to highlight the points described above in the scatterplot.
- ⓑ With these points highlighted, select **Rows > Label/Unlabel** to identify them.
- ⓒ Select **Rows > Exclude** while the rows are highlighted.

Notice in the data table that these points are now marked with a label tag ( ) and the do not symbol ( ).



- Again, select a 0.90 **Density Ellipse** from the top red triangle menu.

Now the ellipse and the Correlation table show the relationship without the excluded points. The new correlation is 0.95, and the ellipse is much narrower, as shown here.

Now, run the regression model again to see the effect of excluding these points that created collinearity:

- Click **Run** again in the Model Specification dialog (with the same model as before).

or

- If the Model Specification dialog has been closed, go to **Analyze > Fit Model** and click **Recall**.

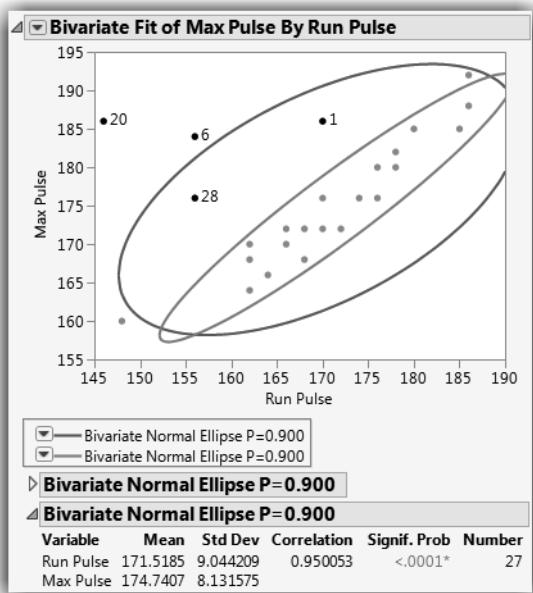
The model should have O2 as the response (Y) variable and Run Pulse, Run Time, and Max Pulse as the model effects.

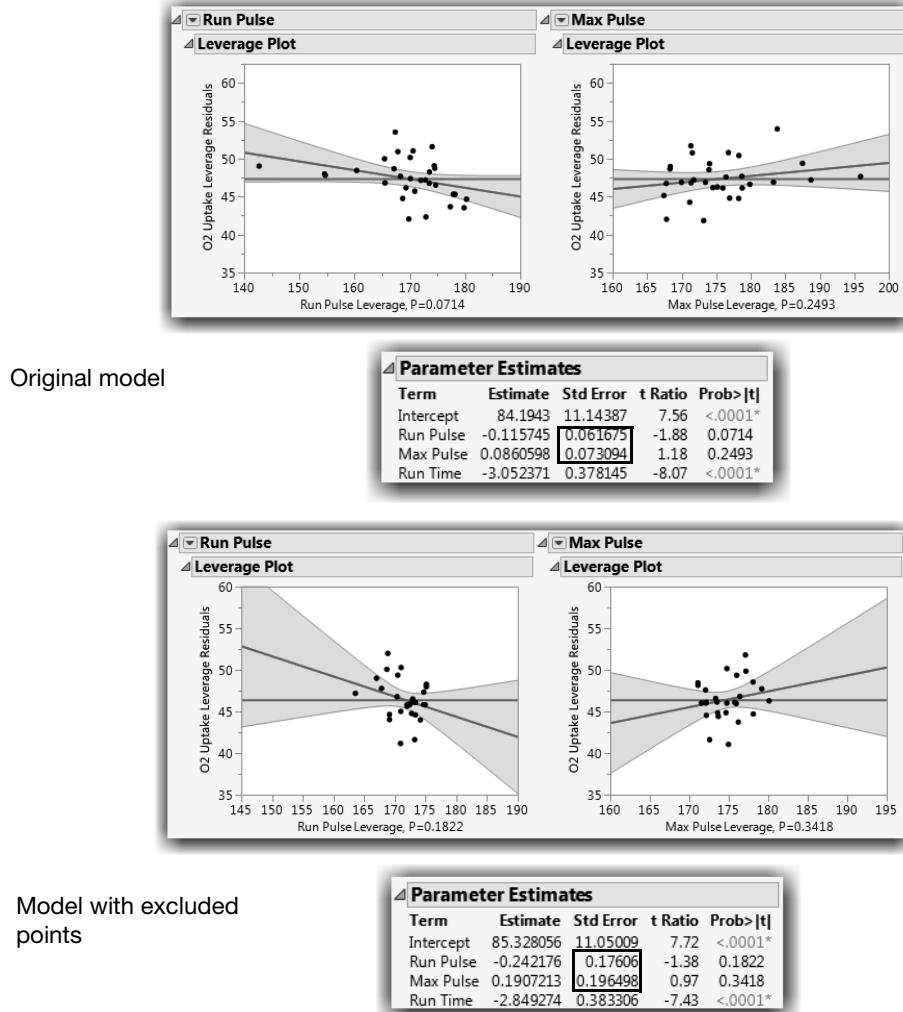
- Click **Run**.

Examine both the Parameter Estimates table and the leverage plots for Run Pulse and Max Pulse, comparing them with the previous report (see **Figure 13.11**).

The parameter estimates and standard errors for the last two regressors have nearly tripled in size.

The leverage plots now have confidence curves that flare out because the points themselves collapse toward the middle. When a regressor suffers collinearity, the other variables have already absorbed much of that variable's variation, and there is less left to help predict the response. Another way of thinking about this is that the points have less leverage on the hypothesis. Points that are far out horizontally are said to have high leverage on the hypothesis test; points in the center have little leverage.



**Figure 13.11** Comparison of Model Fits

## Exact Collinearity, Singularity, and Linear Dependency

Here we construct a variable to show what happens when there is an exact linear relationship, the extreme of collinearity, among the effects.

- ⌚ Return to the Linnerud.jmp sample data table and select **File > Revert** (or **File > Revert to Saved** on Macintosh) to reopen the data table in its original state.
- ⌚ Change the modeling type for Age from Ordinal to Continuous by selecting the continuous icon next to Age in the Columns panel.

- ☞ Select **Cols > New Columns** to add a new column (call it Run-Rest) to the data table.
- ☞ Use the Formula Editor to create a formula that computes the difference between Run Pulse and Rest Pulse.
- ☞ Now run a model of O<sub>2</sub> Uptake against all the response variables, including the new variable Run-Rest.

The report in **Figure 13.12** shows the signs of trouble. In the parameter estimates table, there are notations on Rest Pulse and Run Pulse that the estimates are biased, and on Run-Rest that it is zeroed. With exact linear dependency, the least squares solution is no longer unique. JMP chooses the solution that zeros out the parameter estimate for variables that are linearly dependent on other variables.

**Figure 13.12** Report When There Is a Linear Dependency

Term	Estimate	Std Error	t Ratio	Prob> t	
Intercept	84.269017	11.37593	7.41	<.0001*	
Run Time	-3.069814	0.419438	-7.32	<.0001*	
Rest Pulse	Biased	0.007994	0.075996	0.11	0.9170
Run Pulse	Biased	-0.116706	0.063497	-1.84	0.0775
Max Pulse		0.0851818	0.074936	1.14	0.2660
Run-Rest	Zeroed	0	0	.	.

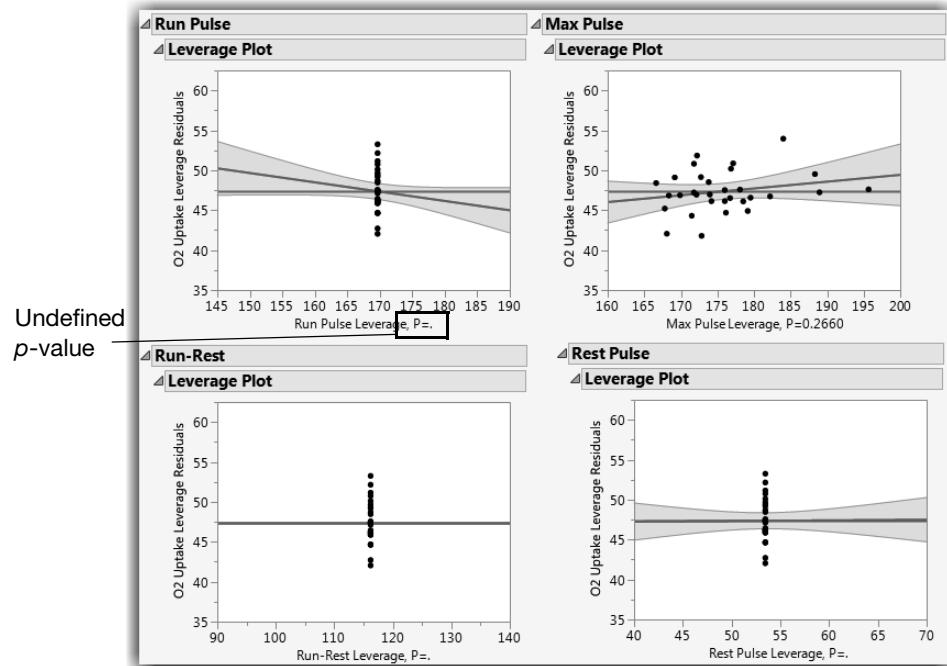
  

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Run Time	1	1	397.86637	53.5659	<.0001*
Rest Pulse	1	0	0.00000	.	.
Run Pulse	1	0	0.00000	.	.
Max Pulse	1	1	9.59746	1.2921	0.2660
Run-Rest	1	0	0.00000	.	.

The Singularity Details report at the top shows what the exact relationship is, in this case expressed in terms of Rest Pulse. The *t*-tests for the parameter estimates must now be interpreted in a conditional sense. JMP refuses to conduct *F*-tests for the non-estimable hypotheses for Rest Pulse, Run Pulse, and Run-Rest, and shows them with no degrees of freedom in the Effect Tests table.

You can see in the leverage plots for the three variables involved in the exact dependency, Rest Pulse, Run Pulse, and Run-Rest, that the points have completely collapsed horizontally—there are no points that have any leverage for these effects (see **Figure 13.13**). However, you can still test the unaffected regressors, like Max Pulse, and make good predictions.

**Figure 13.13** Leverage Plots When There Is a Linear Dependency



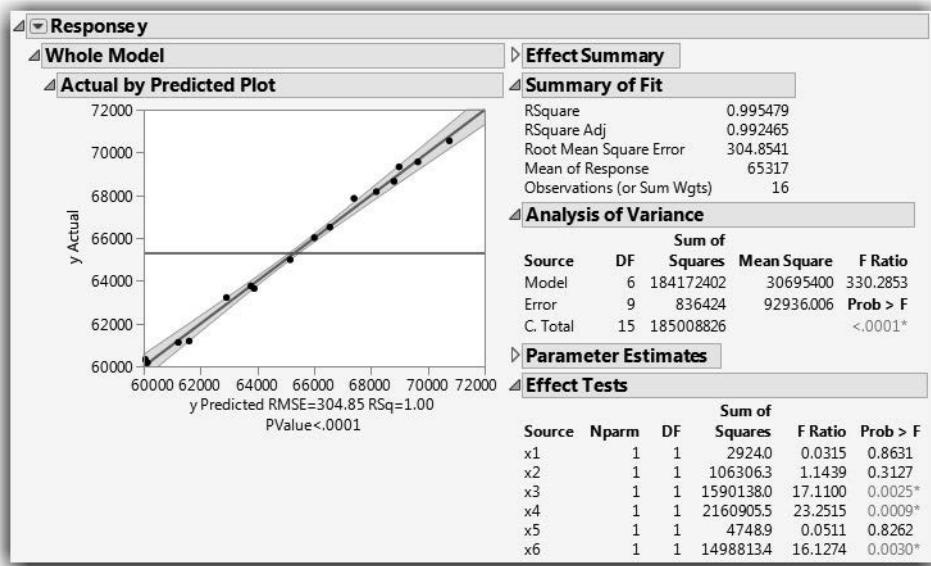
## The Longley Data: An Example of Collinearity

The Longley data set is famous, and is run routinely on most statistical packages to test accuracy of calculations. Why is it a challenge? Look at the data:

- ✓ Select **Help > Sample Data Library** and open Longley.jmp.
- ✓ Select **Analyze > Fit Model**.
- ✓ Assign y to **Y** and all the **X** columns as the model effects.
- ✓ Select **Effect Leverage** as the Emphasis and click **Run** to see results shown in **Figure 13.15**.

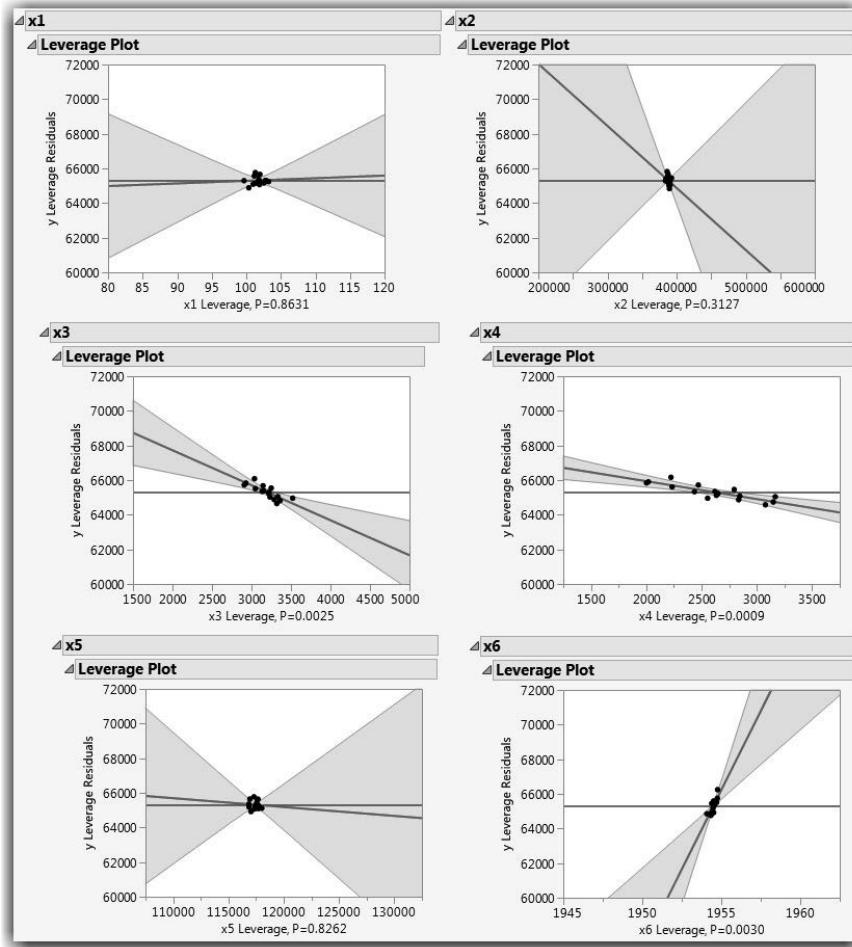
**Figure 13.14** shows the whole-model regression analysis. If you looked only at this overall picture, you would not see information about which (if any) of the six regressors are affected by collinearity. The data appear to fit well in the Actual by Predicted plot, the model has a significant *F* test, and there appear to be several significant factors. There is nothing in these reports to lead you to believe there might be a problem with collinearity.

**Figure 13.14** Analysis of Variance Report for the Longley Data



Now look at the leverage plots for the factors in **Figure 13.15**. The leverage plots show that  $x_1$ ,  $x_2$ ,  $x_5$ , and  $x_6$  have collinearity problems. Their leverage plots appear very unstable with the points clustered at the center of the plot and the confidence lines showing no confidence at all.

If you had looked only at the overall ANOVA picture, you would not have seen information about which (if any) of the six regressors were affected by collinearity.

**Figure 13.15** Leverage Plots for the Six Factors in the Longley Data

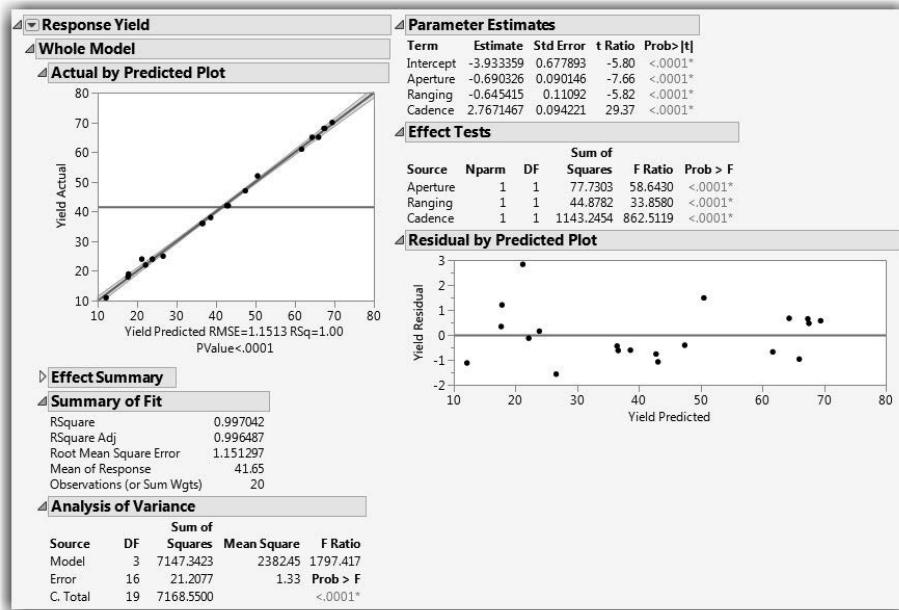
## The Case of the Hidden Leverage Point

Data were collected in a production setting where the yield of a process was related to three variables called Aperture, Ranging, and Cadence. Suppose you want to find out which of these effects are important, and in what direction:

- ☛ Select **Help > Sample Data Library** and open Ro.jmp.
- ☛ Select **Analyze > Fit Model**.
- ☛ Assign Yield to **Y**, Aperture, Ranging, and Cadence as the model effects, and then click **Run**.

JMP produces all the standard regression results, and many more graphics, including the residual plot on the lower right in **Figure 13.16**. Everything looks fine in the tables. The Summary of Fit table shows an  $R^2$  of 99.7%, which makes the regression model look like a great fit. All  $t$ -statistics are highly significant—but don't stop there.

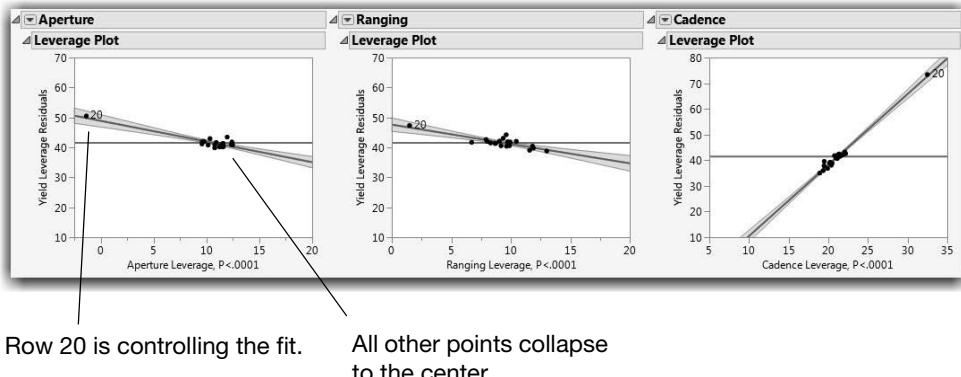
**Figure 13.16** Tables and Plots for Model with Collinearity



For each regression effect, there is a leverage plot showing what the residuals would be without that effect in the model. Note in **Figure 13.17** that row 20, which appeared unremarkable in the whole-model leverage and residual plots, is far out into the extremes of the effect leverage plots.

**Note:** The output in **Figure 13.17** has been rearranged for illustration.

It turns out that row 20 has monopolistic control of the estimates on all the parameters. All the other points appear to have little influence on the regression fit because they track the same part of the shrunken regression space.

**Figure 13.17** Leverage Plots That Detect Unusual Points

In a real analysis, row 20 would warrant special attention. Suppose, for example, that row 20 had an error, and its value was really 32 instead of 65.

Change the value of Yield in row 20 from 65 to 32 and run the model again.

The Parameter Estimates for both the modified table and original table are shown in **Figure 13.18**. The top table shows the parameter estimates computed from the data with an incorrect point. The bottom table has the corrected estimates. In high response ranges, the first prediction equation would give very different results than the second equation. The  $R^2$  is again high, and the parameter estimates are all significant—but every estimate is completely different even though only one point changed!

**Figure 13.18** Comparison of Analysis for Data with Outlier and Correct Data

The figure displays three JMP output tables side-by-side:

- Parameter Estimates - Original Data**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-3.933359	0.877893	-5.80	<.0001*
Aperture	-0.690326	0.090146	-7.66	<.0001*
Ranging	-0.645415	0.11092	-5.82	<.0001*
Cadence	2.7671467	0.094221	29.37	<.0001*

- Analysis of Variance - Modified Data**

Source	DF	Sum of		F Ratio
		Squares	Mean Square	
Model	3	6,659.1859	2,219.73	12620.66
Error	16	2.8141	0.18	Prob > F
C. Total	19	6,662.0000		<.0001*

- Parameter Estimates - Modified Data**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-0.035267	0.246935	-0.14	0.8882
Aperture	1.7097579	0.032837	52.07	<.0001*
Ranging	1.7318952	0.040405	42.86	<.0001*
Cadence	0.2853072	0.034322	8.31	<.0001*

## Mining Data with Stepwise Regression

Let's try a regression analysis on the O2Uptake variable with a set of 30 randomly generated columns as regressors. It seems like all results should be nonsignificant with random regressors, but that's not always the case.

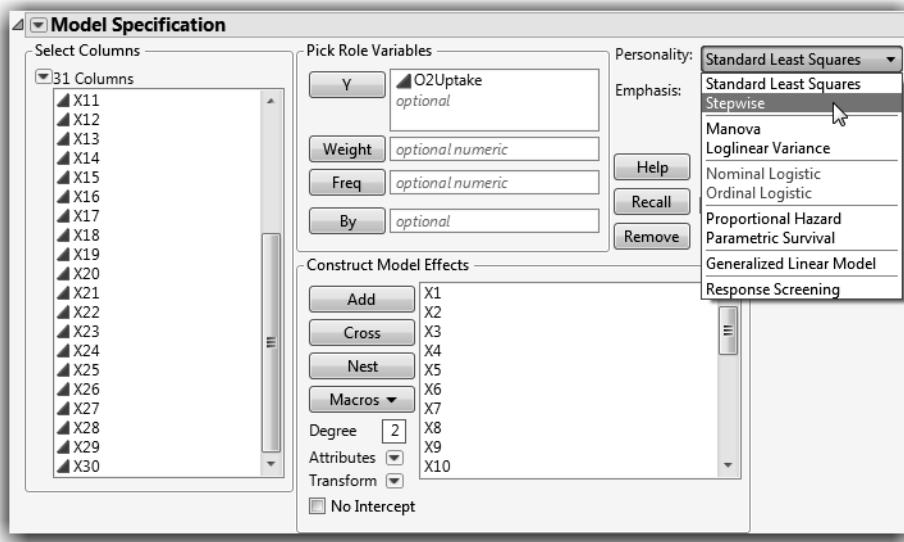
☞ Select **Help > Sample Data Library** and open Linnerand.jmp.

The data table has 30 columns named X1 to X30. Each column contains a uniform random number generator stored as a column formula. In other words, the data table is filled with random variables.

☞ Select **Analyze > Fit Model**.

☞ Assign O2Uptake to **Y** and X1 through X30 as the model effects.

☞ Select **Stepwise** from the **Personality** menu, as shown in **Figure 13.19** and then click **Run**.

**Figure 13.19** Model Specification Window for Stepwise Regression

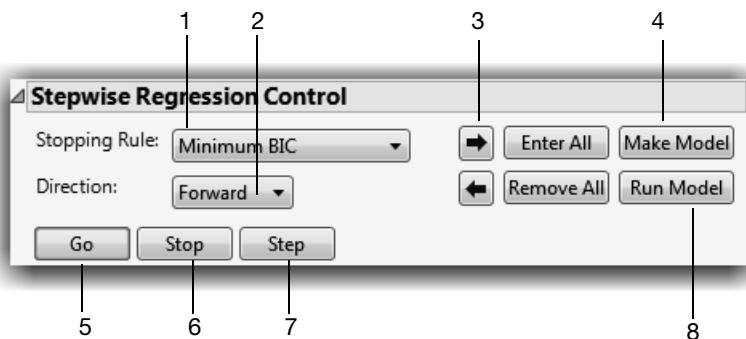
This Stepwise personality launches the Stepwise regression platform, which enables you to explore different combinations of effects. To run a stepwise regression, use the control panel that appears after you run the model (see **Figure 13.20**).

- ☞ Click **Go** in the Stepwise Regression Control panel to begin the stepwise variable selection process.

The recommended (default) stopping rule is **Minimum BIC**, which uses the minimum Bayesian Information Criterion to choose the best model. Minimum BIC is defined as  $-2\loglikelihood + k \ln(n)$  where  $k$  is the number of parameters and  $n$  is the sample size.

By default, stepwise runs a **Forward** selection process. You can also select **Backward** or **Mixed** as the selection process from the Direction menu. The Forward selection process adds the variable to the regression model that is most significant and computes the BIC. The process continues to add variables, one at a time, until the BIC is minimized.

**Note:** You could select the more familiar **P-value Threshold** stopping rule, and choose your own probability to enter and probability to remove a factor during the fitting process. Additional stopping rules are available in JMP Pro.

**Figure 13.20** Stepwise Regression Control Panel**Legend:**

1. Choose stopping rule.
2. Select type of selection process.
3. Arrows: Step forward and backward one step in the selection process.
4. Generate a completed Model Specification window using stepwise results.
5. Start stepwise process.
6. Halt the fitting process.
7. Step through process one step at a time.
8. Run Model using stepwise results.

The Current Estimates table shows the selected terms and resulting  $p$ -values (see **Figure 13.21**). You also see a Step History table (not shown here) that lists the order in which the variables entered the model.

The example in **Figure 13.21** was run with **Forward** direction to enter variables with the **Minimum BIC** stopping rule. Using different stopping rules and directions might lead to different models.

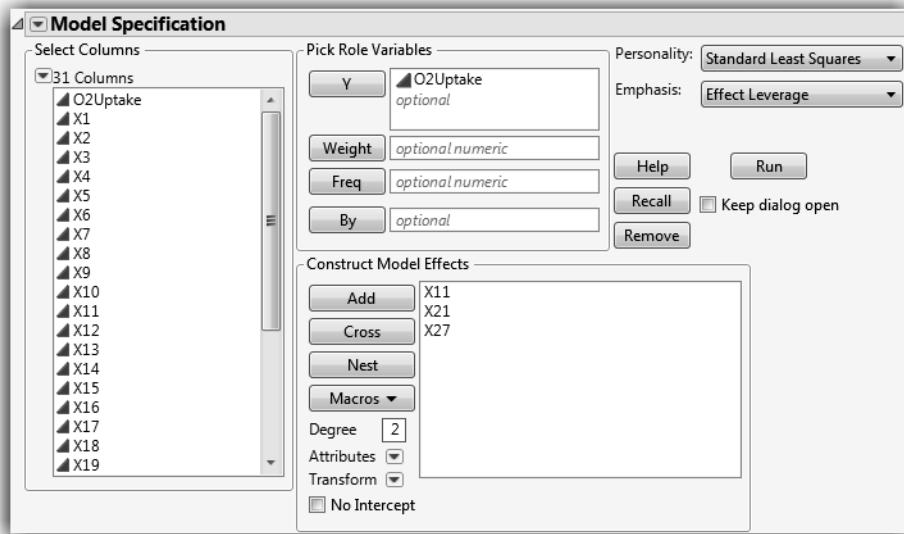
**Figure 13.21** Current Estimates Table Showing Selected Variables

	SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
	572.6814	27	4.6054774	0.3274	0.2526	.	4	190.7808	195.5507
Current Estimates									
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"		
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	58.5418447	1	0	0.000	1		
<input type="checkbox"/>	<input type="checkbox"/>	X1		0	1 13.47955	0.627	0.43572		
<input type="checkbox"/>	<input type="checkbox"/>	X2		0	1 4.413051	0.202	0.65691		
<input type="checkbox"/>	<input type="checkbox"/>	X3		0	1 0.000964	0.000	0.99477		
<input type="checkbox"/>	<input type="checkbox"/>	X4		0	1 0.010357	0.000	0.98287		
<input type="checkbox"/>	<input type="checkbox"/>	X5		0	1 7.122481	0.327	0.57209		
<input type="checkbox"/>	<input type="checkbox"/>	X6		0	1 19.41186	0.912	0.34832		
<input type="checkbox"/>	<input type="checkbox"/>	X7		0	1 18.95353	0.890	0.35417		
<input type="checkbox"/>	<input type="checkbox"/>	X8		0	1 21.86183	1.032	0.31907		
<input type="checkbox"/>	<input type="checkbox"/>	X9		0	1 0.597614	0.027	0.87037		
<input type="checkbox"/>	<input type="checkbox"/>	X10		0	1 3.075725	0.140	0.71093		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X11	-10.122243	1	154.102	7.265	0.01195		
<input type="checkbox"/>	<input type="checkbox"/>	X12		0	1 32.26746	1.552	0.22389		
<input type="checkbox"/>	<input type="checkbox"/>	X13		0	1 14.1288	0.658	0.42474		
<input type="checkbox"/>	<input type="checkbox"/>	X14		0	1 3.848537	0.176	0.67836		
<input type="checkbox"/>	<input type="checkbox"/>	X15		0	1 44.82726	2.208	0.14932		
<input type="checkbox"/>	<input type="checkbox"/>	X16		0	1 38.05512	1.851	0.18538		
<input type="checkbox"/>	<input type="checkbox"/>	X17		0	1 3.035589	0.139	0.71274		
<input type="checkbox"/>	<input type="checkbox"/>	X18		0	1 45.68236	2.254	0.14534		
<input type="checkbox"/>	<input type="checkbox"/>	X19		0	1 0.127703	0.006	0.93988		
<input type="checkbox"/>	<input type="checkbox"/>	X20		0	1 19.38053	0.911	0.34872		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X21	-5.9041302	1	97.96504	4.619	0.04075		
<input type="checkbox"/>	<input type="checkbox"/>	X22		0	1 25.25693	1.200	0.28345		
<input type="checkbox"/>	<input type="checkbox"/>	X23		0	1 8.459115	0.390	0.53784		
<input type="checkbox"/>	<input type="checkbox"/>	X24		0	1 2.255702	0.103	0.75104		
<input type="checkbox"/>	<input type="checkbox"/>	X25		0	1 0.6444316	0.029	0.86545		
<input type="checkbox"/>	<input type="checkbox"/>	X26		0	1 1.691257	0.077	0.78358		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X27	-6.3297117	1	97.57539	4.600	0.04112		
<input type="checkbox"/>	<input type="checkbox"/>	X28		0	1 5.718568	0.262	0.61291		
<input type="checkbox"/>	<input type="checkbox"/>	X29		0	1 17.80754	0.834	0.36939		
<input type="checkbox"/>	<input type="checkbox"/>	X30		0	1 0.000883	0.000	0.995		

- After the stepwise selection finishes selecting variables, click **Make Model** on the control panel.

The Model Specification window shown in **Figure 13.22** appears. You can run a standard least squares regression with the effects that were selected by the stepwise process as most active.

- Select **Run** on the Model Specification window.

**Figure 13.22** Model Specification Window Created by Make Model Option

When you run the model, you get the standard regression reports, partially shown here. The Parameter Estimates table shows all effects significant at the 0.05 level.

However, we created enough data to generate a number of coincidences, and then gathered those coincidences into one analysis and ignored the rest of the variables. This is like gambling all night in a casino, but exchanging money only for those hands where you win. When you mine data to the extreme, you get results that are too good to be true.

Summary of Fit	
RSquare	0.32735
RSquare Adj	0.252612
Root Mean Square Error	4.605477
Mean of Response	47.37581
Observations (or Sum Wgts)	31

Analysis of Variance	
Term	Estimate

Parameter Estimates					
Term	Estimate	Std Error	t Ratio	Prob> t	
Intercept	58.541845	3.22447	18.16	<.0001*	
X11	-10.12224	3.755323	-2.70	0.0119*	
X21	-5.90413	2.74723	-2.15	0.0407*	
X27	-6.329712	2.951131	-2.14	0.0411*	

## Exercises

1. The sample data table *Grandfather Clocks.jmp* (Smyth, 2000) contains data on grandfather clocks sold at open auction. Included are the selling price, age of the clock, and number of bidders on the clock. You are interested in predicting price based on the other two variables.

- (a) Use the Fit Model platform to construct a model using age as the only predictor of price. What is the  $R^2$  for this single predictor model?
- (b) Add the number of bidders to the model. Does the  $R^2$  increase markedly?
2. In *Gulliver's Travels*, the Lilliputians make an entire set of clothes for the (giant) Gulliver by taking only a few measurements from his body:

*"The seamstresses took my measure as I lay on the ground, one standing at my neck, and another at my mid-leg, with a strong cord extended, that each held by the end, while a third measured the length of the cord with a rule of an inch long. Then they measured my right thumb, and desired no more; for by a mathematical computation, that twice round the thumb is once round the wrist, and so on to the neck and the waist, and by the help of my old shirt, which I displayed on the ground before them for a pattern, they fitted me exactly."* (Swift, 1735)

Is there a relationship among the different parts of the body? The sample data table Body Measurements.jmp (Larner, 1996) contains measurements collected as part of a statistics project in Australia from 22 male subjects. In this exercise, you construct a model to predict the mass of a person based on other characteristics.

- (a) Using the Fit Model platform with personality **Standard Least Squares**, construct a model with mass as Y, and all other variables as effects in the model.
- (b) Examine the resulting report and determine the effect that has the least significance to the model. In other words, find the effect with the largest  $p$ -value, and use the Effects Summary table to remove this effect from the model. Note that the  $p$ -values reported in the Effects Summary table are the Prob > F values that are reported in the Effects Tests table.
- (c) Repeat part (b) until all effects have significance at the 0.05 level.
- (d) Now, use the Fit Model platform with personality **Stepwise** to produce a similar model. Enter all effects into a Backward stepwise model with Prob to Leave set at 0.05. Compare this model to the one you generated in part (c).
3. Biologists are interested in determining factors that predict the amount of time an animal sleeps during the day. To investigate the possibilities, Allison and Ciccetti (1976) gathered information about 62 different mammals. Their data is presented in the sample data table Sleeping Animals.jmp. The variables describe body weight, brain weight, total time spent sleeping in two different states (Dreaming and NonDreaming),

life span, and gestation time. The researchers also calculated indices to represent predation (1 meaning unlikely to be preyed upon, 5 meaning likely to be preyed upon), exposure (1 meaning that the animal sleeps in a well-protected den, 5 meaning most exposure), and an overall danger index, based on predation, exposure, and other factors (1 meaning least danger from other animals, 5 meaning most danger).

- (a) Use the Fit Y By X platform to examine the single-variable relationships between TotalSleep and the other variables. Which two variables look like they have the highest correlation with TotalSleep?
  - (b) If you remove NonDreaming from consideration, which two variables appear to be most correlated with TotalSleep?
  - (c) Construct a model using the two explanatory variables that you found in part (b). How well does this model predict TotalSleep?
  - (d) Construct a model using **P-Value Threshold** as the **Stopping Rule** and forward stepwise regression (still omitting NonDreaming), with 0.10 as the probability to enter and leave the model. Compare this model to the one you constructed in part (c).
  - (e) Construct a new model using mixed stepwise regression, with 0.10 as the probability to enter and leave the model. Compare this model to the other models that you have found. Which is the most effective at predicting total amount of sleep?
  - (f) Comment on the generalizability of this model. Would it be safe to use it to predict sleep times for a llama? Or a gecko? Explain your reasoning.
  - (g) Explore models that predict sleep in the dreaming and non-dreaming stages. Do the same predictors appear to be valid?
4. The sample data table Cities.jmp contains a collection of pollution data for 52 cities around the country.
- (a) Use the techniques of this chapter to build a model predicting Ozone for the cities listed. Use any of the continuous variables as effects.
  - (b) After you are satisfied with your model, determine whether there is an additional effect caused by the region of the country the city is in.
  - (c) In your model, interpret the coefficients of the significant effects.
  - (d) Comment on the generalizability of your model to other cities.





# 14

## Fitting Linear Models

### Overview

Several techniques, of increasing complexity, have been covered in this book. From fitting single means, to fitting multiple means, to fitting situations where the regressor is a continuous variable, specific techniques have been demonstrated to address a wide variety of statistical settings. This chapter introduces an approach involving *general linear models*, which encompasses all the models covered so far and extends to many more situations. They are all unified under the technique of least squares, which fits parameters to minimize the sum of squared residuals.

The techniques can be generalized even further to cover categorical response models, and other more specialized applications.

## Chapter Contents

Overview .....	377
The General Linear Model .....	379
Types of Effects in Linear Models.....	380
Coding Scheme to Fit a One-Way ANOVA as a Linear Model.....	381
Regressor Construction .....	384
Interpretation of Parameters .....	385
Predictions Are the Means.....	385
Parameters and Means .....	385
Analysis of Covariance: Continuous and Categorical Terms in the Same Model .....	386
The Prediction Equation.....	389
The Whole-Model Test and Leverage Plot.....	390
Effect Tests and Leverage Plots .....	391
Least Squares Means.....	393
Lack of Fit.....	394
Separate Slopes: When the Covariate Interacts with a Categorical Effect ..	396
Two-Way Analysis of Variance and Interactions .....	400
Optional Topic: Random Effects and Nested Effects .....	406
Nesting .....	407
Repeated Measures.....	409
Method 1: Random Effects-Mixed Model .....	409
Method 2: Reduction to the Experimental Unit .....	414
Method 3: Correlated Measurements-Multivariate Model.....	416
Varieties of Analysis .....	418
Closing Thoughts .....	418
Exercises.....	419

# The General Linear Model

Linear models are the sum of the products of coefficient parameters and factor columns. The linear model is rich enough to encompass most statistical work. By using a coding system, you can map categorical factors to regressor columns. You can also form interactions and nested effects from products of coded terms.

**Table 14.1** lists many of the situations handled by the general linear model approach. To read the model notation in **Table 14.1**, suppose that factors A, B, and C are categorical factors, and that X1, X2, and so on, are continuous factors.

**Table 14.1.** Different Linear Models

Situation	Model Notation	Comments
One-way ANOVA	$Y = A$	Add a different value for each level.
Two-way ANOVA no interaction	$Y = A, B$	Additive model with terms for A and B.
Two-way ANOVA with interaction	$Y = A, B, A*B$	Additive model with terms for A, B, and the combination of A and B.
Three-way factorial	$Y = A, B, A*B, C, A*C, B*C, A*B*C$	For $k$ -way factorial, $2^k - 1$ terms. The higher order terms are often dropped.
Nested model	$Y = A, B[A]$	The levels of B are only meaningful within the context of A levels (for example, City[State], pronounced “city within state”).
Simple regression	$Y = X_1$	An intercept plus a slope coefficient times the regressor.
Multiple regression	$Y = X_1, X_2, X_3, X_4, \dots$	There can be dozens of regressors.
Polynomial regression	$Y = X_1, X_{12}, X_{13}, X_{14}$	Linear, quadratic, cubic, quartic, and so on.
Quadratic response surface model	$Y = X_1, X_2, X_3, X_1^2, X_2^2, X_3^2, X_1*X_2, X_1*X_3, X_2*X_3$	All squares and cross products of effects define a quadratic surface with a unique critical value where the slope is zero (minimum, maximum, or a saddle point).
Analysis of covariance	$Y = A, X_1$	Main effect (A), adjusting for the covariate (X1).

Situation	Model Notation	Comments
Analysis of covariance with different slopes	$Y = A, X_1, A*X_1$	Tests that the covariate slopes are different in different A groups.
Nested slopes	$Y = A, X_1[A]$	Separate slopes for separate groups.
Multivariate regression	$Y_1, Y_2 = X_1, X_2 \dots$	The same regressors affect several responses.
MANOVA	$Y_1, Y_2, Y_3 = A$	A categorical variable affects several responses.
Multivariate repeated measures	sum and contrasts of $(Y_1 \ Y_2 \ Y_3) = A$ and so on.	The responses are repeated measurements over time on each subject.

## Types of Effects in Linear Models

The richness of the general linear model results from the type of effects you can include as columns in a coded model. Special sets of columns can be constructed to support various effects.

### Intercept Term

Most models have an intercept term to fit the constant in the linear equation. This is equivalent to having a regressor variable whose value is always 1. If this is the only term in the model, its purpose is to estimate the mean. This part of the model is so automatic that it becomes part of the background, the submodel to which the rest of the model is compared.

### Continuous Effects

These values are direct regressor terms, used in the model without modification. If all the effects are continuous, then the linear model is called multiple regression. (See Chapter 13 for an extended discussion of multiple regression.)

The only case in which intercepts are not used is the one in which the surface of fit must go through the origin. This happens in mixture models, for example. If you suppress the intercept term, then certain statistics (such as the whole-model  $F$ -test and the  $R^2$ ) do not apply because the question is no longer of just fitting a grand mean submodel against a full model.

In some cases (like mixture models), the intercept is suppressed, but there is a hidden intercept in the factors. This case is detected and the  $R^2$  and  $F$  are reported as usual.

### Categorical Effects

The model must fit a separate constant for each level of a categorical effect. These effects are coded columns through an internal coding scheme, which is described in the next section. These are also called *main effects* when contrasted with compound effects, such as interactions.

### Interactions

These are crossings of categorical effects, where you fit a different constant for each combination of levels of the interaction terms. Interactions are often written with an asterisk between terms, such as Age\*Sex. If continuous effects are crossed, they are multiplied together.

### Nested Effects

Nested effects occur when a term is meaningful only in the context of another term and thus is a type of a combined main effect and interaction with the term within which it is nested. For example, city is nested within state, because if you know you're in Chicago, you also know you're in Illinois. If you specify a city name alone, like Trenton, then Trenton, New Jersey, could be confused with Trenton, Michigan. Nested effects are written with the upper level term in parentheses or brackets, like City[State].

It is also possible to have combinations of continuous and categorical effects, and to combine interactions and nested effects.

## Coding Scheme to Fit a One-Way ANOVA as a Linear Model

When you include categorical variables in a model, JMP converts the categorical values (levels) into internal columns of numbers and analyzes the data as a linear model. The rules to make these columns are the *coding scheme*. These columns are sometimes called *dummy* or *indicator* variables. They make up an internal design matrix used to fit the linear model.

Coding determines how the parameter estimates are interpreted. However, note that the interpretation of parameters is different from the construction of the coded columns. In JMP, the categorical variables in a model are described as follows:

- There is an indicator column for each level of the categorical variable except the last level. An indicator variable is 1 for a row whose value is represented by that indicator; -1 for rows that have the last categorical level (for which there is no indicator variable); and 0 otherwise.

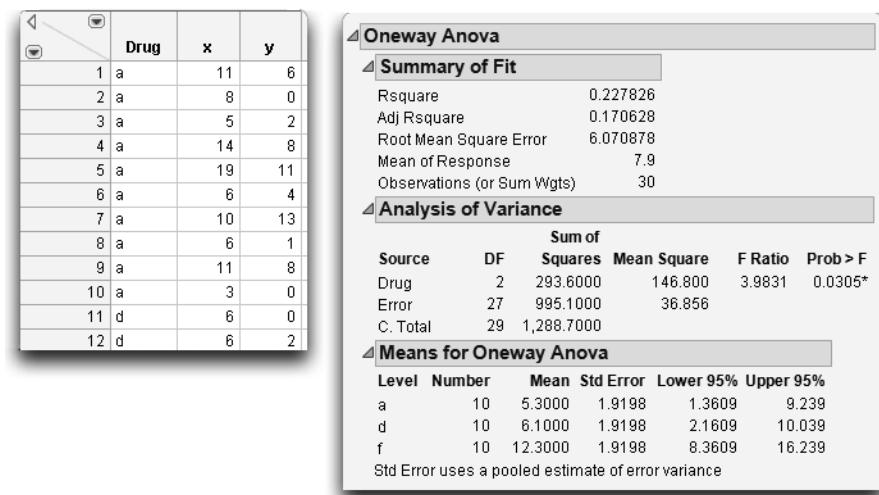
- A parameter is interpreted as the comparison of that level with the average effect across all levels. The effect of the last level is the negative of the sum of all the parameters for that effect. That is why this coding scheme is often called *sum-to-zero coding*.

Different coding schemes are the reason why different answers are reported in different software applications. The coding scheme doesn't matter in many simple cases, but it does matter in more complex cases, because it affects the hypotheses that are tested.

It's best to start learning the new approach covered in this chapter by looking at a familiar model. In Chapter 9, "Comparing Many Means: One-Way Analysis of Variance," you saw the Drug.jmp sample data, comparing three drugs (Snedecor and Cochran, 1967). Let's return to this data table and see how the general linear model handles a one-way ANOVA.

A partial listing of the sample data table called Drug.jmp, shown in **Figure 14.1**, contains the results of a study that measured the response of 30 subjects after treatment by one of three drugs. First, look at the one-way analysis of variance given previously by the Fit Y by X platform (y is the response and Drug is the X, Factor). We see that Drug is significant, with a *p*-value of 0.0305.

**Figure 14.1** Drug Data Table with One-Way Analysis of Variance Report



Now, do the same analysis using the **Fit Model** command.

- >Select **Help > Sample Data Library** and open Drug.jmp, which has variables called Drug, y, and, x.

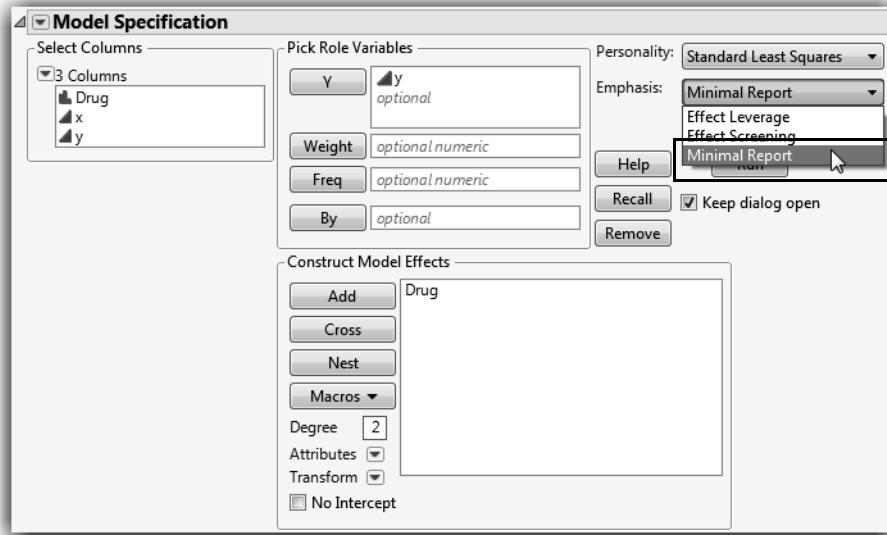
Drug has values “a”, “d”, and “f”, where “f” is really a placebo treatment. y is bacteria count after treatment, and x is a baseline count.

- >Select **Analyze > Fit Model**.
- In the Model Specification window, assign y to **Y (the response)** variable. Select Drug, and click **Add** to include it as a model effect.
- Select **Minimal Report** from the Emphasis menu to display only the basic reports.

The completed window should look like the one in **Figure 14.2**.

- To keep this window open for future use, select the **Keep dialog open** box.
- Click **Run** to see the analysis results in **Figure 14.3**.

**Figure 14.2** Model Specification Window for Simple One-Way ANOVA



Now compare the reports from this new analysis (**Figure 14.3**) with the one-way ANOVA reports in **Figure 14.1**. Note that the statistical results are the same: the same  $R^2$ , ANOVA F-test, means, and standard errors on the means. (The ANOVA F-test is in both the whole-model Analysis of Variance table and in the Effect Tests table because there is only one effect in the model.)

Although the two platforms produce the same results, the way the analyses were run internally was not the same. The Fit Model analysis ran as a regression on an intercept and two regressor variables constructed from the levels of the model main effect. The next section describes how this is done.

**Figure 14.3** ANOVA Results Given by the Fit Model Platform

The screenshot shows the Fit Model platform's output for an ANOVA analysis. The results are organized into several sections:

- Response y:** This section is collapsed.
- Summary of Fit:** This section is expanded, showing the following statistics:
 

RSquare	0.227826
RSquare Adj	0.170628
Root Mean Square Error	6.070878
Mean of Response	7.9
Observations (or Sum Wgts)	30
- Analysis of Variance:** This section is expanded, showing the ANOVA table:
 

Source	DF	Sum of		F Ratio
		Squares	Mean Square	
Model	2	293.6000	146.800	3.9831
Error	27	995.1000	36.856	Prob > F
C. Total	29	1,288.7000		0.0305*
- Parameter Estimates:** This section is expanded, showing the parameter estimates table:
 

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	7.9	1.108386	7.13	<.0001*
Drug[a]	-2.6	1.567494	-1.66	0.1088
Drug[d]	-1.8	1.567494	-1.15	0.2609
- Effect Tests:** This section is collapsed.
- Effect Details:** This section is collapsed.

## Regressor Construction

The terms in the Parameter Estimates table are named according to what level each is associated with.

The terms are called Drug[a] and Drug[d]. Drug[a] means that the regressor variable is coded as 1 when the level is “a”; –1 when the level is “f”; and 0 otherwise.

Drug[d] means that the variable is 1 when the level is “d”; –1 when the level is “f”; and 0 otherwise.

The screenshot shows the Fit Model platform's output for a Parameter Estimates analysis. The results are organized into several sections:

- Response y:** This section is collapsed.
- Summary of Fit:** This section is collapsed.
- Analysis of Variance:** This section is collapsed.
- Parameter Estimates:** This section is expanded, showing the parameter estimates table:
 

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	7.9	1.108386	7.13	<.0001*
Drug[a]	-2.6	1.567494	-1.66	0.1088
Drug[d]	-1.8	1.567494	-1.15	0.2609
- Effect Tests:** This section is collapsed.
- Effect Details:** This section is collapsed.

You can write the notation for Drug[a] as ([Drug=a]-[Drug=f]), where [Drug=a] is a one-or-zero indicator of whether the drug is “a” or not. The regression equation then looks like this:

$$y = b_0 + b_1 * ([Drug=a] - [Drug=f]) + b_2 * ([Drug=d] - [Drug=f]) + \text{error}$$

So far, the parameters associated with the regressor columns in the equation are represented by the names  $b_0$ ,  $b_1$ , and so on.

## Interpretation of Parameters

What is the interpretation of the parameters for the two regressors,  $b_1$ , and  $b_2$ ? The equation can be rewritten as

$$y = b_0 + b_1 * [Drug=a] + b_2 * [Drug=d] + (-b_1 - b_2) * [Drug=f] + \text{error}$$

The sum of the coefficients ( $b_1$ ,  $b_2$ , and  $-b_1 - b_2$ ) on the three indicators is always zero (again, sum-to-zero coding). The advantage of this coding is that the regression parameter tells you immediately how its level differs from the average response across all the levels.

## Predictions Are the Means

To verify that the coding system works, calculate the means (the predicted values for the levels “a”, “d”, and “f”) by substituting the parameter estimates shown previously into the regression equation

$$\text{Pred } y = b_0 + b_1 * ([Drug=a] - [Drug=f]) + b_2 * ([Drug=d] - [Drug=f])$$

For the “a” level,

$$\text{Pred } y = 7.9 + -2.6 * (1 - 0) + -1.8 * (0 - 0) = 5.3, \text{ which is the mean } y \text{ for "a".}$$

For the “d” level,

$$\text{Pred } y = 7.9 + -2.6 * (0 - 0) + -1.8 * (1 - 0) = 6.1, \text{ which is the mean } y \text{ for "d".}$$

For the “f” level,

$$\text{Pred } y = 7.9 + -2.6 * (0 - 1) + -1.8 * (0 - 1) = 12.3, \text{ which is the mean } y \text{ for "f".}$$

These means are shown in the Effect Details table.

## Parameters and Means

Now, substitute the means symbolically and solve for the parameters as functions of these means. First, write the equations for the predicted values for the three levels, called A for “a”, D for “d”, and F for “f”.

$$\text{MeanA} = b_0 + b_1 * 1 + b_2 * 0; \text{MeanD} = b_0 + b_1 * 0 + b_2 * 1; \text{MeanF} = b_0 + b_1 * (-1) + b_2 * (-1)$$

After solving for the  $b$  values, the following coefficients result:

$$b_1 = \text{MeanA} - (\text{MeanA} + \text{MeanD} + \text{MeanF})/3$$

$$b_2 = \text{MeanD} - (\text{MeanA} + \text{MeanD} + \text{MeanF})/3$$

$$(-b_1 - b_2) = \text{MeanF} - (\text{MeanA} + \text{MeanD} + \text{MeanF})/3$$

**Note:** Each level's parameter is interpreted as how different the mean for that group is from the *mean of the means for each level*.

In the next sections, you meet the generalization of this and other coding schemes, with each coding scheme having a different interpretation of the parameters.

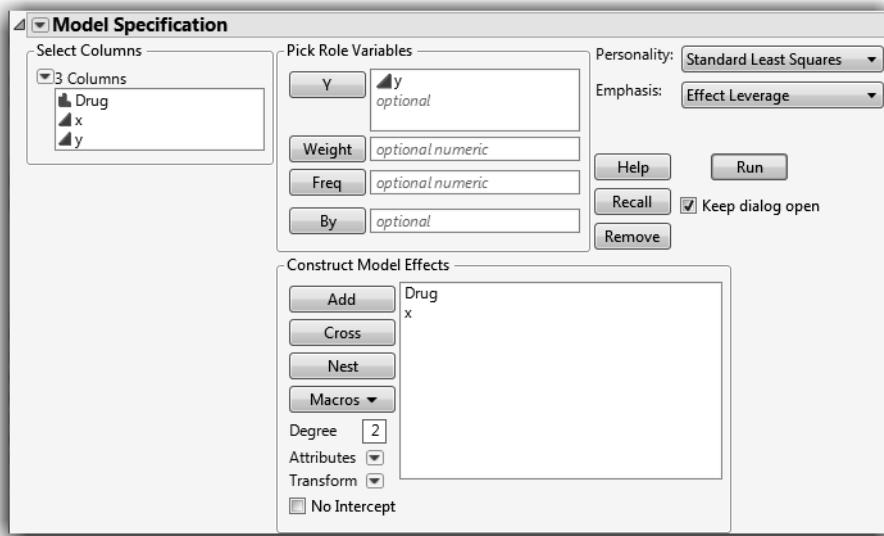
Keep in mind that the coding of the regressors does not necessarily follow the same rule as the interpretation of the parameters. (This is a result from linear algebra, resting on the fact that the inverse of a matrix is its transpose only if the matrix is orthogonal).

Overall, analysis using this coding technique is a way to use a regression model to estimate group means. It's all the same least squares results using a different approach.

## Analysis of Covariance: Continuous and Categorical Terms in the Same Model

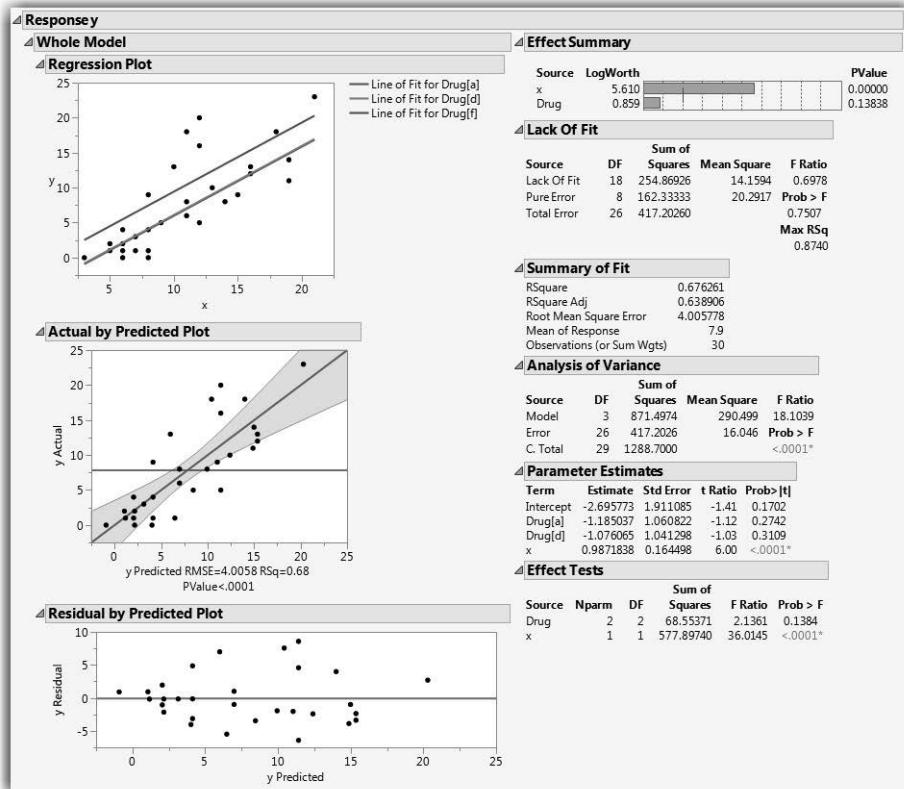
Now take the previous drug example that had one main effect (Drug), and add the other term ( $x$ ) to the model.  $x$ , the initial bacteria count, is a regular regressor, meaning that it is a continuous effect, and is called a *covariate*.

- ⌚ In the Model Specification window used earlier, click  $x$  and then **Add**. Now both Drug and  $x$  are effects, as shown in **Figure 14.4**.
- ⌚ Change the Emphasis back to the default, **Effect Leverage**, and click **Run** to see the results in **Figure 14.5**. The results are reorganized for illustrative purposes.

**Figure 14.4** Model Specification Window for Analysis of Covariance

This new model is a hybrid between the ANOVA models with nominal effects and the regression models with continuous effects. Because the analysis method uses a coding scheme, the categorical term can be put into the model with the regressor.

The new results show that adding the covariate  $x$  to the model raises the  $R^2$  from 22.78% (from **Figure 14.3**) to 67.62%. The parameter estimate for  $x$  is 0.987. This estimate is not unexpected because the response is the pre-treatment bacteria count, and  $x$  is the baseline count before treatment. With a coefficient of nearly 1 for  $x$ , the model is really fitting the difference in bacteria counts. The difference in counts has a smaller variation than the absolute counts.

**Figure 14.5** Analysis of Covariance Results Given by the Fit Model Platform

The  $t$ -test for  $x$  is highly significant. Because the Drug effect uses two parameters, refer to the Effect Tests or Effect Summary table to see whether Drug is significant. The null hypothesis for the  $F$ -test is that both parameters are zero. Surprisingly, the  $p$ -value for Drug changed from 0.0305 to 0.1384.

The Drug effect, which was significant in the previous model, is no longer significant! How could this be? The error in the model has been reduced, so it should be easier for differences to be detected.

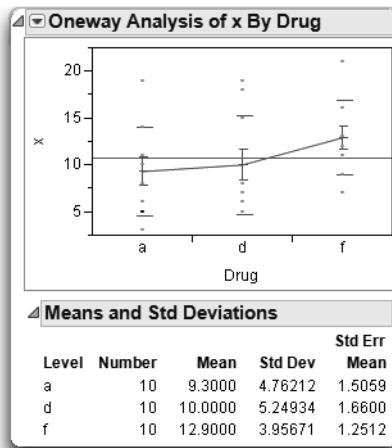
One possible explanation is that there might be a relationship between  $x$  and Drug.

- ✓ Select **Analyze > Fit Y by X**,
- ✓ Assign **x** to **Y, Response**, **Drug** to **X, Factor**, and then click **OK**.
- ✓ Select **Means and Std Dev** and **Display Options > Connect Means** from the red triangle menu next to Oneway Analysis

Look at the results, shown here, to examine the relationship of the covariate  $x$  to Drug.

It appears that the drugs have not been randomly assigned—or, if they were, they drew an unlikely unbalanced distribution. The toughest cases (with the most bacteria) tended to be given the inert drug “f.” This gave the “a” and “d” drugs a head start at reducing the bacteria count until  $x$  was brought into the model.

When fitting models where you don’t control all the factors, you might find that the factors are interrelated, and the significance of one depends on what else is in the model.



## The Prediction Equation

As shown earlier, the prediction equation generated by JMP can be stored as a formula in its own column.

- ⇨ Close the Fit Y by X window and return to the Fit Model results.
- ⇨ Select **Save Columns > Prediction Formula** from the red triangle menu next to Response.

This creates a new column in the data table called Pred Formula y. To see this formula:

- ⇨ Right-click in the column heading area and select **Formula**. Or, click on the plus sign next to the column in the Columns panel.

This opens a formula editor window with the following formula for the prediction equation:

$$-2.695772906 + \text{Match}(\text{Drug}) \begin{cases} "a" \Rightarrow -1.185036537 \\ "d" \Rightarrow -1.076065205 \\ "f" \Rightarrow 2.2611017426 \\ \text{else} \Rightarrow . \end{cases} + 0.9871838111 \cdot x$$

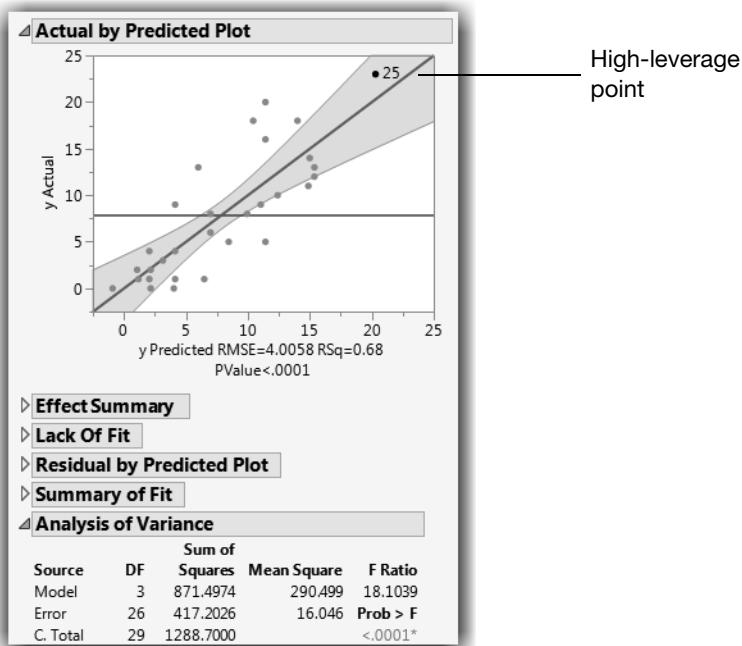
**Note:** The prediction equation can also be displayed in the Fit Model results window by selecting **Estimates > Show Prediction Expression** from the top red triangle menu.

## The Whole-Model Test and Leverage Plot

The whole-model test shows how the model fits as a whole compared with the null hypothesis of fitting only the mean. This is equivalent to testing the null hypothesis that all the parameters in the linear model are zero except for the intercept. This fit has three degrees of freedom, 2 from Drug and 1 from the covariate x. The  $F$  of 18.1 is highly significant (see **Figure 14.6**).

The whole-model leverage plot (produced when **Effect Leverage** is selected rather than **Minimal Report** in the Model Specification window) is a plot of the actual value versus its predicted value. A residual is the distance from each point to the  $45^{\circ}$  line of fit (where the actual is equal to the predicted).

**Figure 14.6** Whole-Model Test and Leverage Plot for Analysis of Covariance



The leverage plot in **Figure 14.6** shows the hypothesis test point by point. The points that are far out horizontally (like point 25) tend to contribute more to the test; the predicted values from the two models differ more there (the points are farther away from the mean). Points like this are called *high-leverage points*.

## Effect Tests and Leverage Plots

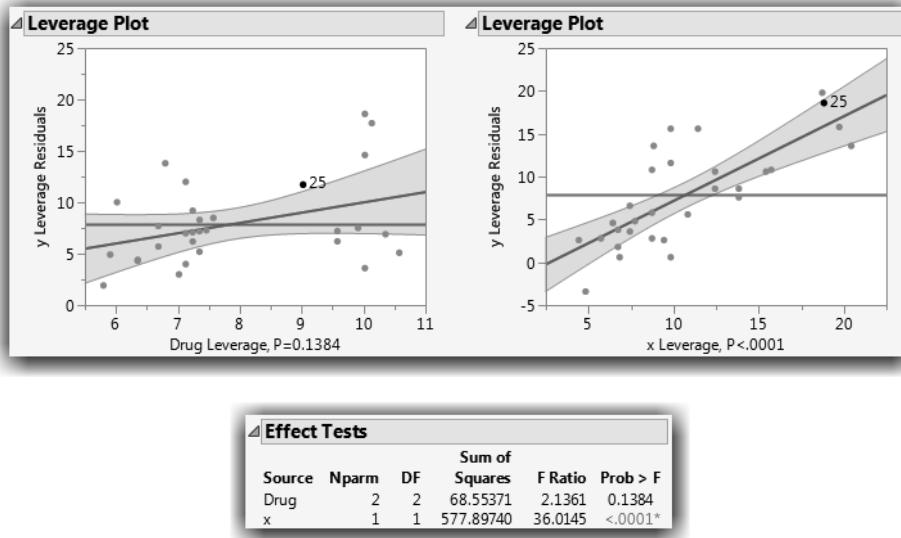
Now look at the effect leverage plots in **Figure 14.7** to examine the details for testing each effect in the model. Each effect test is computed from a difference in the residual sums of squares that compare the fitted model to the model without that effect.

For example, the sum of squares (SS) for  $x$  can be calculated by noting that the SS(error) for the full model is 417.2. But the SS(error) was 995.1 for the model that had only the Drug main effect (see **Figure 14.3**). The SS(error) for the model that includes the covariate is 417.2. The reduction in sum of squares is the difference,  $995.1 - 417.2 = 577.9$ , as you can see in the Effect Tests table (**Figure 14.7**). Similarly, if you remove Drug from the model, the SS(Error) grows from 417.2 to 485.8, a difference of 68.6 from the full model.

The leverage plot shows the composition of these sums of squares point by point. The Drug leverage plot on the left in **Figure 14.7** shows the effect on the residuals that would result from removing Drug from the model. The distance from each point to the sloped line is its residual. The distance from each point to the horizontal line is what its residual would be if Drug were removed from the model. The difference in the sum of squares for these two sets of residuals is the sum of squares for the effect, which, when divided by its degrees of freedom, is the numerator of the  $F$ -test for the Drug effect.

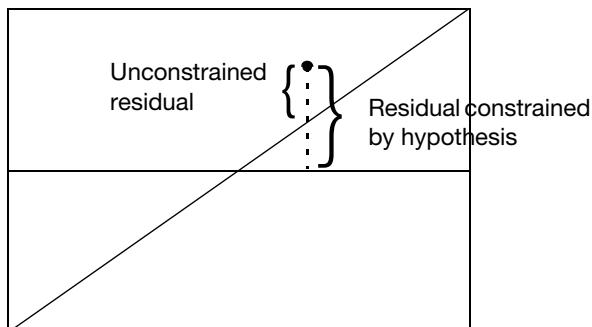
You can evaluate leverage plots in a way that is similar to evaluating a plot for a simple regression that has a mean line and confidence curves on it. The effect is significant if the points are able to pull the sloped line significantly away from the horizontal line. The confidence curves are placed around the sloped line to show the 0.05-level significance test. The curves cross the horizontal line if the effect is significant (as on the right in **Figure 14.7**). The curves encompass the horizontal line if the effect is not significant (as on the left).

Click on points to identify the high-leverage points—those that are away from the middle on the horizontal axis. Note whether they seem to support the test or not. If the points support the test, they are on the side trying to pull the line toward a higher slope.

**Figure 14.7** Effect Tests and Effect Leverage Plots for Analysis of Covariance

**Figure 14.8** summarizes the elements of a leverage plot. A leverage plot for a specific hypothesis is any plot with the following properties:

- There is a sloped line representing the full model and a horizontal line representing a model constrained by a hypothesis.
- The distance from each point to the sloped line is the residual from the full model.
- The distance from each point to the horizontal line is the residual from the constrained model.

**Figure 14.8** Schematic Defining a Leverage Plot

## Least Squares Means

It might not be fair to make comparisons between raw cell means in data that you fit to a linear model. Raw cell means do not compensate for different covariate values and other factors in the model. Instead, construct predicted values that are the expected value of a typical observation from some level of a categorical factor when all the other factors have been set to neutral values. These predicted values are called *least squares means*. There are other terms used for this idea: *marginal means*, *adjusted means*, and *marginal predicted values*.

The role of these adjusted or least squares means is that they allow comparisons of levels with the other factors being held fixed.

In the drug example, the least squares means are the predicted values expected for each of the three values of Drug, given that the covariate  $x$  is held at some constant value. For convenience, the constant value is chosen to be the mean of the covariate, which is 10.733.

The prediction equation gives the least squares means as follows:

fit equation:

$$-2.695 - 1.185 \text{ drug[a-f]} - 1.076 \text{ drug[d-f]} + 0.987 x$$

for "a":

$$-2.695 - 1.185 (1) - 1.076 (0) + 0.987 (10.733) = 6.715$$

for "d":

$$-2.695 - 1.185 (0) - 1.076 (1) + 0.987 (10.733) = 6.824$$

for "f":

$$-2.695 - 1.185 (-1) - 1.076 (-1) + 0.987 (10.733) = 10.161$$

To verify these results, return to the Fit Model platform and open the Least Squares Means table for the Drug effect (shown here).

In the diagram shown on the left in

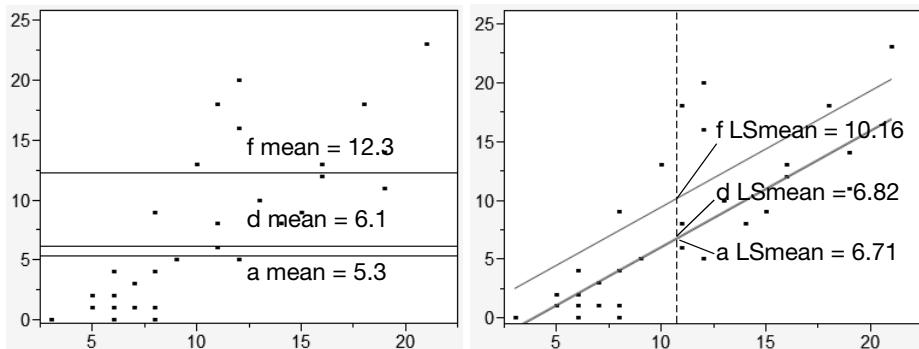
**Figure 14.9**, the ordinary means are taken with different values of the covariate, so it is not fair to compare them.

Least Squares Means Table			
Level	Least Sq Mean	Std Error	Mean
a	6.714963	1.2884943	5.3000
d	6.823935	1.2724690	6.1000
f	10.161102	1.3159234	12.3000

In the diagram to the right in **Figure 14.9**, the least squares means for this model are the intersections of the lines of fit for each level with the  $x$  value of 10.733.

With this data, the least squares means are less separated than the raw means. This LS Means plot is displayed in the Fit Model analysis window (see the Regression Plot at the top right of **Figure 14.5**).

**Figure 14.9** Diagram of Ordinary Means (top) and LS Means (bottom)



## Lack of Fit

The lack-of-fit test is the opposite of the whole-model test. Where the whole-model test (ANOVA) tests whether anything in the model is significant, lack-of-fit tests whether anything left out of the model is significant. Unlike all other tests, it is usually desirable for the lack-of-fit test to be nonsignificant. That is, the null hypothesis is that the model does not need higher-order effects. A significant lack-of-fit test is advice to add more effects to the model using higher orders of terms already in the model.

But how is it possible to test effects that haven't been put in the model? All tests in linear models compare a model with a constrained or reduced version of that model. To test all the terms that could be in the model but are not—now *that* would be amazing!

Lack-of-fit compares the fitted model with a saturated model using the same terms. A *saturated* model is one that has a parameter for each combination of factor values that exists in the data. For example, a one-way analysis of variance is already saturated because it has a parameter for each level of the single factor. A complete factorial with all higher order interactions is completely saturated. For simple regression, saturation would be like having a separate coefficient to estimate for each value of the regressor.

If the lack-of-fit test is significant, there is some significant effect that has been left out of the model. That effect is a function of the factors already in the model.

It could be a higher-order power of a regressor variable, or some form of interaction among categorical variables. If a model is already saturated, there is no lack-of-fit test possible.

The other requirement for a lack-of-fit test in continuous responses is that there be some exact replications of factor combinations in the data table. These exact duplicate rows (except for responses) allow the test to estimate the variation to use as a denominator in the lack-of-fit  $F$ -test. The error variance estimate from exact replicates is called *pure error* because it is independent of whether the model is right or wrong (assuming that it includes all the right factors).

In the drug model with covariate, the observations shown in **Table 14.2** form exact replications of data for Drug and  $x$ . The sum of squares around the mean in each replicate group reveals the contributions to pure error. (This calculation is shown for the first two rows).

This pure error represents the best that can be done in fitting these terms to the model for this data. Whatever is done to the model involving Drug and  $x$ , these replications and this error always exists. Pure error exists in the model regardless of the form of the model.

**Table 14.2.** Lack-of-Fit Analysis

Replicate Rows	Drug	x	y	Pure Error DF	Contribution to Pure Error
6	a	6	4	1	$4.5 = (4-2.5)^2 + (1-2.5)^2$
8	a	6	1		
1	a	11	6	1	$2.0 = (6-7)^2 + (8-7)^2$
9	a	11	8		
11	d	6	0	1	2.0
12	d	6	2		
14	d	8	1	2	32.667
16	d	8	4		
18	d	8	9		
27	f	12	5	2	120.667
28	f	12	16		
30	f	12	20		
21	f	16	13	1	0.5
26	f	16	12		
<b>Total</b>				<b>8</b>	162.333

Pure error can reveal how complete the model is. If the error variance estimate from the model is much greater than the pure error, then adding higher order effects of terms already in the model improves the fit.

The Lack-of-Fit table for this example is shown here. The difference between the total error from the fitted model and pure error is called *lack-of-fit error*. It represents all the terms that might have been added to the model, but were not. The ratio of the lack-of-fit mean square to the pure error mean square is the *F*-test for lack-of-fit. For the covariate model, the lack-of-fit error is not significant, which is good; it is an indication that the model is adequate with respect to the terms included in the model.

Lack Of Fit				
Source	DF	Sum of Squares		F Ratio
		Mean Square	Prob > F	
Lack Of Fit	18	254.86926	14.1594	0.6978
Pure Error	8	162.33333	20.2917	
Total Error	26	417.20260		0.7507
Max RSq				
0.8740				

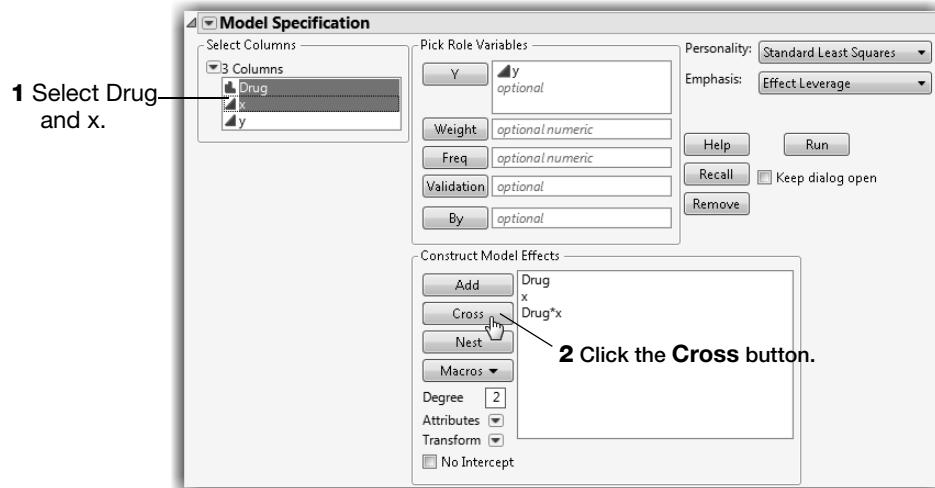
## Separate Slopes: When the Covariate Interacts with a Categorical Effect

When a covariate model includes a main effect and a covariate regressor, the analysis uses a separate intercept for the covariate regressor for each level of the main effect.

If the intercepts are different, could the slopes of the lines also be different? To find out, you need a method to capture the interaction of the regression slope with the main effect. This is accomplished by introducing a crossed term, the interaction of Drug and  $x$ , into the model:

- ⓐ Return to the Model Specification window, which already has Drug and  $x$  as effects in the model.
- ⓐ Hold down Shift and click Drug and  $x$  in the column selection list so that both columns are highlighted as shown in **Figure 14.10**.
- ⓐ Click the **Cross** button to create an effect in the model called Drug\*x.
- ⓐ Click **Run** to see the results shown in **Figure 14.11**. Note that some default output are not shown in the figure.

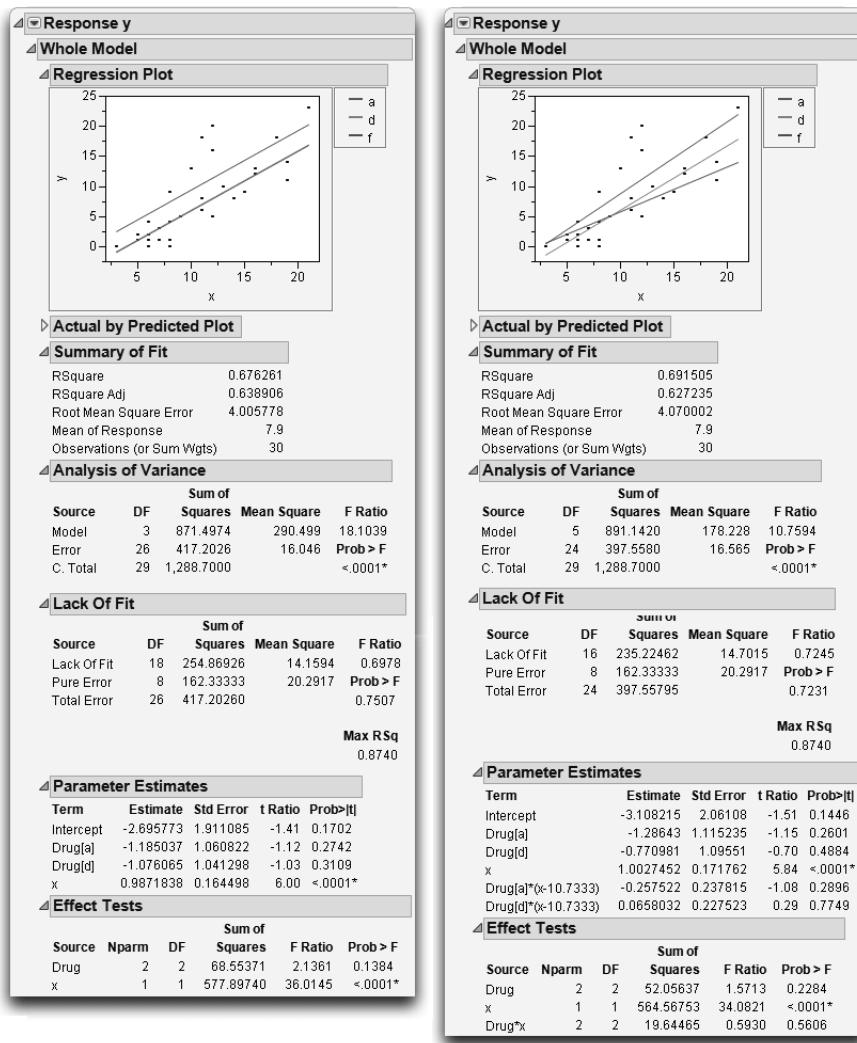
**Figure 14.10** Model Specification Window for Analysis of Covariance with Separate Slopes



This adds two parameters to the linear model that allow the slopes for the covariate to be different for each Drug level. The new variable is the product of the dummy variables for Drug and the covariate values.

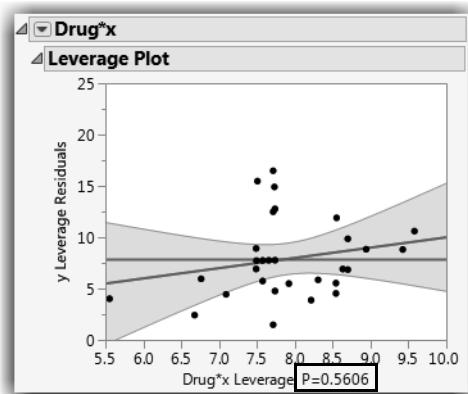
The Summary of Fit tables in **Figure 14.11** compare this separate slopes fit to the same slopes, showing an increase in  $R^2$  from 67.63% to 69.15%.

**Figure 14.11** Analysis of Covariance with Same Slopes (left) and Separate Slopes (right)

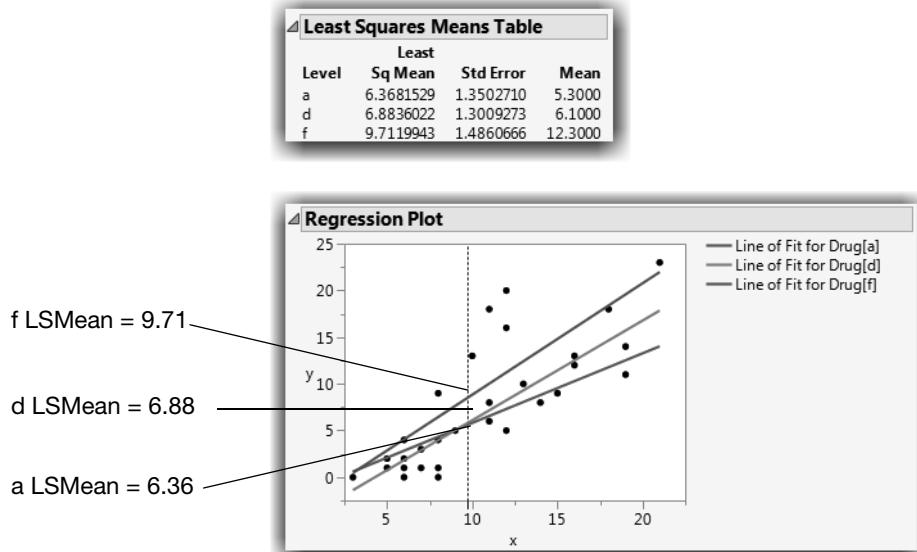


The separate slopes model shifts two degrees of freedom from the lack-of-fit error to the model, increasing the model degrees of freedom in the ANOVA from 3 to 5. The pure error seen in both Lack-of-Fit tables is the same because there are no new variables in the separate slopes covariance model. The new effect in the separate slopes model is constructed from terms already in the original analysis of covariance model.

The Effect Tests table in the report on the right in **Figure 14.11** shows that the test for the new term Drug\*x for separate slopes is not significant; the *p*-value is 0.56 (shown at the bottom of the plot on the right). The confidence curves on the leverage plot for the Effect Test do not cross the horizontal mean line, showing that the interaction term doesn't significantly contribute to the model. The least squares means for the separate slopes model have a more dubious value now.



Previously, with the same slopes on  $x$  as shown in **Figure 14.11**, the least squares means changed with whatever value of  $x$  was used. However, the separation between them did not. Now, with separate slopes as shown in **Figure 14.12**, the separation of the least squares means is also a function of  $x$ . The least squares means are more or less significantly different depending on whatever value of  $x$  is used. JMP uses the overall mean, but this does not represent any magic standard base. Notice that a and d cross near the mean of  $x$ . Their least squares means happen to be about the same only because of where the  $x$  covariate was set. Note that a reference line was added to **Figure 14.12** the X Axis Settings to show the mean of  $x$ .

**Figure 14.12** Illustration of Covariance with Separate Slopes

Interaction effects always have the potential to cloud the main effect, as will be seen again with the two-way model in the next section.

## Two-Way Analysis of Variance and Interactions

This section shows how to analyze a model in which there are two nominal or ordinal classification variables—a two-way model instead of a one-way model.

For example, suppose a popcorn experiment was run, varying three factors and measuring the popped volume yield per volume of kernels. The goal was to see which combination of factors gave the greatest volume of popped corn.

- ☞ To see the data, select **Help > Sample Data Library** and open Popcorn.jmp (shown in **Figure 14.13**).

The variables are popcorn, with values “plain” and “gourmet”; batch, which designates whether the popcorn was popped in a large or small batch; and oil amt with values “lots” or “little.” Start with two of the three factors, popcorn and batch.

**Figure 14.13** Listing of the Popcorn Data Table

	popcorn	oil amt	batch	yield	trial
1	plain	little	large	8.2	1
2	gourmet	little	large	8.6	1
3	plain	lots	large	10.4	1
4	gourmet	lots	large	9.2	1
5	plain	little	small	9.9	1
6	gourmet	little	small	12.1	1
7	plain	lots	small	10.6	1
8	gourmet	lots	small	18.0	1
9	plain	little	large	8.8	2
10	gourmet	little	large	8.2	2
11	plain	lots	large	8.8	2
12	gourmet	lots	large	9.8	2
13	plain	little	small	10.1	2
14	gourmet	little	small	15.9	2
15	plain	lots	small	7.4	2
16	gourmet	lots	small	16.0	2

Rows:  
All rows 16  
Selected 0  
Excluded 0  
Hidden 0  
Labelled 0

- ☞ Select **Analyze > Fit Model** and assign yield to **Y**.
- ☞ Select popcorn and batch and click **Add** to use them as the model effects.
- ☞ Change the Emphasis to **Minimal Report**.
- ☞ Click **Run** to see the analysis.

**Figure 14.14** shows the analysis tables for the two-factor analysis of variance:

- The model explains 56% of the variation in yield (the  $R^2$ ).
- The remaining variation has a standard error of 2.248 (Root Mean Square Error).
- The significant lack-of-fit test ( $p$ -value of 0.0019) says that there is something in the two factors that is not being captured by the model. The factors are affecting the response in a more complex way than is shown by main effects alone. The model needs an interaction term.
- Each of the two effects has two levels, so they each have a single parameter. Thus, the  $t$ -test results are identical to the  $F$ -test results. Both factors are significant.

**Figure 14.14** Two-Factor Analysis of Variance for Popcorn Experiment

**Response yield**

**Effect Summary**

Source	LogWorth	PValue
batch	2.085	0.00823
popcorn	1.677	0.02102

Remove Add Edit FDR

**Lack Of Fit**

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Lack Of Fit	1	37.210000	37.2100	15.6674	0.0019*
Pure Error	12	28.500000	2.3750		
Total Error	13	65.710000			

Max RSq  
0.8094

**Summary of Fit**

RSquare	0.560527
RSquare Adj	0.492916
Root Mean Square Error	2.248247
Mean of Response	10.75
Observations (or Sum Wgts)	16

**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Model	2	83.81000	41.9050	8.2904	
Error	13	65.71000	5.0546	Prob > F	
C. Total	15	149.52000			0.0048*

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	10.75	0.562062	19.13	<.0001*
popcorn[gourmet]	1.475	0.562062	2.62	0.0210*
batch[large]	-1.75	0.562062	-3.11	0.0082*

**Effect Tests**

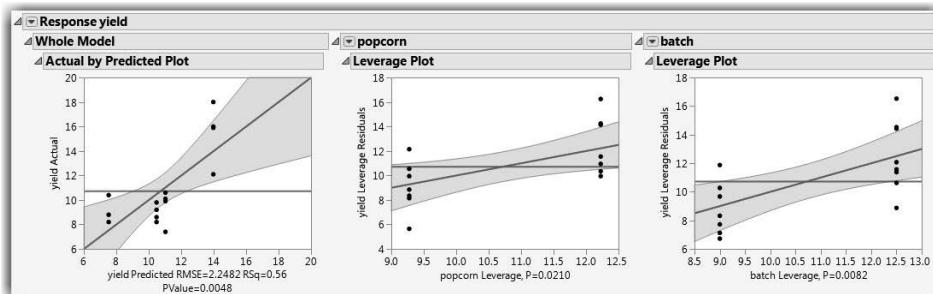
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
popcorn	1	1	34.810000	6.8868	0.0210*
batch	1	1	49.000000	9.6941	0.0082*

**Effect Details**

To see the point-by-point detail for the fit as a whole and the partial fits for each factor, we can add leverage plots to the analysis.

- ⇨ Return to the Fit Model window and change the Emphasis to Effect Leverage to see the reports shown in **Figure 14.15**.

Because this is a balanced design, all the points have the same leverage. Therefore, they are spaced out horizontally the same in the leverage plot for each effect.

**Figure 14.15** Leverage Plots for Two-Factor Popcorn Experiment

The lack-of-fit test shown in **Figure 14.14** is significant. We reject the null hypothesis that the model is adequate, and decide to add an interaction term, also called a *crossed effect*. An interaction means that the response is not simply the sum of a separate function for each term. In addition, each term affects the response differently depending on the level of the other term in the model.

The popcorn by batch interaction is added to the model as follows:

- ⇨ Return to the Model Specification window, which already has the popcorn and batch terms in the model. (Note: Select **Model Dialog** from the red triangle menu next to Response if the window is not open.)

First, select both popcorn and batch in the Select Columns list. To do so:

- ⇨ Click on popcorn and press Ctrl-click ( $\text{⌘}$ -click on the Macintosh) on batch to extend the selection.
- ⇨ Click the **Cross** button to see the popcorn\*batch interaction effect in the window.
- ⇨ Change the Emphasis to **Effect Leverage**, and click **Run** to see the results in **Figure 14.16**. Again, some default output are not shown in the figure.

Including the interaction term increases the  $R^2$  from 56% to 81%. The standard error of the residual (Root Mean Square Error) has gone down from 2.25 to 1.54.

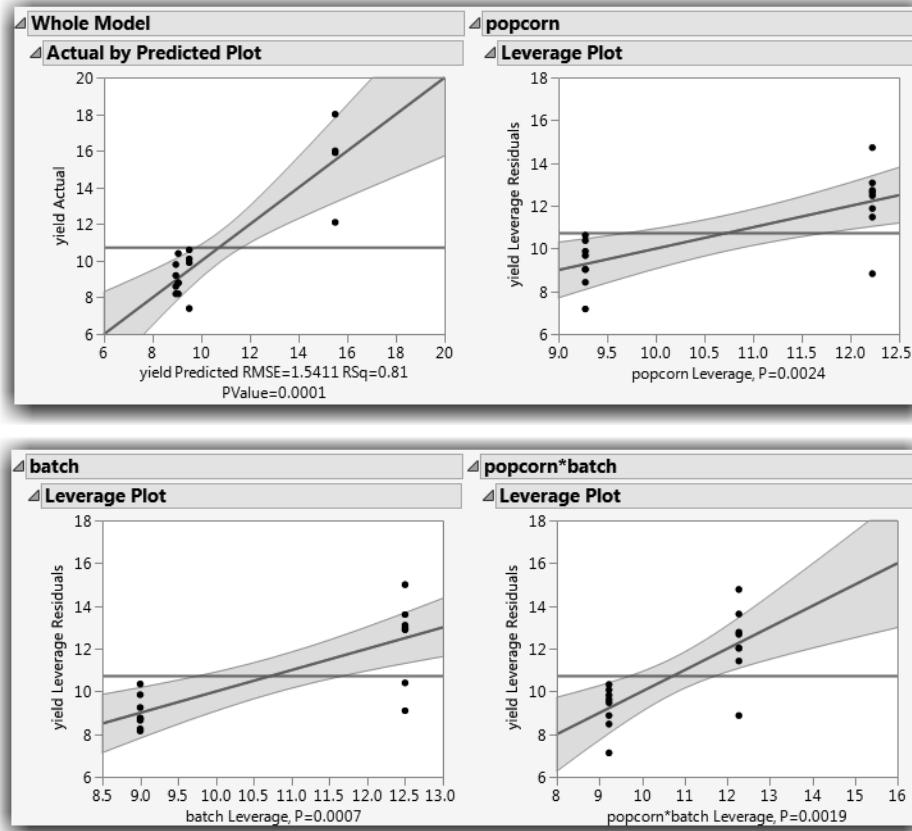
**Figure 14.16** Statistical Analysis of Two-Factor Experiment with Interaction

Response yield					
Summary of Fit					
RSquare	0.80939				
RSquare Adj	0.761738				
Root Mean Square Error	1.541104				
Mean of Response	10.75				
Observations (or Sum Wgts)	16				
Analysis of Variance					
Parameter Estimates					
Term	Estimate	Std Error	t Ratio	Prob> t	
Intercept	10.75	0.385276	27.90	<.0001*	
popcorn[gourmet]	1.475	0.385276	3.83	0.0024*	
batch[large]	-1.75	0.385276	-4.54	0.0007*	
popcorn[gourmet]*batch[large]	-1.525	0.385276	-3.96	0.0019*	
Effect Tests					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
popcorn	1	1	34.810000	14.6568	0.0024*
batch	1	1	49.000000	20.6316	0.0007*
popcorn*batch	1	1	37.210000	15.6674	0.0019*

The Effect Tests table shows that all effects are significant. The popcorn\*batch effect has a  $p$ -value of 0.0019, which is highly significant. The number of parameters (and degrees of freedom) of an interaction are the product of the number of parameters of each term in the interaction. The popcorn\*batch interaction has one parameter (and one degree of freedom) because the popcorn and batch terms each have only one parameter.

Balanced designs display an interesting phenomenon: the parameter estimates and sums of squares for the main effects are the same as in the previous fit without interaction. The  $F$ -tests are different only because the error variance (Mean Square Error) is smaller in the interaction model. The sum of squares and degrees of freedom for the interaction effect test is identical to the lack-of-fit test in the previous model.

Again, the leverage plots (**Figure 14.17**) show the tests in point-by-point detail. The confidence curves clearly cross the horizontal line. The effects tests (shown in **Figure 14.6**) confirm that the model and all effects are highly significant.

**Figure 14.17** Leverage Plots for Two-Factor Experiment with Interaction

You can see some details of the means in the least squares means table. However, in a balanced design (equal numbers in each level and no covariate regressors), the details are equal to the raw cell means.

- To see profile plots for each effect, select **LSMeans Plot** from the red triangle menu at the top of each leverage plot.

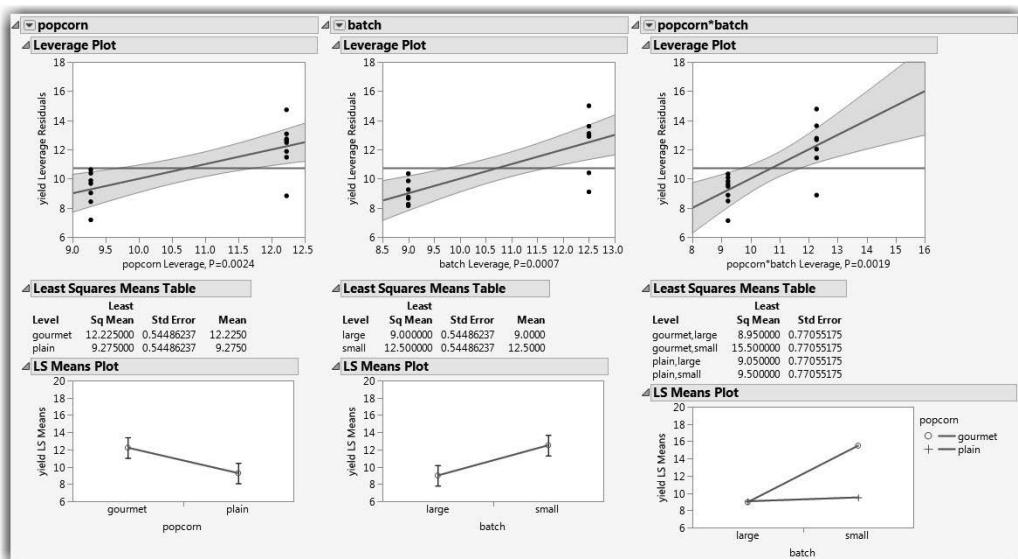
The result is a series of profile plots below each effect's report. Profile plots are a graphical form of the values in the Least Squares Means table.

- The leftmost plot in **Figure 14.18** is the profile plot for the popcorn main effect. The “gourmet” popcorn seems to have a higher yield.
- The middle plot is the profile plot for the batch main effect. It looks like small batches have higher yields.

- The rightmost plot is the profile plot for the popcorn by batch interaction effect.

Looking at the effects together in an interaction plot shows that the popcorn type matters for small batches but not for big ones. Said another way, the batch size is important for gourmet popcorn, but not for plain popcorn (gourmet popcorn does much better in smaller batches). In an interaction profile plot, one interaction term is on the  $x$ -axis, and the other term forms the different lines.

**Figure 14.18** Interaction Plots and Least Squares Means



## Optional Topic: Random Effects and Nested Effects

This section talks about nested effects, repeated measures, and random effects mixed models—a large collection of topics to cover in a few pages. So, hopefully the following overview provides an inspiration to look to other textbooks and study these topics more completely.

For example, consider the following situation. Six animals from two species were tracked, and the diameter of the area that each animal wandered was recorded.

Each animal was measured four times, once per season. **Figure 14.19** shows a partial listing of the animals data. To see this example:

- ⇨ Select **Help > Sample Data Library** and open Animals.jmp.

**Figure 14.19** Partial Listing of the Animals Data Table

	species	subject	miles	season
1	FOX	1	0	fall
2	FOX	1	0	winter
3	FOX	1	5	spring
4	FOX	1	3	summer
5	FOX	2	3	fall
6	FOX	2	1	winter
7	FOX	2	5	spring
8	FOX	2	4	summer
9	FOX	3	4	fall
10	FOX	3	3	winter
11	FOX	3	6	spring
12	FOX	3	2	summer
13	COYOTE	1	4	fall

## Nesting

One feature of the data is that the labeling for each subject animal is nested within species. The data (miles and season) for Fox subject 1 are not the same as the data for Coyote subject 1 (subject 1 occurs twice). The way to express this in a model is to always write the subject effect as subject[species], which is read as “subject nested within species” or “subject within species.”

The rule about nesting is that whenever you refer to a subject with a given level of factor, if that implies what another factor’s level is, then the factor should appear only in nested form.

When the linear model machinery in JMP sees a nested effect such as “B within A”, denoted B[A], it computes a new set of A parameters for each level of B. The Model Specification window allows for nested effects to be specified as in the following example.

- ⇨ Select **Help > Sample Data Library** and open Animals.jmp.
- ⇨ Select **Analyze > Fit Model** and assign miles to **Y**.
- ⇨ Add both species and subject to the model effects.

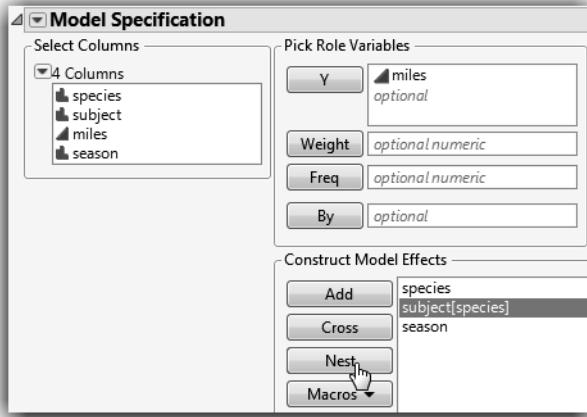
- Now, select species in the Select Columns list and also select subject in the model effects. With both highlighted, click the **Nest** button.

This adds the nested effect subject[species] shown here.

- Add season to the model effects and click **Run** to see the results in

**Figure 14.20.**

This model runs fine, but something is wrong with it. The *F*-tests for all the effects in the model use the residual error in the denominator. The reason that this is incorrect and the solution to this problem are presented in “Repeated Measures” on page 409.



**Figure 14.20** Results for Animal Data Analysis (Incorrect Analysis)

Response miles					
Summary of Fit					
RSquare		0.838417			
RSquare Adj		0.75224			
Root Mean Square Error		1.219062			
Mean of Response		4.458333			
Observations (or Sum Wgts)		24			
Analysis of Variance					
Sum of					
Source	DF	Squares	Mean Square	F Ratio	
Model	8	115.66667	14.4583	9.7290	
Error	15	22.29167	1.4861	Prob > F	
C. Total	23	137.95833		0.0001*	
Parameter Estimates					
Term		Estimate	Std Error	t Ratio	Prob >  t
Intercept		4.4583333	0.24884	17.92	<.0001*
species[COYOTE]		1.4583333	0.24884	5.86	<.0001*
species[COYOTE]:subject[1]		-0.666667	0.49768	-1.34	0.2003
species[COYOTE]:subject[2]		-0.666667	0.49768	-1.34	0.2003
species[FOX]:subject[1]		-1	0.49768	-2.01	0.0628
species[FOX]:subject[2]		0.25	0.49768	0.50	0.6227
season[fall]		-0.625	0.431003	-1.45	0.1676
season[spring]		1.7083333	0.431003	3.96	0.0012*
season[summer]		0.875	0.431003	2.03	0.0605
Effect Tests					
Sum of					
Source	Nparm	DF	Squares	F Ratio	Prob > F
species	1	1	51.041667	34.3458	<.0001*
subject[species]	4	4	17.166667	2.8879	0.0588
season	3	3	47.458333	10.6449	0.0005*

Let's look at the parameters in this nested model (**Figure 14.20**).

- There is one parameter for the two levels of species ("Fox" and "Coyote").
- Subject is nested in species, so there is a separate set of two parameters for three levels of subject within each level of species. This gives a total of four parameters for subject.
- Season, with four levels, has three parameters.

The total parameters for the model (not including the intercept) is:

$$+1 \text{ for species} + 4 \text{ for subject} + 3 \text{ for season} = 8.$$

## Repeated Measures

As previously mentioned, the animal data analysis has a problem. The *F*-test used to test the species effect is constructed using the model residual in the denominator, which isn't appropriate for this situation. The following sections explain this problem and outline solutions.

There are three ways to understand this problem, which correspond to three different (but equivalent) resolutions:

- Method 1: The effects can be declared as random, causing JMP to synthesize special *F*-tests.
- Method 2: The observations can be made to correspond to the experimental unit.
- Method 3: The analysis can be viewed as a multivariate problem.

The key concept here is that there are two layers of variation: *animal to animal* and *season to season within animal*. When you test an effect, the error term used in the denominator of the *F*-test for that effect must refer to the correct layer.

## Method 1: Random Effects-Mixed Model

The subject effect is what is called a *random effect*. The animals were selected randomly from a large population, and the variability from animal to animal is from some unknown distribution. To generalize to the whole population, study the species effect with respect to the variability.

It turns out that if the design is balanced, you can use an appropriate random term in the model as an error term instead of using the residual error to get an appropriate test. In this case, *subject[species]*, the nested effect, acts as an error term for the species main effect.

To construct the appropriate  $F$ -test, do a hand calculation using the results from the Effect Test table shown previously in **Figure 14.20**. Divide the mean square for species by the mean square for subject[species] as shown by the following formula.

$$F = \frac{\frac{51.041667}{1}}{\frac{17.166667}{4}}$$

This  $F$ -test has 1 numerator degree of freedom and 4 denominator degrees of freedom, and evaluates to 11.89.

Random effects in general statistics texts, often described in connection with *split plots* or *repeated measures designs*, describe which mean squares need to be used to test each model effect.

Now, let's have JMP do this calculation instead of doing it by hand. JMP gives the correct tests even if the design is not balanced.

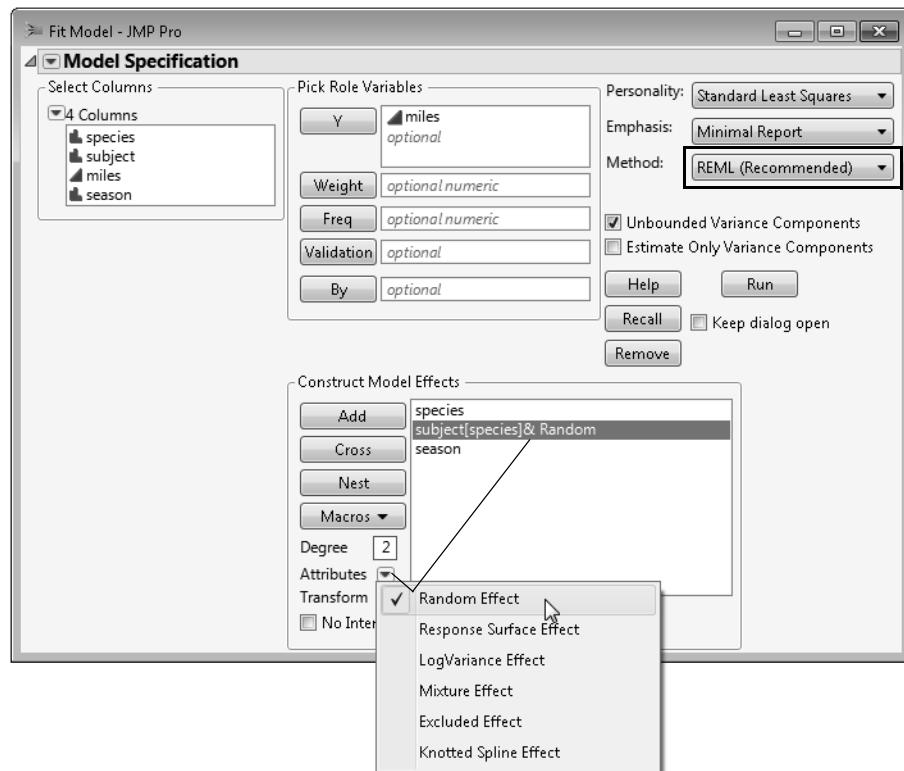
First, specify subject[species] as a random effect.

- ⓐ Return to the Model Specification window for the Animals data.
- ⓐ Click to highlight subject[species] in the model effects list.
- ⓐ Select **Random Effect** from the red triangle menu next to Attributes.

The subject[species] effect then appears with & Random appended to it as shown in **Figure 14.21**.

- ⓐ **REML (Recommended)** is the analysis Method, as shown.
- ⓐ Change the Emphasis to **Minimal Report**.
- ⓐ Click **Run** to see the results in **Figure 14.22**. Again, some default output is not shown in the figure.

**Note:** The Least Squares Means Table shown in **Figure 14.22** is in the Effect Details outline.

**Figure 14.21** Model Specification Window Using a Random Effect

### The REML (Restricted Maximum Likelihood) Method for Fitting Random Effects

Random effects are those in which the levels are chosen randomly from a larger population of levels. These random effects represent a sample from the larger population. In contrast, the levels of fixed effects (nonrandom) are of direct interest and are often specified in advance. If you have both random and fixed effects in a model, it is called a *mixed model*.

This example is a mixed model, and uses the REML method to fit the random effect, `subject[species]`. For historical interest only, the Fit Model platform also offers the Method of Moments (EMS). However, EMS is no longer recommended except in special cases where it is equivalent to REML. The REML method for fitting mixed models is now the mainstream, state-of-the-art method, supplanting older methods.

The REML approach was pioneered by Patterson and Thompson in 1974. See also Wolfinger, Tobias, and Sall (1994) and Searle, Casella, and McCulloch (1992). The reason to prefer REML is that it works without requiring balanced data, or shortcut approximations. And it gets all the tests right, even contrasts that work across interactions. Most packages that use the traditional EMS method either are not able to test some of these contrasts, or compute incorrect variances for them.

As we've discussed, levels in random effects are randomly selected from a larger population of levels. For the purpose of hypothesis testing, the distribution of the effect on the response over the levels is assumed to be normal, with mean zero and some variance (called a variance component). Often, the exact effect sizes are not of direct interest. It is the fact that they represent the larger population that is of interest. What you learn about the mean and variance of the effect tells you something about the general population from which the effect levels were drawn. This is different from fixed effects, where you know only about the levels you actually encounter in the data.

**Negative Variance Estimates** - JMP can produce *negative estimates* for both REML and EMS. For REML, there are two check boxes in the model launch window:

**Unbounded Variance Components** - Deselecting this check box constrains the estimate to be nonnegative. We recommend that you do not deselect this if you are interested in estimating fixed effects.

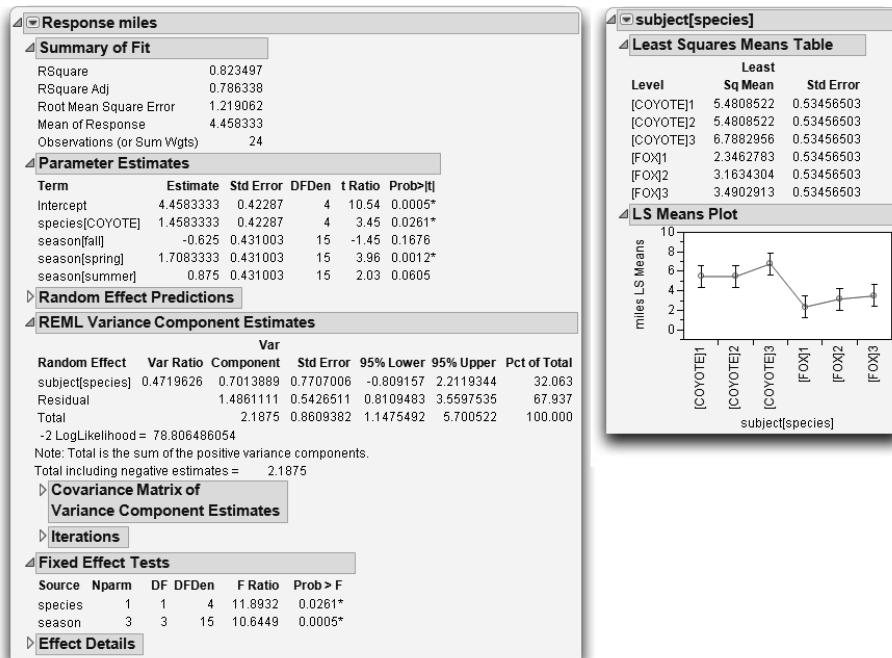
**Estimate Only Variance Components** - Constraining the variance estimates leads to bias in the tests for the fixed effects. If, however, you are interested only in variance components, and you do not want to see negative variance components, then selecting the check box beside Estimate Only Variance Components is appropriate.

### REML Results

The Fit Model Specification in **Figure 14.21** produces the report shown in **Figure 14.22**. A nice feature of REML is that the report doesn't need qualification. The estimates are all properly shrunk and the standard errors are properly scaled (SAS Institute Inc. 2008). The variance components are shown as a ratio to the error variance, and as a portion of the total variance.

**Note:** The Least Squares Means Table shown on the right in **Figure 14.22** is in the Effect Details outline (the bottom left in **Figure 14.22**). The LS Means Plot was selected from the red triangle menu next to subject[species].

**Figure 14.22** REML Results for Animal Data



There is no special table of synthetic test creation, because all the adjustments are automatically taken care of by the model itself. There is no table of expected means squares, because the method does not need this.

If you have random effects in the model, the analysis of variance report is not shown. This is because the variance does not partition in the usual way, nor do the degrees of freedom attribute in the usual way for REML-based estimates. You can obtain the residual variance estimate from the REML report rather than from the analysis of variance report.

The Variance Component Estimates table shows 95% confidence intervals for the variance components using the Satterthwaite (1946) approximation. You can right-click the Variance Components Estimates table to display the Kackar-Harville correction (under **Columns > Norm KHC**). This value is an approximation of the magnitude of the increase in the mean squared errors of the estimators for

the mixed model. See Kackar and Harville (1984) for a discussion of approximating standard errors in mixed models.

#### Select **Estimates > Show Prediction Expression**

from the red triangle menu next to Response miles to see the prediction equation, shown here.

**Note:** The Mixed personality in JMP Pro provides additional options for fitting mixed models.

## Method 2: Reduction to the Experimental Unit

There are only six animals, but there are 24 rows in the data table because each animal is measured four times. However, taking four measurements on each animal doesn't make it legal to count each measurement as an observation. Measuring each animal millions of times and throwing all the data into a computer would yield an extremely powerful—and incorrect—test.

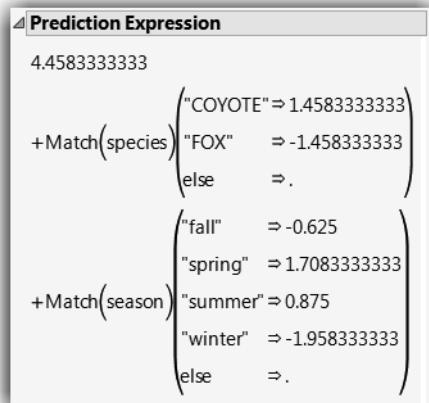
The experimental unit that is relevant to species is the individual animal, not each measurement of that animal. When testing effects that vary only subject to subject, the experimental unit should be a single response per subject, instead of the repeated measurements.

One way to handle this situation is to group the data to find the means of the repeated measures, and analyze these means instead of the individual values.

- ⓐ With the Animals.jmp sample data table active, select **Tables > Summary**.
- ⓐ Assign both species and subject as grouping variables.
- ⓐ Select the miles variable as shown at the top of **Figure 14.23** and select **Mean** from the **Statistics** menu to add Mean(miles).

This notation indicates that you want to see the mean (average) miles for each subject within each species.

- ⓐ Click **OK** to see the summary table shown at the bottom of **Figure 14.23**.



**Figure 14.23** Summary Window and Summary Table for Animals Data

The screenshot shows the SAS interface with three main windows:

- Request Summary Statistics by Grouping Columns.** This dialog box contains settings for summarizing data by species and subject. It includes a list of columns (species, subject, miles, season), a statistics section for 'Mean(miles)', and a list of actions like OK, Cancel, Remove, Recall, and Help.
- Action** window: A vertical list of statistical measures including N, Mean, Std Dev, Min, Max, Range, % of Total, N Missing, N Categories, Sum, Sum Wgt, Variance, Std Err, CV, Median, Geometric Mean, Interquartile Range, Quantiles, and Histogram.
- Summary Table**: A data grid titled 'Animals By (species, subject)'. It has columns for species, subject, N Rows, and Mean(miles). The data shows observations for COYOTE and FOX across three subjects, with mean distances ranging from 2.0 to 7.3 miles.

Now fit a model to the summarized data.

- ⓐ With the new summary table (Animals By (species, subject)) active, select **Analyze > Fit Model**.
- ⓑ Use Mean(miles) as Y, and species as the single model effects variable.
- ⓒ Click **Run** to see the proper *F*-test of 11.89 for species, with a *p*-value of 0.0261.

Effect Tests					
Source	Nparm	DF	Sum of	F Ratio	Prob > F
			Squares		
species	1	1	12.760417	11.8932	0.0261*

Note that this result for Species is the same as the calculation shown in the Fixed Effect Test table in the previous section.

## Method 3: Correlated Measurements-Multivariate Model

In the animal example, there were multiple (four) measurements for the same animal. These measurements are likely to be correlated in the same sense as two measurements are correlated in a paired *t*-test. This situation of multiple measurements of the same subject is called *repeated measures*, or a *longitudinal* situation. This type of experimental situation can be looked at as a multivariate problem. The analysis is also called a Multivariate ANalysis Of VAriance (MANOVA).

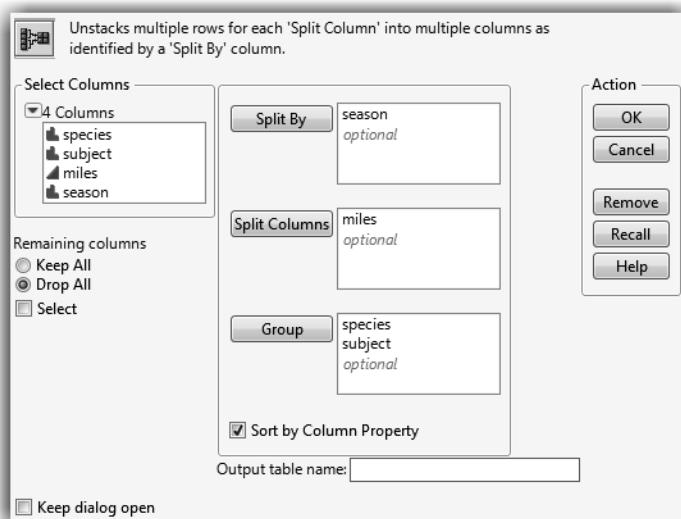
To use this multivariate approach, the data table must be rearranged so that there is only one row for each individual animal, with the four measurements on each animal in four columns.

- ☞ To rearrange the original Animals table, select **Tables > Split** to split the miles column into four columns, one for each season.

To complete the Split window:

- ☞ Assign miles to **Split Columns**.
- ☞ Assign season to **Split By** (its values become the new column names).
- ☞ Assign species and subject to **Group** (**Figure 14.24**) and then click **OK**.

**Figure 14.24** Split Window to Arrange Data for Repeated Measures Analysis



The new table (**Figure 14.25**) shows the values of miles in four columns named by the four seasons. The group variables (species and subject) ensure that the rows are formed correctly.

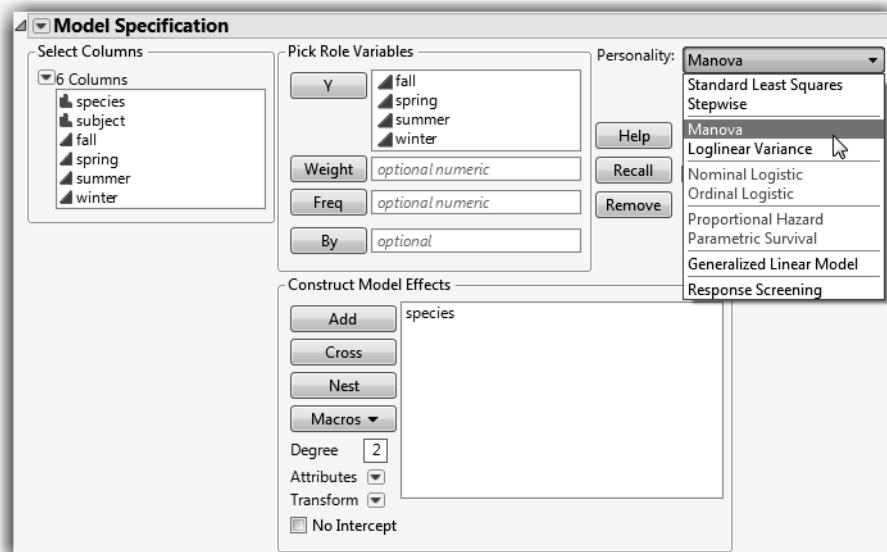
**Figure 14.25** Rearrangement of the Animals Data for MANOVA

	species	subject	fall	spring	summer	winter
1	COYOTE	1	4	7	8	2
2	COYOTE	2	5	6	6	4
3	COYOTE	3	7	8	9	5
4	FOX	1	0	5	3	0
5	FOX	2	3	5	4	1
6	FOX	3	4	6	2	3

Now, fit a multivariate model with four Y variables and a single response:

- ⓐ Select **Analyze > Fit Model** and assign fall, spring, summer, and winter to **Y**.
- ⓑ Select species and click **Add** to assign it as the only model effect.
- ⓒ Select **Manova** from the Personality menu (**Figure 14.26**) and then click **Run**.

**Figure 14.26** Model Specification Window for MANOVA Analysis

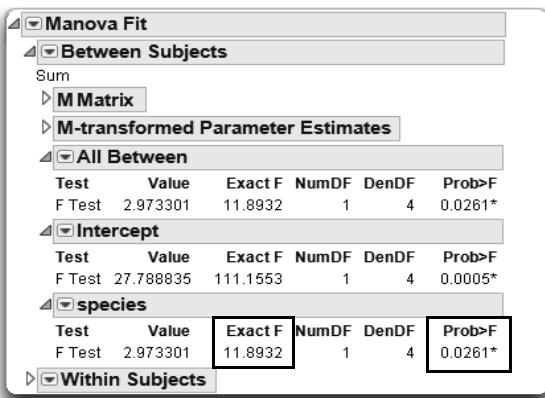


- ⓐ First, the Response Specification control panel appears. On the control panel, select **Choose Response > Repeated Measures**.

✓ Click **OK** to accept the default name Time.

The partial report is shown in **Figure 14.27**.

**Figure 14.27** MANOVA of Repeated Measures Data



The Between Subjects report for species in **Figure 14.27** shows the same  $F$ -test with the value of 11.89 and a  $p$ -value of 0.0261, just the same as the other methods.

## Varieties of Analysis

In the previous cases, all the tests resulted in the same  $F$ -test for species. However, it is not generally true that different methods produce the same answer. For example, if species had more than two levels, you would see the four standard multivariate tests (Wilk's lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's maximum root). Each of these tests would produce a different test result, and none of them would agree with the mixed model approach discussed previously.

Two more tests involving adjustments to the univariate method can be obtained from the multivariate fitting platform (MANOVA). With an unequal number of measurements per subject, the multivariate approach cannot be used. With unbalanced data, the mixed model approach also plunges into a diversity of methods that offer different answers from the method that JMP uses (the Method of Moments).

## Closing Thoughts

When using the residual error to form  $F$ -statistics, ask whether the row in the table corresponds to the unit of experimentation. Are you measuring variation in an appropriate way for the effect that you are examining?

- When the situation does not live up to the framework of the statistical method, the analysis is always incorrect. This happened in the first method shown (treating data table observations as experimental units) in the nesting example for the species test.
- Statistics offers a diversity of methods (and a diversity of results) for the same question. In this example, the different results are not wrong; they are just different.
- Statistics is not always simple. There are many ways to go astray. Common sense goes a long way, but there is no substitute for expert advice when the situation warrants it.

## Exercises

1. Denim manufacturers are concerned with maximizing the comfort of the jeans that they make. In the manufacturing process, starch is often built up in the fabric, creating a stiff jean that must be “broken in” by the wearer before they are comfortable. To minimize the break-in time for each pair of pants, the manufacturers often wash the jeans to remove as much starch as possible. This example concerns three methods of this washing. The sample data table Denim.jmp (Creighton, 2000) contains four columns: Method (describing the method used in washing), Size of Load (in pounds), Sand Blasted (recording whether the fabric was sand blasted prior to washing), Thread Wear (a measure of the destruction of threads during the process), and Starch Content (the starch content of the fabric after washing).
  - (a) The three methods of washing consist of using an enzyme (alpha amylase) to destroy the starch, chemically destroying the starch (caustic soda), and washing with an abrasive compound (pumice stone, hence the term “stone-washed jeans”). Determine whether there is a significant difference in starch content due to washing method in two ways: using the Fit Y by X platform and the Fit Model platform. Compare the Summary of Fit and Analysis of Variance displays for both analyses.
  - (b) Produce an LSMeans Plot for the Method factor in the above analysis. What information does this tell you? Compare the results with the means diamonds produced in the Fit Y By X plot.
  - (c) Fit a model of Starch Content using all three factors Size of Load, Sand Blasted, and Method. Which factors are significant?

- (d) Produce LSMeans Plots for Method and Sand Blasted and interpret them.
- (e) Use LSMeans Contrasts to determine whether alpha amalyze is significantly different from caustic soda in its effect on Starch Content.
- (f) Examine all first-level interactions of the three factors by running Fit Model and requesting **Factorial to Degree** from the Macros menu with the three factor columns selected. Which interactions are significant?
2. The sample data table Titanic Passengers.jmp contains information about the passengers of the RMS Titanic. Use JMP to answer the following questions:
- Using the Fit Model platform, find a nominal logistic model to predict survival based on passenger class, age, and sex. Use the default target level, 0.
  - To evaluate the effectiveness of your model, save the probability formula from the model to the data table. Then, using Fit Y By X, prepare a contingency table to discover the percentage of times the model made the correct prediction. (**Note:** You can also request a Confusion Matrix from the red triangle menu for the model report.)
  - Now, fit a model to predict survival based on passenger class, age, sex, and their two-way interactions. Are any of the interactions significant?
  - In the report window, select **Profiler** from the top red triangle menu. Drag the vertical lines in the profiler to explore the fitted model. The predicted survival rates appear on the left side of the profiler. Which group had the highest survival rate? The lowest?
3. The sample data table Decathlon.jmp (Perkiomäki, 1995) contains decathlon scores from several winners in various competitions. Although there are 10 events, they fall into groups of running, jumping, and throwing.
- Suppose you were going to predict scores on the 100m running event. Which of the other nine events do you think would be the best indicators of the 100m results?
  - Run a Stepwise regression (from the Fit Model Platform) with 100m as the Y and the other nine events as X. Use the default value of 0.25 probability to enter and the Forward method. Which events are included in the resulting model?
  - Complete a similar analysis on Pole Vault. Examine the events included in the model. Are you surprised?



15

## Design of Experiments

### Overview

Designed experiments are an important methodology in science and engineering. We learn by trial and error—*that is*, experimentation, and experimental design make the learning process efficient and reliable.

Design of Experiments (DOE) is probably the single most powerful technique that you can use for product and process development, refinement, problem solving, scientific inquiry, and optimization.

The DOE platform in JMP provides tools for creating designed experiments and saving them in a JMP data table. JMP supports two ways to make a designed experiment:

- Build a new design that matches the description of your engineering problem and remains within your budget for time and material. This method uses the **Custom Design** platform and is recommended for creating all experimental designs.
- Select a pre-formulated design from a list of designs. These lists of designs include **Screening**, **Response Surface**, **Full Factorial**, **Mixture**, **Space Filling**, **Taguchi**, and other specialized designs.

This chapter shows you how JMP generates and analyzes common experimental designs using the Custom Design platform. This introductory DOE chapter shows examples of screening, response surface, and split plot designs.

See the *Design of Experiments Guide* in the online Help for details about designs not covered in this chapter.

## Chapter Contents

Overview .....	421
Introduction.....	424
Key Concepts.....	424
JMP DOE .....	425
A Simple Design.....	426
The Experiment .....	426
Enter the Response and Factors .....	427
Define the Model.....	429
Is the Design Balanced? .....	432
Perform Experiment and Enter Data .....	432
Analyze the Model .....	433
Flour Paste Conclusions.....	439
Details of the Design: Confounding Structure .....	439
Using the Custom Designer .....	440
How the Custom Designer Works .....	440
Choices in the Custom Designer.....	441
An Interaction Model: The Reactor Data .....	442
Analyzing the Reactor Data.....	444
Where Do We Go from Here? .....	450
Some Routine Screening Examples .....	452
Main Effects Only (a Review) .....	452
All Two-Factor Interactions Involving a Single Factor .....	453
Alias Optimal Designs .....	455
Response Surface Designs.....	456
The Odor Experiment.....	456
Response Surface Designs in JMP.....	456
Analyzing the Odor Response Surface Design.....	458
Plotting Surface Effects.....	461
Specifying Response Surface Effects Manually .....	462
The Custom Designer versus the Response Surface Design Platform .....	463
Split-Plot Designs .....	464
The Box Corrosion Split-Plot Experiment .....	465

Designing the Experiment .....	465
Analysis of Split-Plot Designs.....	467
Design Strategies .....	471
DOE Glossary of Key Terms.....	472
Exercises.....	476

# Introduction

Experimentation is the fundamental tool of the scientific method. In an experiment, a *response* of interest is measured as some *factors* are changed systematically through a series of *runs*, also known as *trials*. Those factors that are not involved in the experiment are held (as much as is practical) at a constant level. Then any variation or effect produced by the experiment can be attributed to the changes in the design factors and natural variability. The goal is to determine whether and how the factors affect the response.

Experimental design addresses this goal—it allows us to learn the most about the relationship between the response and factors for a given number of runs. Compared to ad hoc trial and error, experimental design saves money and resources by reducing the number of necessary trials.

## Key Concepts

DOE, and statistics in general, enable us to make decisions and develop product and process knowledge in light of random variation. Experimentation is a true journey of discovery.

### **Experimentation Is Learning**

The word *experiment* has the same root as the words *expert* and *experience*. They all derive from the Latin verb *experior*, which means to try, to test, to experience, or to prove.

Typical experimentation is a familiar process of trial and error. Try something and see what happens. Learn from experience. Learn from changing factors and observing the results. This is the *inductive method* that encompasses most learning. Designed experiments, on the other hand, add to this method by giving us a framework to direct our experiences in a meaningful way.

### **Controlling Experimental Conditions Is Essential**

We are easily fooled unless we take care to control the experimental conditions and environment. This control is the critical first step in doing scientific experiments. This is the step that distinguishes experimental results from observational, happenstance data. You can obtain clues of how the world works from non-experimental data. However, you cannot put full trust into learning from observational phenomena because you are never completely sure why the response changed. Were the response differences due to changes in the factor of

interest, or some change in uncontrolled variables? The goal is to obtain knowledge of cause and effect.

### Experiments Manage Random Variation within a Statistical Framework

Experiments are never exactly repeatable because of the inevitable uncontrollable, random component of the response. Understanding how to model this random variation was one of the first triumphs of the new “statistical science” of the 1920s. Important techniques like randomization and blocking, standard statistical distributions, and tests that were developed are still in use today.

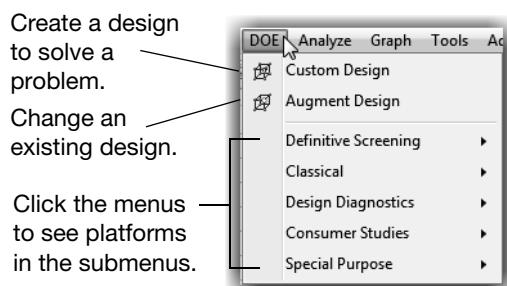
## JMP DOE

JMP has tools that enable you to produce almost any type of experimental design. Several commonly used “classical” designs are as follows:

- *Screening Designs* for scouting many factors: Screening designs examine many factors to see which factors have the greatest effect on the results of a process. To economize on the number of runs needed, each factor is usually set at only two levels, and response measurements are not taken for all possible combinations of levels. Screening designs are generally a prelude to further experiments.
- *Response Surface Designs* for optimization: Response surface experiments try to focus on the optimal values for a set of continuous factors. They are modeled with a curved surface so that the maximum point of the surface (optimal response) can be found mathematically.
- *Factorial Designs*: A complete factorial experiment includes a run for all possible combinations of factor levels.

**Note:** The designs in this chapter are commonly used in industrial settings, but there are many other designs used in other settings. The DOE facility can produce additional designs such as Mixture, Choice (for market research), Space Filling (computer simulation), Nonlinear, and more, as shown on the next page. JMP can also produce and analyze other designs, such as split plots, repeated measures, crossovers, complete and incomplete blocks, and Latin squares.

The DOE platform in JMP is an environment for describing the factors, responses, and other specifications, creating a designed experiment, and saving it in a JMP table. When you select the DOE menu, you see the list of designs shown here.



The JMP Custom Designer builds a design for your specific problem that is consistent with your resource budget. It can be used for routine factor screening, response optimization, and mixture problems. In many situations, the Custom Designer produces designs that require fewer runs than the corresponding classical design. Also, the Custom Designer can find designs for special conditions not covered in the lists of predefined classical and specialized designs.

## A Simple Design

The following example helps acquaint you with the design and analysis capabilities of JMP.

### The Experiment

Acme Piñata Corporation discovered that its piñatas were too easily broken. The company wants to perform experiments to discover what factors might be important for the peeling strength of flour paste.

In this design, we explore nine factors in an attempt to discover the following:

- which factors actually affect the peel strength, and
- what settings should those factors take in order to optimize the peel strength?

### The Response

Strength refers to how well two pieces of paper that are glued together resist being peeled apart.

### The Factors

Batches of flour paste were prepared to determine the effect of the following nine factors on peeling strength:

- Liquid: 4 teaspoons of liquid or 5 teaspoons of liquid (4 or 5)
- Sugar: formula contained no sugar or 1/4 teaspoon of sugar (0 or 0.25)
- Flour: 1/8 cup of white unbleached flour or 1/8 cup of whole wheat flour (White or Wheat)
- Sifted: flour was not sifted or was sifted (No or Yes)
- Type: water-based paste or milk-based paste (Water or Milk)
- Temp: mixed when liquid was cool or when liquid was warm (Cool or Warm)
- Salt: formula had no salt or a dash of salt (No or Yes)
- Clamp: pasted pieces were loosely clamped or tightly clamped together during drying (Loose or Tight)
- Coat: whether the amount of paste applied was thin or thick (Thin or Thick)

### The Budget

There are many constraints that can be put on a design, arising from many different causes. In this case, we are allotted only enough time to complete 16 runs.

## Enter the Response and Factors

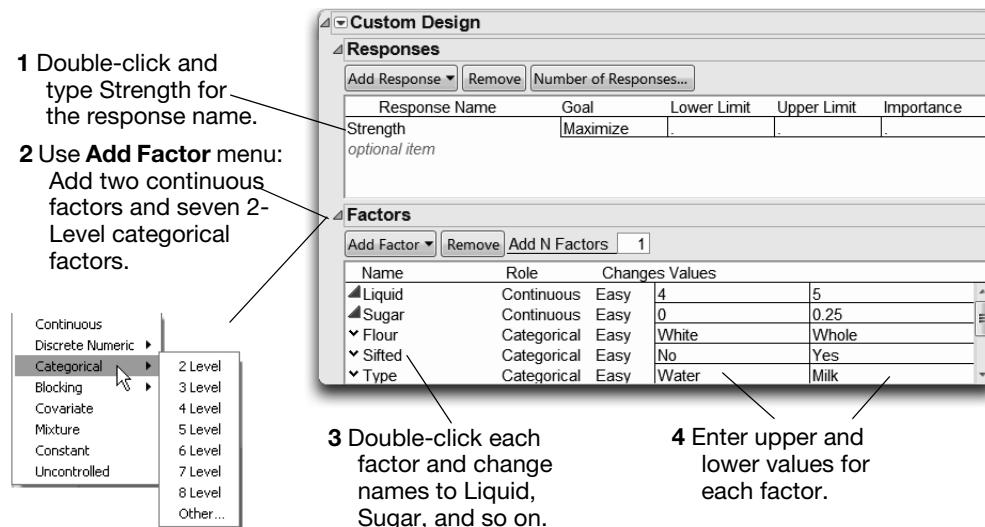
Use the Custom Design platform to define and name the response and the factors of the experiment.

⇨ Select **DOE > Custom Design**.

You see one default response called Y in the **Responses** panel.

⇨ Double-click the Y and change it to Strength.

Your Responses panel should look like the one showing at the top in **Figure 15.1**.

**Figure 15.1** Window for Designating Responses and Factors

Now, complete the following steps to add the nine factors:

- ⓐ Click the **Add Factor** button and select **Continuous** to add 1 continuous factor. Repeat to add a second continuous factor.
- ⓑ Type 7 into the Add N Factors field and select **Add Factor > Categorical > 2 Level**.
- ⓒ Double-click each factor and enter its name.
- ⓓ Add the appropriate values to specify the levels of each factor. For example, Liquid can take levels 4 and 5, Sugar is 0 and 0.25, and Flour is White and Whole. Continue this for each of the nine factors.
- ⓔ Click **Continue**.

Your **Factors** panel should look like the one in **Figure 15.2**.

**Figure 15.2** Factors Panel

Factors				
		Add Factor ▾	Remove	Add N Factors
Name	Role	Changes	Values	
▲ Liquid	Continuous	Easy	4	5
▲ Sugar	Continuous	Easy	0	0.25
▼ Flour	Categorical	Easy	White	Whole
▼ Sifted	Categorical	Easy	No	Yes
▼ Type	Categorical	Easy	Water	Milk
▼ Temp	Categorical	Easy	Cool	Warm
▼ Salt	Categorical	Easy	No	Yes
▼ Clamp	Categorical	Easy	Loose	Tight
▼ Coat	Categorical	Easy	Thin	Thick

## Define the Model

After you click Continue, you will see new panels, including **Define Factor Constraints** and **Model** as shown in **Figure 15.3**. There are no mathematical constraints on allowable factor level combinations, so there is no need to use the **Define Factor Constraints** panel (it is minimized by default).

Look at the **Model** panel. By default, the model contains the intercept and all the main effects. Other model terms, like interactions and powers, could be entered at this point. We examine only main effects in this first example (typical of screening designs), so there is no need to modify the **Model** panel.

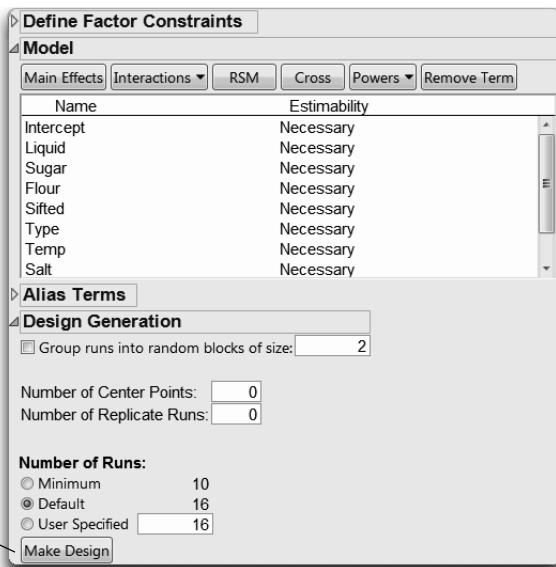
The **Design Generation** panel enables you to specify the number of experimental runs (or *trials*). JMP has several suggestions to select from.

- **Minimum** is the smallest number of runs that allow estimation of the main effects. The minimum is always the number of effects plus one.
  - **Default** is a suggestion that is larger than minimum but smaller than a full factorial design, often giving balanced designs using an economical number of runs.
  - Depending on your budget, you might be able to afford a different number of runs than those suggested. The **Number of Runs** box accepts any number of runs. You are in no way limited to the suggestions that JMP makes.
- ☞ Make sure the **Default** selection of 16 runs is selected and then click the **Make Design** button (see **Figure 15.3**).

**Figure 15.3** Model Definition Panel

Main Effects model,  
typical of a screening  
design

Click **Make Design**  
to continue.



JMP now searches for an optimal design. Once completed, the design shows in the Design outline node (**Figure 15.4**).

**Note:** Your design might not look like the one below because JMP optimizes the design based on a mathematical criterion. There are often several equivalent designs. For example, in this case there are 87 different equivalent designs.

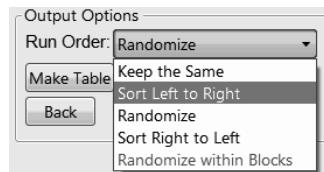
**Figure 15.4** An Optimal Main Effects Design

Run	Liquid	Sugar	Flour	Sifted	Type	Temp	Salt	Clamp	Coat
1	4	0.25	White	No	Water	Warm	Yes	Tight	Thick
2	5	0.25	Whole	Yes	Water	Cool	Yes	Loose	Thick
3	4	0	White	Yes	Water	Cool	Yes	Loose	Thin
4	5	0	Whole	No	Water	Cool	No	Tight	Thick
5	4	0.25	Whole	Yes	Milk	Cool	Yes	Tight	Thick
6	4	0	Whole	Yes	Milk	Warm	No	Loose	Thick
7	5	0	Whole	No	Water	Warm	Yes	Tight	Thin
8	5	0.25	Whole	Yes	Water	Warm	No	Loose	Thin
9	5	0.25	White	No	Milk	Cool	Yes	Loose	Thin
10	5	0	White	No	Milk	Cool	No	Loose	Thick
11	4	0	Whole	No	Milk	Warm	Yes	Loose	Thin
12	4	0.25	Whole	No	Milk	Cool	No	Tight	Thin
13	5	0	White	Yes	Milk	Warm	Yes	Tight	Thick
14	5	0.25	White	Yes	Milk	Warm	No	Tight	Thin
15	4	0	White	Yes	Water	Cool	No	Tight	Thin
16	4	0.25	White	No	Water	Warm	No	Loose	Thick

**Design Evaluation**

- Output Options  
Run Order: Randomize

JMP is ready to create the data table. In most designs, it is desirable to randomize the order of the trials. For this illustration, we sort the run order to better see the design. Select **Sort Left to Right** from the **Run Order** menu, as shown here. Then, click **Make Table**.



JMP generates a data table with several convenient features.

- All the factors have an equal number of trials at each setting.
- The response column is present, waiting for the results of your experiment.
- A note is present (upper left corner) that shows the type of design.
- Limits and coding information are stored as column properties for each variable.
- A **Model** script holds all the information about the requested model. The **Fit Model** window looks for this script and completes itself automatically based on the script's contents.

**Figure 15.5** Flour Data Table

	Liquid	Sugar	Flour	Sifted	Type	Temp	Salt	Clamp	Coat	Strength
1	4	0	White	Yes	Water	Warm	No	Tight	Thick	*
2	4	0	White	Yes	Milk	Warm	Yes	Tight	Thin	*
3	4	0	Whole	No	Water	Cool	No	Loose	Thick	*
4	4	0	Whole	No	Milk	Cool	Yes	Loose	Thin	*
5	4	0.25	White	No	Water	Cool	Yes	Tight	Thick	*
6	4	0.25	White	No	Milk	Cool	No	Tight	Thin	*
7	4	0.25	Whole	Yes	Water	Warm	Yes	Loose	Thick	*
8	4	0.25	Whole	Yes	Milk	Warm	No	Loose	Thin	*
9	5	0	White	No	Water	Warm	No	Loose	Thin	*
10	5	0	White	No	Milk	Warm	Yes	Loose	Thick	*
11	5	0	Whole	Yes	Water	Cool	No	Tight	Thin	*
12	5	0	Whole	Yes	Milk	Cool	Yes	Tight	Thick	*
13	5	0.25	White	Yes	Water	Cool	Yes	Loose	Thin	*
14	5	0.25	White	Yes	Milk	Cool	No	Loose	Thick	*
15	5	0.25	Whole	No	Water	Warm	Yes	Tight	Thin	*
16	5	0.25	Whole	No	Milk	Warm	No	Tight	Thick	*

Asterisk indicates information stored with column.

Each run has its factor settings as values.

Response column ready for data

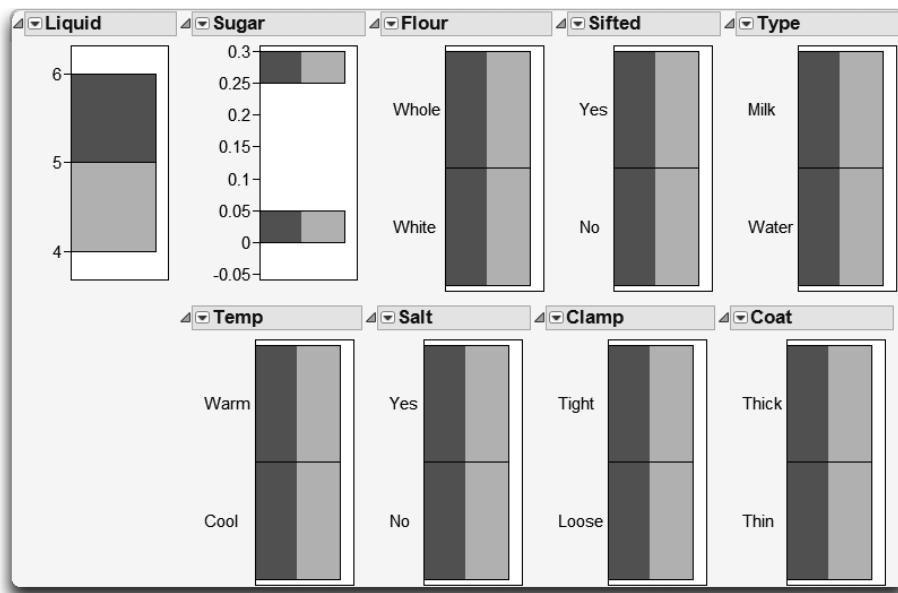
- ☞ Now, if this is the final design, use **File > Save As** to save the table. Note: You need this table again later.

## Is the Design Balanced?

It is easy to show that the Custom designer produces balanced designs when possible. To check that this design is balanced:

- ☛ Select **Analyze > Distribution** for all nine factor variables.
- ☛ Select the **Histograms Only** box, on the lower left corner of the launch window, and click **OK**.
- ☛ Select **Arrange in Rows** from the red triangle menu next to Distributions, type 5 in the box provided, and click **OK**.
- ☛ Click in each histogram bar to see that the highlighted distribution is flat for all the other variables, as illustrated in **Figure 15.6**.

**Figure 15.6** Histograms Verify That the Design Is Balanced



## Perform Experiment and Enter Data

At this point, you run the 16 trials of the experiment, noting the strength readings for each trial, and then entering these results into the Strength column. The experiment shown in **Figure 15.5**, along with the results shown here, have been saved in the Flrpaste.jmp sample data table in the Sample Data Library.

- ☞ Select **Help > Sample Data Library** and open Flrpaste.jmp.

### Examine the Response Data

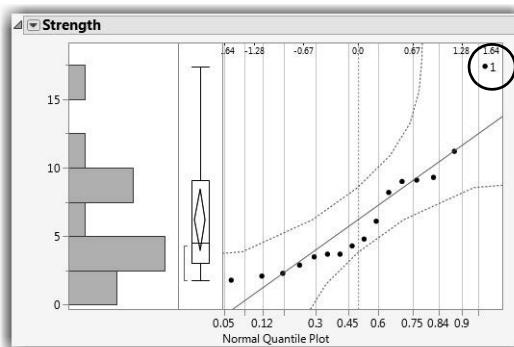
As usual, a good place to start is by examining the distribution of the response, which in our example is the peel strength.

- ☞ Select **Analyze > Distribution**.
- ☞ Assign Strength to **Y, Columns** and then click **OK**.
- ☞ When the histogram appears, select **Normal Quantile Plot** from the red triangle menu next to Strength.

Strength
17.4
8.2
9
1.8
9.3
6.1
4.8
2.9
9.1
4.3
3.5
2.3
11.2
3.7
3.7
2.1

You should see the plots shown here.

The box plot and the normal quantile plot are useful for identifying runs that have extreme values. In this case, run 1 has a peel strength that appears higher than the others.

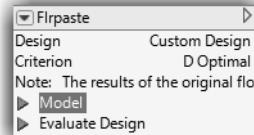


### Analyze the Model

Of the nine factors in the flour paste experiment, there might be only a few that stand out in comparison with the others. The goal of this experiment is to find out which of the factors are most important, and to identify the factor combinations that optimize the predicted response (peel strength). This type of experimental situation lends itself to an effect screening analysis.

When an experiment is designed in JMP, a Model script is automatically saved in the Tables panel of the design data table.

- ☞ Click the green arrow next to the Model script to fit the model (or, select **Analyze > Fit Model**).

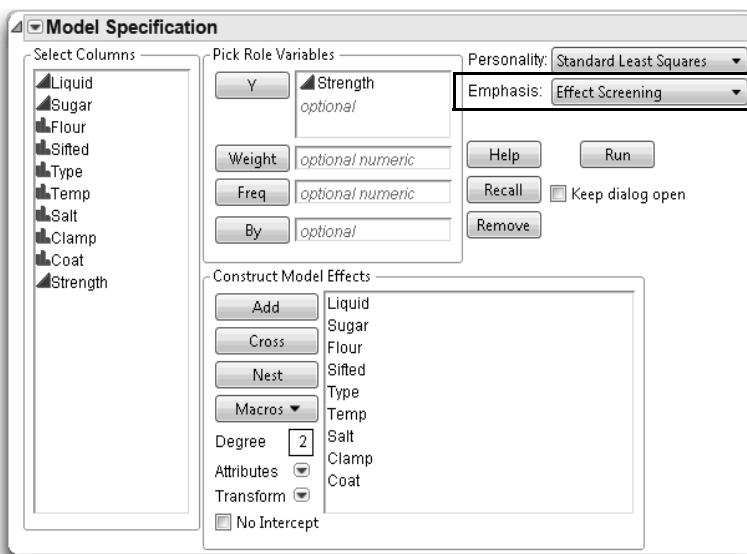


- ⌚ The Model Specification window appears, with Strength as the response (Y) variable and all of the factor columns as effects in the model. Note that **Effect Screening** is selected on the Emphasis menu.

**Figure 15.7** shows the completed window.

- ⌚ Click **Run** to see the analysis results shown in **Figure 15.8**. Again, note that some default output is not displayed in the figure.

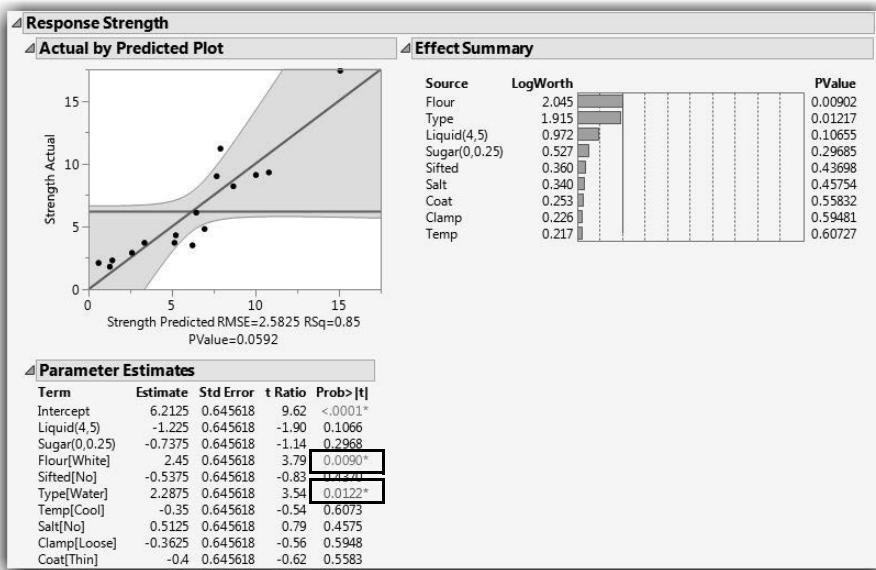
**Figure 15.7** Model Specification Window for Flour Paste Main Effects Analysis



A summary of whole model fit statistics is provided at the bottom of the Actual by Predicted plot. The p-value shows that the model as a whole is borderline significant ( $p = 0.0592$ ). The standard deviation of the error (Root Mean Square Error) is estimated as 2.58, a high value relative to the scale of the response. The  $R^2$  of 0.85 tells us that our model explains a decent amount of the variation in peel strength values.

**Note:** Additional whole model fit statistics are reported in the Summary of Fit and Analysis of Variance regression reports. To request these reports, select the options from the Regression Reports red triangle menu next to Response.

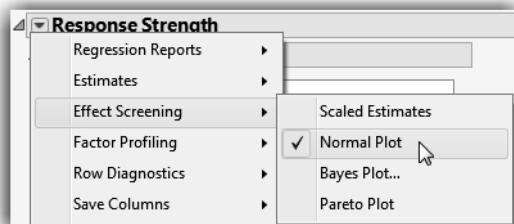
The most significant factors, shown in the Effects Summary and Parameter Estimates tables, are Flour and Type (of liquid).

**Figure 15.8** Flour Paste Main Effects Analysis Output

### Continuing the Analysis with the Normal Plot

Commands in the red triangle menu on the Response Strength title bar give you many analysis and display options. The normal plot of parameter estimates is commonly used in screening experiments to understand which effects are most important. The normal plot is about the only way to evaluate the results of a *saturated design* with no degrees of freedom for estimating the error, which is common in screening experiments.

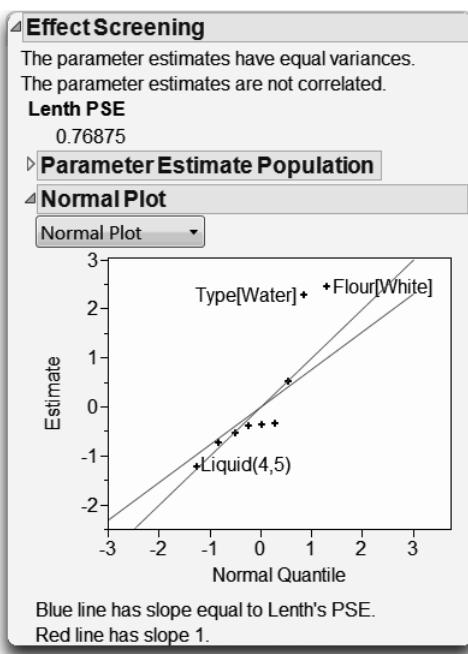
- From the red triangle menu next to Response Strength, select **Effect Screening > Normal Plot** to display the normal plot shown in **Figure 15.9**.



The *normal plot* is a normal quantile plot (Daniel 1959), which shows the parameter estimates on the vertical axis and the normal quantiles on the horizontal axis. In a screening experiment, you expect most of the effects to be inactive, to have little or no effect on the response. If that is true, then the estimates for those effects are a realization of random noise centered at zero. Normal plots give a sense of the magnitude of an

effect that you should expect when it is truly active, rather than just noise. The active effects appear as labeled outliers.

**Figure 15.9** Normal Plot Shows Most Influential Effects



If labels overlap,  
click the labels  
and drag to make  
them visible.

The normal plot shows a straight line with slope equal to the *Lenth's PSE* (pseudo standard error) estimate (Lenth 1989). Lenth's PSE is formed as follows:

Take 1.5 times the median absolute value of the estimates after removing all the estimates greater than 2.75 times the median absolute estimate in the complete set.

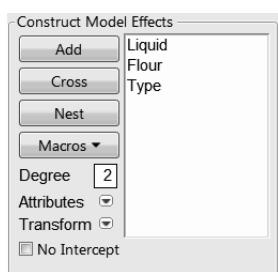
Lenth's PSE is computed using the normalized estimates. It disregards the intercept. Effects that deviate substantially from this normal line are automatically labeled on the plot.

Usually, most effects in a screening analysis have small values; a few have larger values. In the flour paste example, Type and Flour separate from the normal lines more than would be expected from a normal distribution of the estimates. Liquid falls close to the normal lines, but is also labeled as a potentially important factor.

### Visualizing the Results with the Prediction Profiler

To better visualize the most important factors, and how they effect Strength, we look at the Prediction Profiler. First, we remove the other effects from the model and rerun the analysis.

- ✓ Either select **Analyze > Fit Model**, or select **Model Dialog** from the top red triangle menu on the report to see the Model Specification window.



- ✓ Remove all effects except Liquid, Flour, and Type, as shown here.
- ✓ Click **Run** and then scroll to the bottom of the report.

**Note:** You can also remove effects directly in the Fit Least Squares window using the **Remove** button on the Effect Summary table.

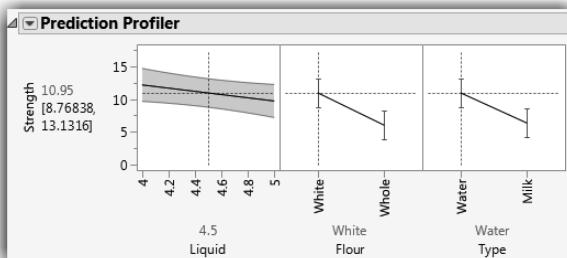
- ✓ From the red triangle menu next to Prediction Profiler, select **Optimization and Desirability** and deselect **Desirability Functions**. We'll discuss optimization later in this chapter.

The Prediction Profiler shows the predicted response for different factor settings.

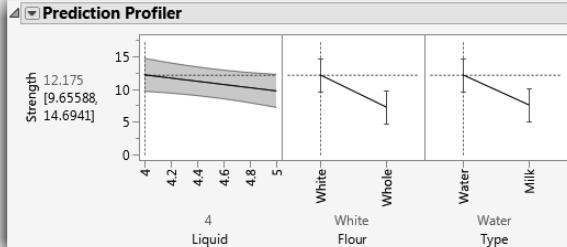
**Figure 15.10** shows three manipulations of the Prediction Profiler for the flour experiment.

**Figure 15.10** Screening Model Prediction Profiler for Flour Paste Experiment

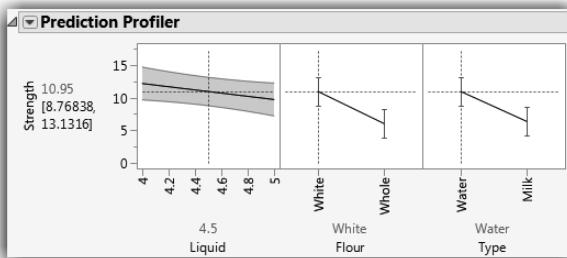
All factors at the initial settings



All factors at the low settings



All factors at the high settings



The settings of each factor are connected by a line, called the *prediction trace* or *effect trace*. You can grab and move each vertical dotted line in the Prediction Profile plots to change the factor settings. The predicted response automatically recomputes and shows on the vertical axis, and the prediction traces are redrawn. The Prediction Profiler lets you look at the effect on the predicted response of changing one factor setting while holding the other factor settings constant. It can be useful for judging the importance of the factors.

The effect traces in the plot at the top of **Figure 15.10** show higher predicted strength when there is the following:

- four teaspoons of liquid rather than five
- white flour instead of whole wheat flour
- water rather than milk

This indicates that changing the three effects to these values increases peel strength.

The second plot in **Figure 15.10** shows what happens if you click and move the effect trace to the settings listed above: four teaspoons of water and white flour. The predicted response changes from 10.95 to 12.175.

If you had the opposite settings for these factors (bottom plot in **Figure 15.10**), then the linear model would predict a surface strength of 0.25.

## Flour Paste Conclusions

We designed and conducted a 16-run screening experiment to explore the effect of nine factors on peel strength. Although the initial Whole-Model *F*-Test was marginally significant ( $p = 0.0592$ ), we found three important factors: Liquid, Flour, and Type. We then used the prediction profiler to explore the optimal settings for these three factors.

Since the unexplained variation was relatively high, the experiment might bear repeating with better control of variability, a better experimental procedure, and an improved measurement system.

## Details of the Design: Confounding Structure

There are a few details of this design that we did not discuss during the analysis. For example, how the nine factors plus an intercept can be investigated with only 16 runs. Since this experiment needs to look at many factors in very few runs, many main effects are confounded with two-way interactions.

Confounding, or aliasing, means that the estimate of one effect includes the influence of one or more other effects. We don't have enough runs (information) to uniquely estimate these effects. In main effects screening experiments, we assume that there are no interactions, and we do not include interactions in the model. However, it is informative to look at the alias structure of the design, which tells which effects are aliased or partially aliased with other effects. To do this:

- ⇨ Select **DOE > Design Diagnostics > Evaluate Design**.
- ⇨ Specify Strength as **Y, Response**.
- ⇨ Specify the factors as **X, Factor**, and click **OK**.

**Note:** You can also run the DOE Dialog Evaluate Design script in the Table panel of your design data table.

☛ Open the **Alias Matrix** outline under Design Evaluation.

**Figure 15.11** Alias Matrix

Effect	Liquid*Sugar	Liquid*Flour	Liquid*Sifted	Liquid>Type
Intercept	0	0	0	0
Liquid	0	0	0	0
Sugar	0	0	0	0
Flour	0	0	0	0
Sifted	0	0	0	0
Type	0	0	0	0
Temp	0	0	-1	0
Salt	0	0	0	0
Clamp	0	1	0	0
Coat	0	0	0	1

Main effects are aliased with the two-way interactions.

In **Figure 15.11**, you can see that the Clamp effect is aliased with at least one second-order interaction. See the JMP *Design of Experiments Guide* for more information about interpreting the Alias Matrix.

Luckily, you do not have to understand the details of aliasing in order to use the Custom Designer.

## Using the Custom Designer

There is a **Back** button at several stages in the design window that enables you to go back to a previous step and modify the design. For example, you can modify a design by adding or removing quadratic terms, center points, or replicated runs.

### How the Custom Designer Works

The Custom Designer starts with a random design where each point is inside the range of each factor. The computational method is an iterative algorithm called *coordinate exchange*. Each iteration of the algorithm involves testing every value of each factor in the design to determine whether replacing that value increases the optimality criterion. If so, the new value replaces the old. This process continues until no replacement occurs in an entire iteration.

To avoid converging to a local optimum, the whole process is repeated several times using a different random start. The Custom Designer displays the most optimal of these designs.

Sometimes a design problem can have several equivalent solutions. Equivalent solutions are designs having equal precision for estimating the model coefficients as a group. When this is true, the design algorithm generates different (but equivalent) designs if you click the **Back** and **Make Design** buttons repeatedly.

## Choices in the Custom Designer

Custom designs give the most flexibility of all design choices. The Custom Designer gives you the following options:

- continuous factors
- discrete numeric factors with arbitrary numbers of levels
- categorical factors with arbitrary numbers of levels
- blocking with arbitrary numbers of runs per block
- covariates (factors that already have unchangeable values and design around them)
- mixture ingredients
- inequality constraints on the factors, including disallowed combinations
- interaction terms
- polynomial terms for continuous factors
- selecting factors (or combinations of factors) whose parameters are estimated only if possible
- choice of number of center points and replicated runs
- choice of number of experimental runs to do, which can be any number greater than or equal to the number of terms in the model
- choice of optimality criteria

After specifying all of your requirements, the Custom Designer generates an appropriate optimal design for those requirements. In cases where a classical design (such as a factorial) is optimal, the Custom Designer finds it. Therefore, the Custom Designer can serve any number or combination of factors. For a complete discussion of optimal designs, optimality criteria, and design evaluation, refer to the *Design of Experiments Guide*.

## An Interaction Model: The Reactor Data

Now that you've seen the basic flow of experimental design and analysis in JMP, we illustrate a more complicated design that involves both main effects and second-order interactions.

Box, Hunter, and Hunter (2005, p. 259) discuss a study of chemical reactors that has five two-level continuous factors, Feed Rate, Catalyst, Stir Rate, Temperature, and Concentration. The purpose of the study is to find the best combination of settings for optimal reactor output, measured as percent reacted. It is also known that there might be interactions among the factors.

A full factorial for five factors requires  $2^5 = 32$  runs. Since we're only interested in main effects and second-order interactions, a smaller screening design can be used. The **Screening Design** option on the **DOE > Classical** menu provides a selection of fractional factorial and Plackett-Burman designs. The 16-run fractional factorial design allows us to estimate all 5 main effects and 10 two-factor interactions. Unfortunately, this leaves no runs for the estimation of the error variance.

So, we'll design this experiment using the Custom Designer, which provides more flexibility in choosing the number of runs than traditional screening designs. For fitting the model with all the two-factor interactions, by default the Custom Designer produces a design with 20 runs. This gives us four runs to estimate the error variance, which enables us to perform more rigorous significance tests of the various effects.

- ☞ Select **DOE > Custom Design** and specify the response and factor settings as shown at the top of **Figure 15.12**.
- ☞ Click **Continue**.

**Figure 15.12** Add Response and Factors for Reactor Experiment

The screenshot shows the 'Custom Design' dialog box. The 'Responses' section contains a table with one row: 'Percent Reacted' with 'Maximize' goal. The 'Factors' section contains a table with five rows: Feed Rate, Catalyst, Stir Rate, Temperature, and Concentration, all set to 'Continuous' role and 'Easy' changes.

Response Name	Goal	Lower Limit	Upper Limit	Importance
Percent Reacted	Maximize			

Name	Role	Changes	Values
Feed Rate	Continuous	Easy	10 15
Catalyst	Continuous	Easy	1 2
Stir Rate	Continuous	Easy	100 120
Temperature	Continuous	Easy	140 180
Concentration	Continuous	Easy	3 6

- Click **Interactions > 2nd** to add all second-order interactions. Under **Number of Runs**, the default changes to 20.

**Figure 15.13** Designing the Reactor Experiment

The screenshot shows the 'Model' dialog box. The 'Interactions' tab is selected, showing a dropdown menu with '2nd' highlighted. The 'Design Generation' section includes options for grouping runs into random blocks of size 2, specifying 0 center points, and 0 replicate runs. The 'Number of Runs:' section shows 'Default' selected with a value of 20, and 'User Specified' also set to 20. A 'Make Design' button is at the bottom.

Name	2nd	Estimability
Intercept	3rd	Necessary
Feed Rate	4th	Necessary
Catalyst	5th	Necessary
Stir Rate		Necessary
Temperature		Necessary
Concentration		Necessary
Feed Rate*Catalyst		Necessary
Feed Rate*Stir Rate		Necessary

**Design Generation**

Group runs into random blocks of size:

Number of Center Points:

Number of Replicate Runs:

**Number of Runs:**

Minimum 16

Default 20

User Specified

**Make Design**

- Click **Make Design** to compute the design, and then click **Make Table** to generate the design data table.

The design generated by JMP and the results of this experiment are in the sample data table called Reactor 20 Custom.jmp. Note that the design that you generated might be different.

- ✓ Select **Help > Sample Data Library** and open Design Experiment/Reactor 20 Custom.jmp (shown in **Figure 15.14**).

**Figure 15.14** Design Table and Data for Reactor Example

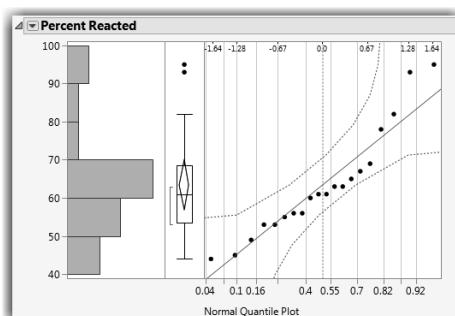
	Feed Rate	Catalyst	Stir Rate	Temperature	Concentration	Percent Reacted
1	10	1	100	140	6	56
2	10	1	100	180	3	69
3	10	1	100	180	6	44
4	10	1	120	140	3	53
5	10	1	120	180	6	49
6	10	2	100	140	3	63
7	10	2	100	180	6	78
8	10	2	120	140	6	67
9	10	2	120	180	3	95
10	15	1	100	140	3	53
11	15	1	100	140	6	63
12	15	1	100	180	6	45
13	15	1	120	140	3	56
14	15	1	120	140	6	55
15	15	1	120	180	3	60
All rows	20					
Selected	0					
Excluded	0					
Hidden	0					
Labelled	0					
	15	2	100	140	3	61
	17	15	2	100	6	65
	18	15	2	100	3	93
	19	15	2	120	3	61
	20	15	2	120	6	82

## Analyzing the Reactor Data

We begin the analysis with a quick look at the response data.

- ✓ Select **Analyze > Distribution** to look at the distribution of the response variable Percent Reacted.
- ✓ Select **Normal Quantile Plot** from the red triangle menu next to Percent Reacted to see the normal quantile plot shown here.

There do not appear to be any unusual observations, although some trials resulted in high Percent Reacted (this is good!).

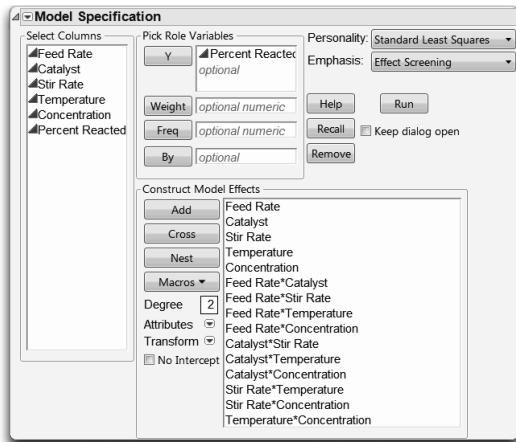


To analyze this experiment, we use the Fit Model platform. Since the design was generated using JMP, a Model script for the analysis is saved in the Tables panel.

- ⌚ Click the green triangle next to the Model script, or select **Analyze > Fit Model**.

We designed this experiment to estimate all main effects and two-factor interactions. So, the Model Specification window is set up to estimate these effects.

- ⌚ Confirm that Percent Reacted is **Y**, that the five main effects and 10 two-way interactions are listed as model effects, and that the Emphasis is **Effect Screening**.

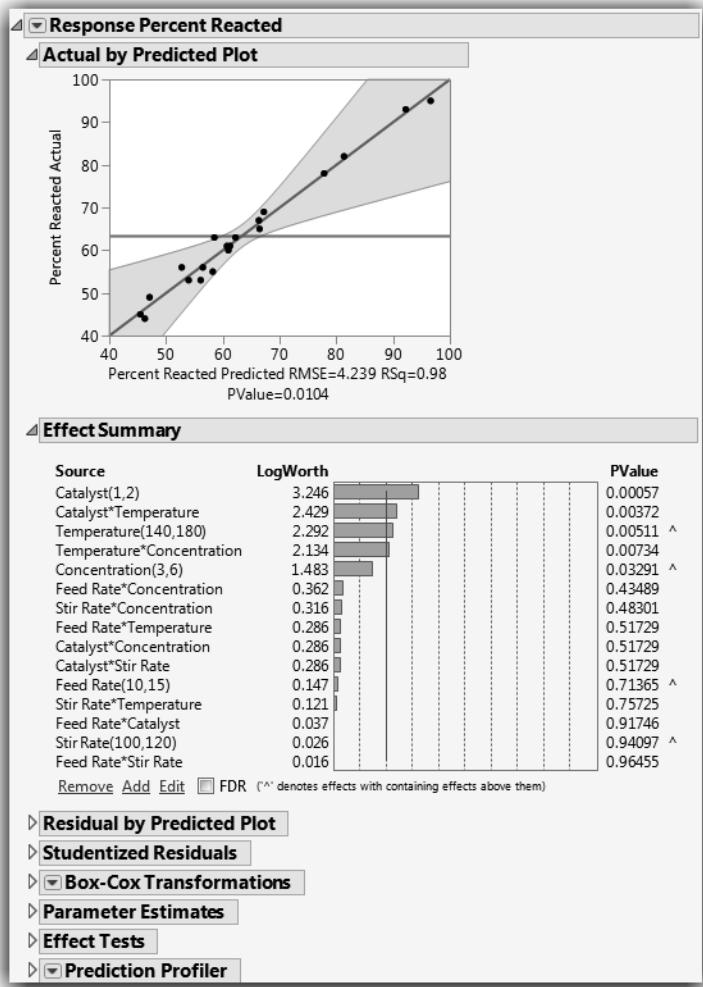


- ⌚ Click **Run** to see the results in **Figure 15.15**.

As we saw with the Flour Paste example, JMP produces an Actual by Predicted Plot and a variety of statistical tables.

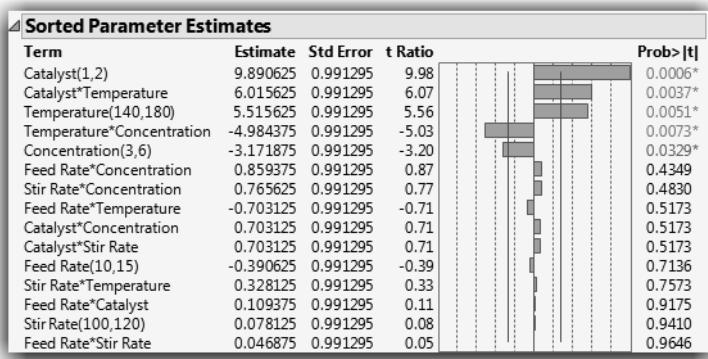
The results summarized below the Actual by Predicted Plot indicate that the whole model is significant, with a  $p$ -value of 0.0104. The Root Mean Square Error is 4.239. This value is relatively low given the scale of the response, and the  $R^2$  is a healthy 98%.

**Note:** To see the Summary of Fit and Whole Model tests results, select the options from the Regression Reports red triangle menu next to Response Percent Reacted.

**Figure 15.15** Reactor Analysis Output

The Effect Summary table sorts the effects in descending order of significance. As we can see, there are several significant terms. Catalyst ( $p$ -value = 0.00057), and the Catalyst\*Temperature interaction ( $p$ -value = 0.00372) are both highly significant. Temperature, the Temperature\*Concentration interaction, and Concentration are also significant (at the 0.05 level).

The parameter estimates for the model are shown in the Parameter Estimates table. To display these estimates in descending order of significance, select **Estimates > Sorted Estimates** from the red triangle menu next to Response Percent Reacted. The Sorted Parameters Estimates table is shown in **Figure 15.16**.

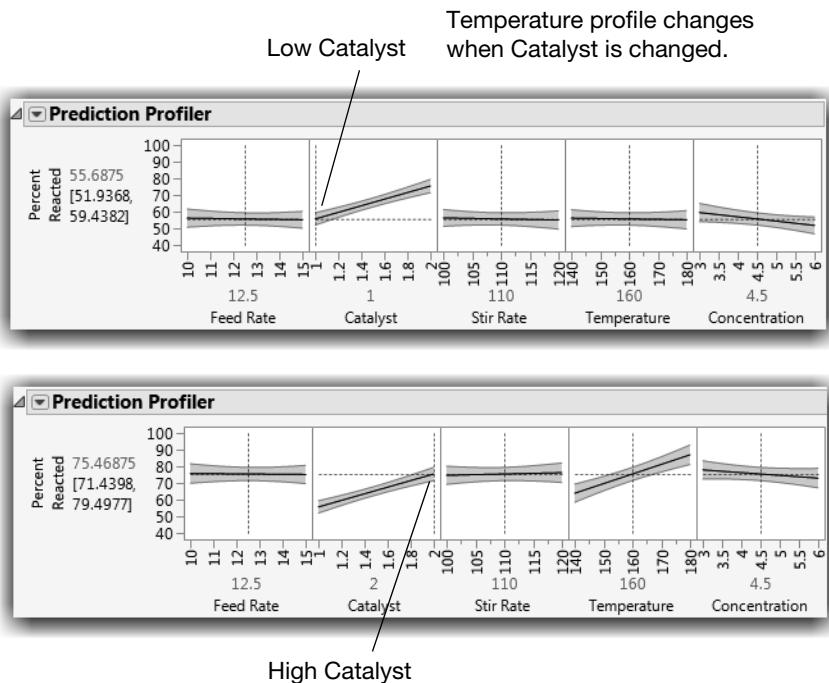
**Figure 15.16** Reactor Sorted Parameter Estimates

Let's look more closely at the interaction of Catalyst and Temperature using the Prediction Profiler, which appears at the bottom of the report. Select **Factor Profiling > Profiler** from the red triangle menu next to Response Percent Reacted if the profiler isn't displayed.

- ❖ For now, let's hide the Desirability plots at the bottom of the profiler. Select **Optimization and Desirability > Desirability Functions** in the red triangle menu next to Prediction Profiler.
- ❖ In the profiler, change the setting of Catalyst from 1 to 2 repeatedly (drag the vertical red line back and forth, or click at either end of the profile trace).

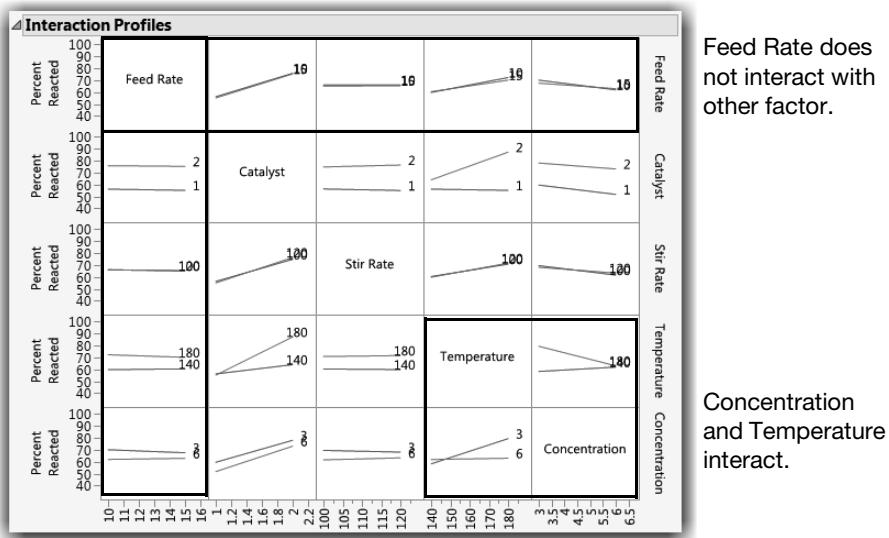
Watch the slope of the profile for Temperature as you change the setting of Catalyst. The slope changes dramatically as catalyst is changed, which indicates an interaction. When there is no interaction, only the heights of the profile should change, not the slope. For example, watch the profile for Stir Rate, which does not interact with Catalyst. The profile height changes as the level of Catalyst is changed, but the slope stays relatively constant.

The Prediction Profiler at the top of **Figure 15.17** shows the response when catalyst is at its low setting. The lower set of plots show what happens when catalyst is at its higher setting.

**Figure 15.17** Effect of Changing Temperature Levels

This slope change can be seen for all interactions in one picture as follows:

- ☞ Select **Factor Profiling > Interaction Plots** from the red triangle menu next to Response Percent Reacted to display the plot in **Figure 15.18**.

**Figure 15.18** Interaction Plots for Five-Factor Reactor Experiment

These profile plots show the interactions involving all pairs of variables in the reactor experiment.

In an interaction plot, the  $y$ -axes are the response. Each small plot shows the effect of two factors on the response. One factor (labeled in the matrix of plots) is on the  $x$ -axis. This factor's effect is displayed as the slope of the lines in the plot. The other factor, labeled on the right  $y$ -axis, becomes multiple prediction profiles (lines) as it varies from low to high. This factor shows its effect on the response as the vertical separation of the profile lines. If there is an interaction, then the slopes are different for different profile lines, like those in the Temperature by Concentration plot (the lines are non-parallel).

Recall that Temperature also interacted with Catalyst. This is evident by the different slopes showing in the Temperature by Catalyst interaction plots. On the other hand, Feed Rate did not show in any significant interactions. See the  $p$ -values in the Parameter Estimates table.) The lines for all of the interactions involving Feed Rate are nearly parallel.

**Note:** The lines of a cell in the interaction plot are dotted when there is no corresponding interaction term in the model (since we've included all interactions, all of the lines are solid).

## Where Do We Go from Here?

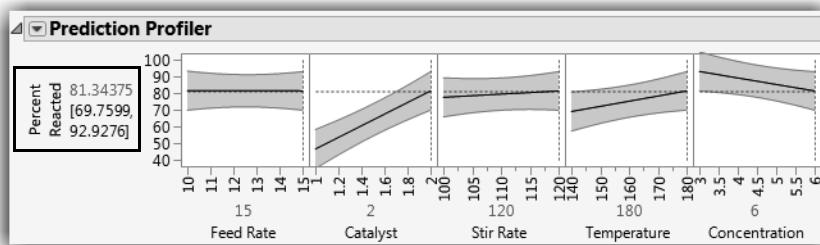
Recall that the purpose of the study was to find the best combination of settings for optimal reactor output, measured as percent reacted. We've built a linear model for percent reacted, as a function of all five main effects and all two-factor interactions. We've found many significant effects, including important two-factor interactions.

### Finding the Best Settings

As we saw earlier, we can change factor settings in the Prediction Profiler to see corresponding changes in the predicted response. Changing all of the factors to their high levels results in a predicted Percent Reacted of 81.34 (see **Figure 15.19**). Hopefully, we can do much better than this!

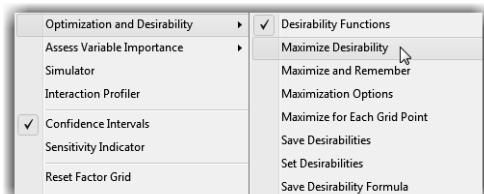
The bracketed numbers, 69.76 and 92.93, provide a confidence interval for the predicted response.

**Figure 15.19** Percent Reacted when Factors Are at High Levels



We could change factor settings, one at a time, to see whether we can get a better predicted response. But, the presence of two-factor interactions makes it difficult to determine the optimal settings manually. Fortunately, JMP has a built-in optimizer that facilitates this process for us.

- ⌚ Select **Optimization and Desirability > Desirability Functions** from the red triangle menu next to Prediction Profiler to open the Desirability plots. **(Note:** These plots might be displayed by default.)
  
- ⌚ Select **Optimization and Desirability > Maximize Desirability** from the same red triangle menu.



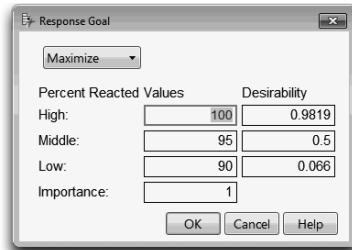
Recall that, when we designed the experiment, we set the response goal to maximize Percent Reacted. By selecting **Maximize Desirability**, JMP finds factor settings that maximize the response goal.

**Note:** To change the response goal, double-click in the last box in the top row of Prediction Profiler. This opens the Response Goal panel (shown here).

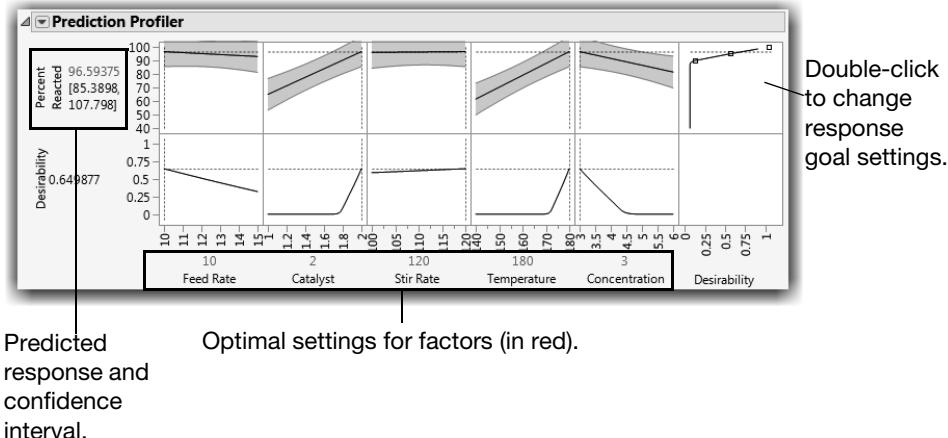
The factor settings that produce the maximum predicted response are displayed at the bottom of the Profiler. These settings are as follows:

- Feed Rate - 10
- Catalyst - 2
- Stir Rate - 120
- Temperature - 180
- Concentration - 3

The predicted Percent Reacted, at these settings, is 96.59.



**Figure 15.20** Optimal Settings for the Reactor Factors



### Validate the Results

Can we really see nearly 97% yields if we implement these factor settings? Can we do better? Before making any process changes, it's always a good idea to validate your experimental findings. One strategy is to use "check points," trials that were not included in the analysis, to see how well the model predicts those points. Another strategy is to use a small number of confirmation runs to test the factor settings on a small scale. In either case, if our validation produces results inconsistent with our model, we have more work to do.

### Reduce the Model First?

We found optimal settings using the full model. However, many of the terms in the model are not significant. Instead of optimizing on the full model, we might first reduce the model by eliminating nonsignificant terms. Model reduction involves returning to the Model Specification window, eliminating interactions with the highest  $p$ -values, rerunning the analysis, and repeating until only significant effects remain.

If you're curious about this experiment, try reducing the model to see whether the prediction results change.

**Note:** The Reactor 20 Custom.jmp experiment is a subset of the Reactor 32 Runs.jmp  $2^5$  full factorial experiment. As we will see in an exercise, the efficient 20-run design generated from the Custom Designer produces results similar to the much larger 32-run full factorial experiment.

## Some Routine Screening Examples

The Flour Paste and Reactor experiments were examples of screening designs. In one case (Flour Paste) we were interested in screening for main effects. In the other (Reactor), we estimated only main effects and second-order interactions. This section gives short examples showing how to use the Custom Designer to generate specific types of screening designs.

### Main Effects Only (a Review)

- ❖ Select **DOE > Custom Design**.
- ❖ Enter the number of factors you want (six for this example) into the Factors panel, select **Add Factor > Continuous**, and then click **Continue**.

Because there are no higher order terms in the model, no further action is needed in the Model panel. The default number of runs (12) enables us to estimate the intercept and the six main effects, with five degrees of freedom for error.

- ⌚ Click **Make Design** to see the Design table in **Figure 15.21**.

The result is a resolution-three screening design. All main effects are estimable but are confounded with two-factor interactions (see the Alias Matrix on the bottom of **Figure 15.21**). Note that your results might differ.

**Figure 15.21** A Main Effects Only Screening Design

**Design**

Run	X1	X2	X3	X4	X5	X6
1	-1	1	-1	1	1	1
2	1	1	-1	1	1	-1
3	-1	1	1	-1	-1	1
4	1	1	1	1	-1	-1
5	-1	-1	-1	1	-1	1
6	1	-1	-1	-1	-1	-1
7	-1	1	-1	-1	-1	-1
8	1	-1	1	1	-1	1
9	-1	-1	1	-1	1	-1
10	1	-1	-1	-1	1	1
11	-1	-1	1	1	1	-1
12	1	1	1	-1	1	1

**Design Evaluation**

- ▷ Power Analysis
- ▷ Prediction Variance Profile
- ▷ Fraction of Design Space Plot
- ▷ Prediction Variance Surface
- ▷ Estimation Efficiency

**Alias Matrix**

Effect	X1*X2	X1*X3	X1*X4	X1*X5	X1*X6	X2*X3	X2*X4	X2*X5	X2*X6	X3*X4	X3*X5	X3*X6	X4*X5	X4*X6	X5*X6
Intercept	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X1	0	0	0	0	0	0.333	0.333	0.333	0.333	-0.33	0.333	-0.33	0.333	-0.33	0.333
X2	0	0.333	0.333	0.333	-0.33	0	0	0	0	-0.33	-0.33	0.333	0.333	-0.33	0.333
X3	0.333	0	0.333	-0.33	0.333	0	-0.33	-0.33	0.333	0	0	0	-0.33	-0.33	-0.33
X4	0.333	0.333	0	-0.33	-0.33	0	0.333	-0.33	0	-0.33	-0.33	0	0	0	-0.33
X5	0.333	-0.33	-0.33	0	0.333	-0.33	0.333	0	0.333	-0.33	0	-0.33	0	-0.33	0
X6	-0.33	0.333	-0.33	0.333	0	0.333	-0.33	0.333	0	-0.33	-0.33	0	-0.33	0	0

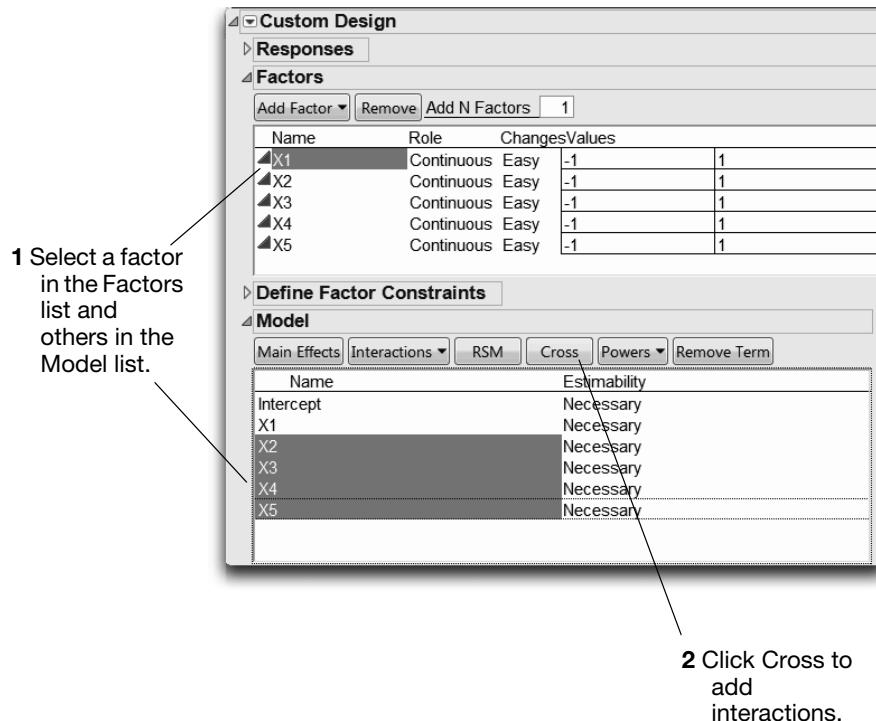
## All Two-Factor Interactions Involving a Single Factor

Sometimes there is reason to believe that some, but not all, two-factor interactions might be important. The following example illustrates adding all two-factor interactions involving a single factor. The example has five continuous factors.

- ⌚ In a new Custom Design window, enter five continuous factors and click **Continue**.

To get a specific set of crossed factors (rather than all interaction terms):

- ⌚ Select the factor to cross (X1, for example) in the Factors table.
- ⌚ Select the other factors in the Model Table and click **Cross**, as shown in **Figure 15.22**.

**Figure 15.22** Crossing Terms with One Factor

This design is a resolution-four design equivalent to folding over on the factor for which all two-factor interactions are estimable.

**Note:** The default sample size for designs with only two-level factors is the nearest multiple of 4 that has at least five more runs than there are effects in the model. For example, in **Figure 15.23** there are nine effects. The smallest multiple of four that is five runs higher than nine is 16.

✓ Select **Make Design**.

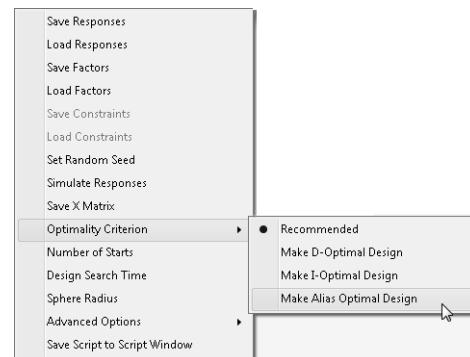
**Figure 15.23** Two-Factor Interactions for Only One Factor

Model		Design					
Name	Estimability	Run	X1	X2	X3	X4	X5
Intercept	Necessary	1	1	1	-1	1	-1
X1	Necessary	2	1	-1	-1	-1	-1
X2	Necessary	3	-1	1	1	1	1
X3	Necessary	4	-1	-1	1	-1	-1
X4	Necessary	5	1	1	1	-1	-1
X5	Necessary	6	-1	1	1	-1	-1
X1*X2	Necessary	7	-1	-1	-1	-1	-1
X1*X3	Necessary	8	1	-1	1	-1	1
X1*X4	Necessary	9	-1	1	1	1	1
X1*X5	Necessary	10	1	1	1	1	1
		11	1	1	-1	-1	1
		12	-1	-1	-1	1	1
		13	1	-1	1	1	-1
		14	1	-1	-1	1	1
		15	-1	-1	1	-1	1
		16	-1	1	-1	1	-1

## Alias Optimal Designs

Suppose we are interested in estimating main effects, but are concerned that there might be aliasing between main effects and two-factor interactions. In the Alias Matrix for the main effects only design (**Figure 15.21**), we see that main effects and two-factor interactions are aliased. For such screening design scenarios, an alias optimal design allows for the estimation of main effects while minimizing (and sometimes eliminating) the aliasing of main effects and two-factor interactions.

- ⓐ In a new Custom Design window, enter six continuous factors and click **Continue**.
- ⓑ Under the red triangle menu next to Custom Design, select **Make Alias Optimal Design** from the **Optimality Criterion menu**, as shown here.
- ⓒ Select **Make Design**.



The resulting 12-run design (see **Figure 15.24**) looks similar to the main effects design shown in **Figure 15.21**. But, look at the Alias Matrix at the bottom of **Figure 15.24**. Main effects are not aliased with any of the two-factor interactions. As a result, any of the two-way interactions can be uniquely estimated.

**Figure 15.24** Alias Optimal Design

Design						
Run	X1	X2	X3	X4	X5	X6
1	-1	-1	-1	1	1	1
2	-1	-1	1	1	-1	-1
3	1	1	-1	-1	1	1
4	-1	1	-1	1	-1	1
5	-1	1	-1	-1	1	-1
6	1	-1	1	-1	1	-1
7	1	-1	1	1	-1	1
8	1	-1	-1	1	-1	-1
9	-1	-1	-1	-1	-1	1
10	1	1	1	-1	-1	-1
11	-1	1	1	-1	1	1
12	1	1	1	1	1	-1

Alias Matrix															
Effect	X1*X2	X1*X3	X1*X4	X1*X5	X1*X6	X2*X3	X2*X4	X2*X5	X2*X6	X3*X4	X3*X5	X3*X6	X4*X5	X4*X6	X5*X6
Intercept	0	0.333	0	0	-0.33	0	-0.33	0.333	0	0	0	-0.33	-0.33	0	0
X1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

**Note:** For a complete discussion of screening designs and design evaluation, see the Screening Designs chapter in the JMP *Design of Experiments Guide*.

## Response Surface Designs

Response surface designs are useful for modeling a curved (quadratic) surface to continuous factors. If a minimum or maximum response exists inside the factor region, a response surface model can pinpoint it. The standard two-level designs cannot fit curved surfaces; a minimum of three distinct values for each factor are necessary to fit a quadratic function.

### The Odor Experiment

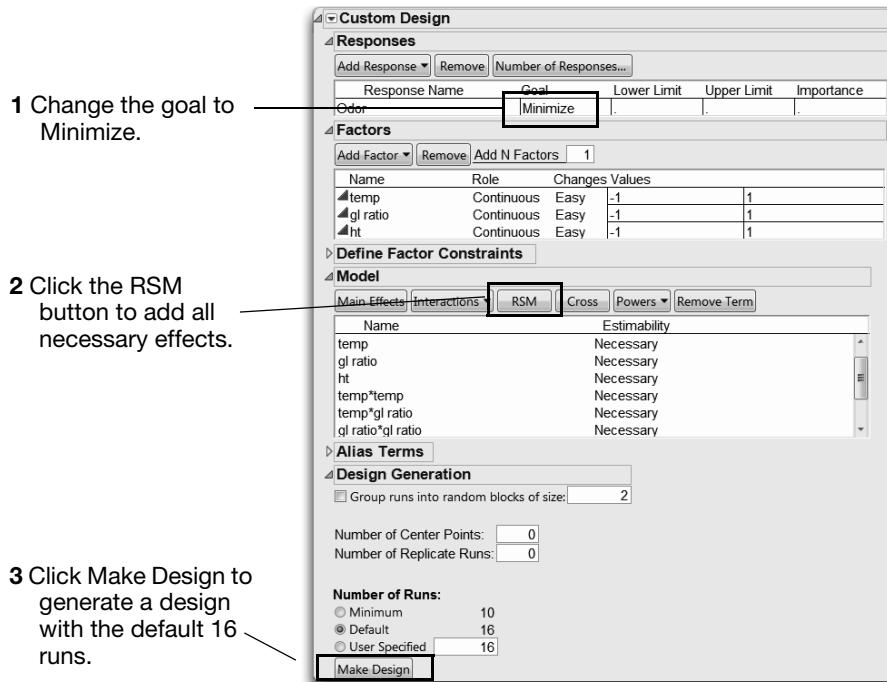
Suppose the objective of an industrial experiment is to minimize the unpleasant odor of a chemical. It is known that the odor varies with temperature (temp), gas-liquid ratio (gl ratio), and packing height (ht). The experimenter wants to collect data over a wide range of values for these variables to see whether a response surface can identify values that give a minimum odor (adapted from John, 1971).

### Response Surface Designs in JMP

To generate a standard response surface design, select **Response Surface Design** from the **DOE > Classical** menu. However, the Custom Designer can also produce standard response surface designs, and provides much more flexibility in design selection.

- ☛ Select **DOE > Custom Design**.
- ☛ In the resulting window, enter the response and three continuous factors (shown in **Figure 15.25**). Change the response goal to **Minimize**.
- ☛ Click **Continue**.
- ☛ Click the RSM button to add the interaction and power terms that are necessary to model the quadratic surface. The **Number of Runs** defaults to 16.
- ☛ Click **Make Design**.
- ☛ When the design appears, select **Sort Left to Right** from the **Run Order** menu.
- ☛ Click **Make Table** to see the JMP design table in **Figure 15.26**.

**Figure 15.25** Design Window to Specify Factors



The resulting design table and experimental results are in the JMP sample data table called **Odor JSS.jmp** (**Figure 15.26**). JMP has produced a 16-run Central Composite design.

**Figure 15.26** Odor JSS Response Surface Design and Values

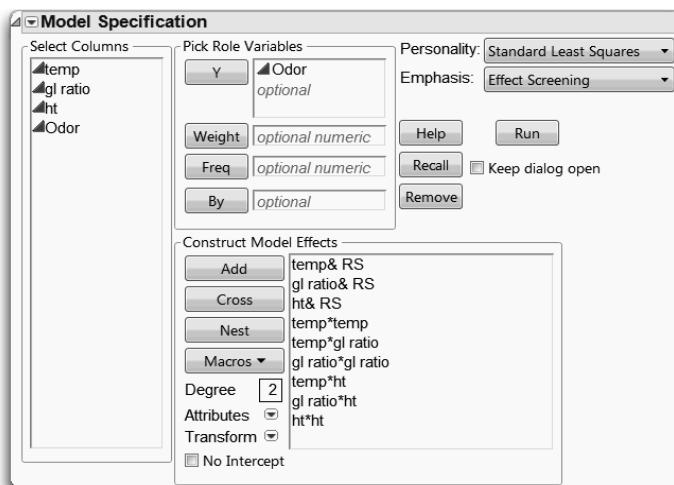
The screenshot shows the JMP software's 'Tables' panel with a table titled 'Odor JSS'. The table has four columns: 'temp', 'gl ratio', 'ht', and 'Odor'. The 'temp' column has values -1, -1, -1, 1, -1, 0, 0, 0, 0, 0, 1, 0, 1, -1, 1, 1, 1. The 'gl ratio' column has values -1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, -1, 1, 1, 1, 1. The 'ht' column has values -1, 1, 0, 0, 1, -1, 0, 0, 0, 0, 0, 0, 1, 0, -1, 0, 1. The 'Odor' column has values 150, 109, 43, 84, 63, 77, 35, 4, 10, 12, 28, 99, 83, 42, 87, 61. The table also includes sections for 'Columns (4/0)', 'Rows', and summary statistics like 'All rows' (16), 'Selected' (0), 'Excluded' (0), 'Hidden' (0), and 'Labelled' (0).

		temp	gl ratio	ht	Odor
Design	Custom Design	1	-1	-1	150
Criterion	I Optimal	2	-1	-1	109
▶ Screening		3	-1	0	43
▶ Model		4	-1	1	84
▶ DOE Dialog		5	-1	1	63
Columns (4/0)		6	0	-1	77
▲ temp *		7	0	0	35
▲ gl ratio *		8	0	0	4
▲ ht *		9	0	0	10
▲ Odor *		10	0	0	12
Rows		11	0	1	28
All rows	16	12	1	-1	99
Selected	0	13	1	-1	83
Excluded	0	14	1	0	42
Hidden	0	15	1	1	87
Labelled	0	16	1	1	61

## Analyzing the Odor Response Surface Design

Like all JMP tables generated by the Custom Designer, the Tables panel contains a **Model** script that generates the completed Fit Model launch window for the design.

- ❖ Select **Help > Sample Data Library** and open Odor JSS.jmp if you haven't already opened it.
- ❖ Click the green triangle next to the Model script, or select **Analyze > Fit Model** to see the completed window in **Figure 15.27**.

**Figure 15.27** Odor Fit Model Window

The effects appear in the model effects list as shown in **Figure 15.27**, with the &RS notation on the main effects (temp, gl ratio, and ht). This notation indicates that these terms are to be subjected to a curvature analysis.

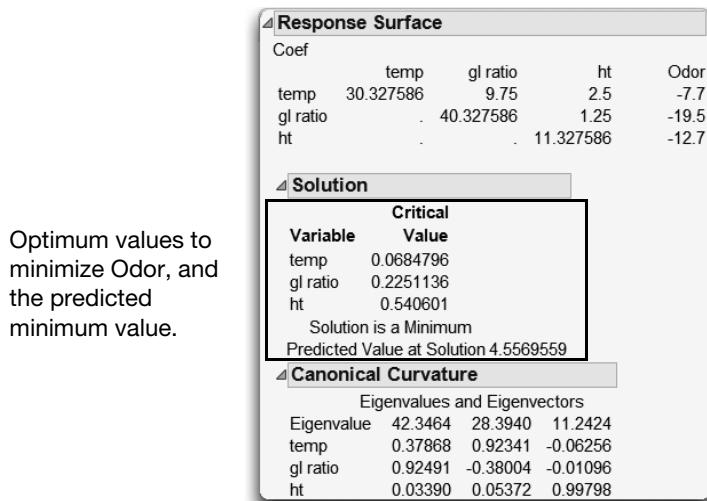
☞ Click **Run** to see the analysis.

The standard plots and least squares analysis tables appear. The additional report outline level called Response Surface is included.

☞ Open the Response Surface outline level to see the tables shown in **Figure 15.28**.

- The first table is a summary of the parameter estimates.
- The Solution table lists the critical values of the surface and tells the type of solution (maximum, minimum, or saddle point). The critical values are where the surface has a slope of zero, which could be an optimum depending on the curvature.
- The Canonical Curvature table shows eigenvalues and eigenvectors of the effects. The eigenvectors are the directions of the principal curvatures. The eigenvalue associated with each direction tells whether it is a decreasing slope, like a maximum (negative eigenvalue), or an increasing slope, like a minimum (positive eigenvalue).

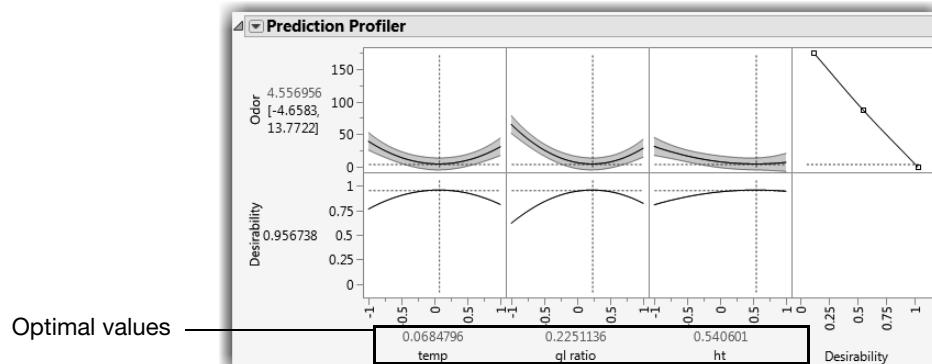
The Solution table in this example shows the solution to be a minimum.

**Figure 15.28** Response Surface Model and Analysis Results

The Prediction Profiler enables us to visualize the response surface model, and confirms the optimum results in **Figure 15.28**.

- ☞ Select **Optimization and Desirability > Maximize Desirability** from the red triangle menu next to Prediction Profiler.

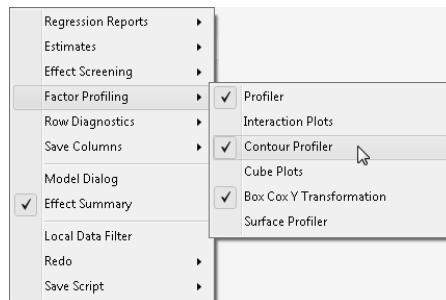
Since we specified Minimize as the response goal when we designed the experiment, the Prediction Profiler finds settings of the factors that minimize the predicted response (**Figure 15.29**).

**Figure 15.29** Prediction Profiler

## Plotting Surface Effects

If there are more than two factors, you can see a contour plot of any two factors at intervals of a third factor by using the Contour Profiler. This profiler is useful for graphically optimizing response surfaces.

- ⓐ Select **Factor Profiling > Contour Profiler** from the red triangle menu next to Response Odor, as shown here.



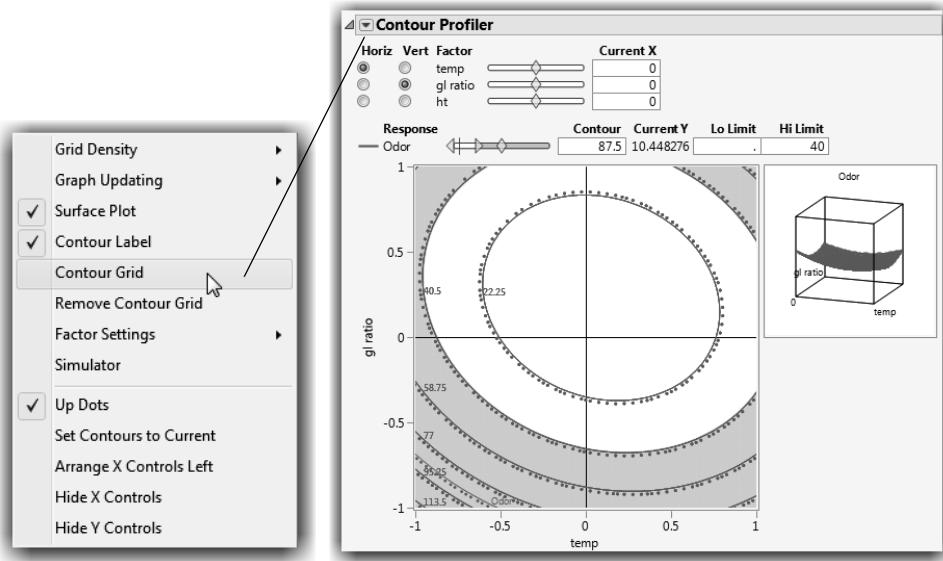
The **Contour Profiler** displays a panel that lets you use interactive sliders to vary one factor's values and observe the effect on the other two factors. You can also vary one factor and see the effect on a mesh plot of the other two factors. **Figure 15.30** shows contours of ht as a function of temp and gl ratio.

- ⓐ Enter 40 as the Hi Limit specification for Odor and press Enter.

Entering a value defines and shades the region of acceptable values for the three predictor variables.

Optionally, use the **Contour Grid** option to add grid lines with specified values to the contour plot, as shown in **Figure 15.30**

- ⓐ Click on the red triangle next to Contour Profiler, and select **Contour Grid**.
- ⓐ In the Contour Grid panel, change the values as desired, and click **OK** (the default values were used in **Figure 15.30**).

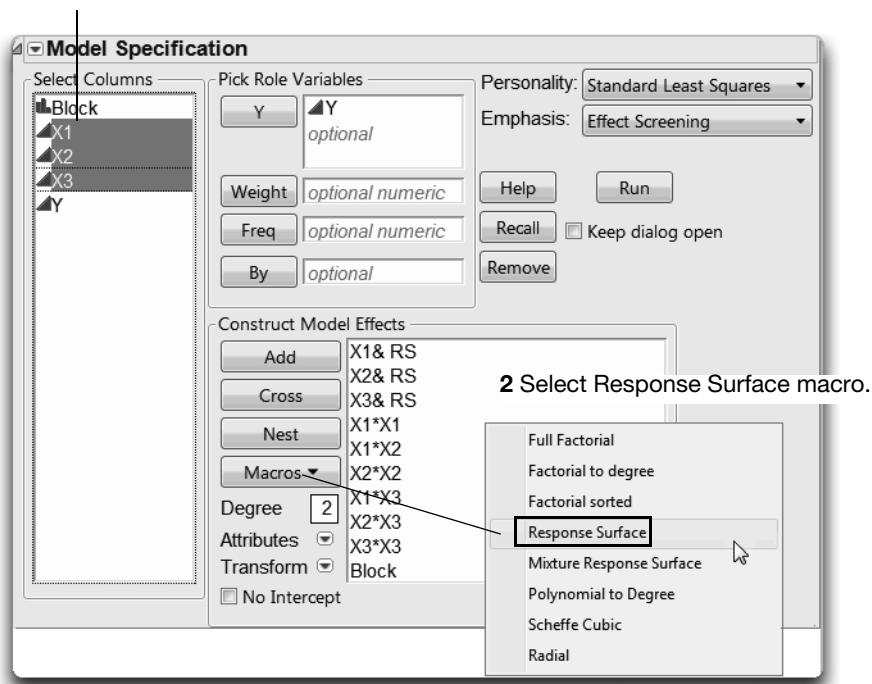
**Figure 15.30** Contour Profiler with Contours and Mesh Plot

## Specifying Response Surface Effects Manually

JMP completes the Fit Model's Model Specification window automatically when you build the design using one of the JMP DOE design options. If you're working with experimental results and the data table does not have a saved model script, you can generate a response surface analysis manually. In the Model Specification window, select effects in the column selection list and select **Response Surface** from the effect **Macros**, as illustrated in **Figure 15.31**.

**Figure 15.31** Fit Model Window for Response Surface Design

1 Select variables.



2 Select Response Surface macro.

## The Custom Designer versus the Response Surface Design Platform

We designed the Odor experiment using the Custom Designer. The resulting design was a 16-run Central Composite Design (CCD). Why didn't we just use the Response Surface Design Platform? In three-factor response surface designs, the default is, indeed, a CCD. However, classical response surface designs involving four or more factors tend to be unwieldy, growing rapidly as the number of factors increases. The Custom Designer generally produces smaller designs by default. For example, compare four-factor response surface designs generated from the Custom Designer and the Response Surface platform (**Figure 15.32**).

Although the smallest classical design is 27 runs (a Box-Behnken), the Custom Designer defaults to 21 runs. In addition, the Custom Designer enables the user to design the experiment to support the proposed model. The size of the design can be minimized by specifying only those cross products and quadratic effects that need to be estimated.

**Figure 15.32** Four-Factor Response Surface Designs

The image displays two side-by-side JMP software windows illustrating four-factor response surface designs.

**Left Window: Custom Design**

- Responses:** Y
- Factors:** X1, X2, X3, X4, X1\*X1, X1\*X2, X2\*X2, X1\*X3
- Model:** Main Effects, Interactions, RSM, Cross, Powers, Remove Term
- Estimability:** Necessary for all terms except X1\*X2 which is "Necessary If Possible".
- Alias Terms:** None listed.
- Design Generation:**
  - Group runs into random blocks of size: 2
  - Number of Center Points: 0
  - Number of Replicate Runs: 0
- Number of Runs:**
  - Minimum: 15
  - Default: 21 (selected)
  - User Specified: 21

**Right Window: Response Surface Design**

- Responses:** Y, Maximize
- Factors:**

Name	Role	Values
X1	Continuous	-1
X2	Continuous	-1
X3	Continuous	-1
X4	Continuous	-1
- 4 Factors - Choose a Design:**

Number Of Runs	Block Size	Center Points	Design Type
27	3	3	Box-Behnken
27	9	3	Box-Behnken
26	2	2	Central Composite Design
30	10	6	CCD-Orthogonal Blocks
31	7	7	CCD-Uniform Precision
36	12	12	CCD-Orthogonal

## Split-Plot Designs

In the experimental designs presented so far, all of the trials were run in random order (although we sorted the trials for illustration). Randomization is an experimental strategy for averaging out the influences of uncontrolled variables. For example, changes in weather or equipment conditions during an experiment can easily impact results and cloud findings.

In a completely randomized design, all of the factor settings can change from one run to the next. However, in many design situations, some factors are hard to change, and it is difficult or impossible to run the trials in a completely random order. It can be more convenient, and even necessary, to conduct the experiment using a split-plot design structure. Split plot experiments are structured with groups of runs called *whole plots*, where one or more factors stay constant within each run group. In JMP these factors are designated as **Hard** to change factors, and factors that can be fully randomized within the experiment are designated as **Easy** to change.

Split plot experiments occur often in industrial settings, as illustrated by the next example.

## The Box Corrosion Split-Plot Experiment

Box et al. (2005) discuss an experiment to improve the corrosion resistance of steel bars in an industrial setting. A coating was applied to the bars, which are then heated in a furnace for a fixed amount of time. The factors of interest are furnace temperature and coating type. Three temperatures and four coatings were tested:

- Furnace Temp: 360, 370, and 380 degrees C
- Coating: C1, C2, C3, and C4

Since it is difficult and time-consuming to change the furnace temperature, a split plot design was used. The experiment was run in temperature groups, or “heats.” For each heat, the furnace is held constant at one of the three temperatures. Within each heat, four bars treated with one of the four coatings are placed randomly in the furnace and heated to the designated temperature.

The experiment consists of two different types of factors:

- Furnace Temp is a whole plot, or *hard to change* factor. It is held constant within each heat.
- Coating is a split plot, or *easy to change* factor. It is fully randomized within each heat.

Each of the temperature settings is run twice, resulting in six heats and a total of 24 treated bars.

## Designing the Experiment

To generate the design, we again use the Custom Design platform.

- ☞ Select **DOE > Custom Design**.
- ☞ Enter the response, Corrosion Resistance.
- ☞ Enter Furnace Temp as a 3-Level categorical factor.
- ☞ Enter Coating as a 4-Level categorical factor.
- ☞ Enter factor settings as shown in **Figure 15.33**.

By default, both factors show as **Easy** to change in the Factors panel under Changes. To identify Furnace Temp as a hard to change factor:

- ☞ Click on the word **Easy** next to Furnace Temp in the Changes column and select **Hard** from the Changes menu.

✓ Click **Continue**.

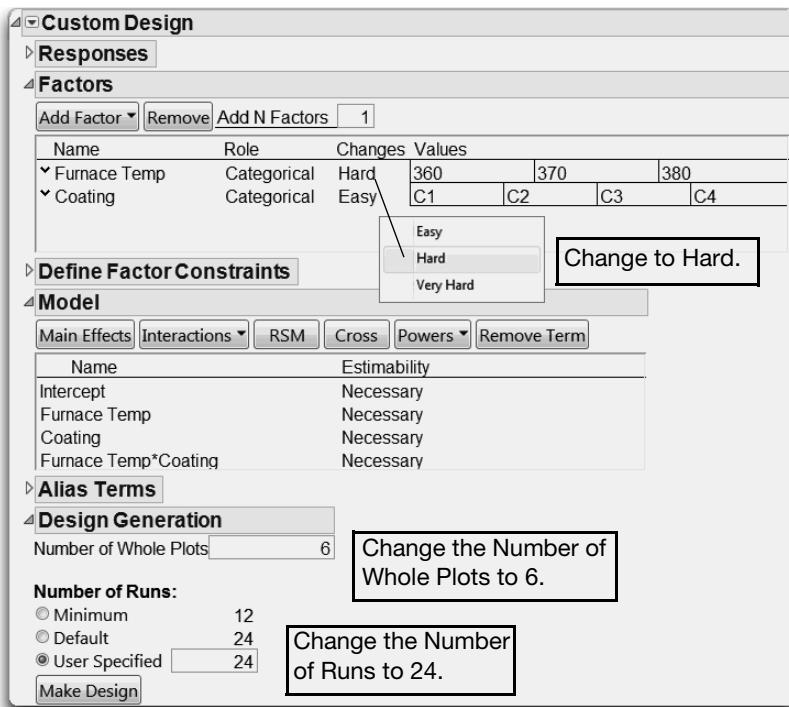
✓ Click **Interactions** > **2nd** to add the Furnace Temp\*Coating interaction.

The Model panel now contains two main effects and the interaction (**Figure 15.33**).

Finally, specify the number of heats, which are the whole plots, and the total number of runs.

- ✓ Under **Design Generation**, enter 6 for the number of whole plots (two whole plots for each temperature level).
- ✓ Change the **Number of Runs** to 24 (four bars in each of the six heats).

**Figure 15.33** Split Plot Design



✓ Click **Make Design**, to see the design in the Custom Designer.

✓ Click **Make Table** to generate the design table.

The design table and results are in the JMP sample data table Design Experiment/Box Corrosion Split-Plot Design.jmp (**Figure 15.34**).

**Figure 15.34** Box Corrosion Split-Plot Data Table

	Whole Plots	Furnace Temp	Coating	Corrosion Resistance
1	1	370	C2	91
2	1	370	C3	87
3	1	370	C1	65
4	1	370	C4	86
	5	370	C3	121
	6	370	C4	150
	7	370	C2	142
	8	370	C1	140
	9	360	C3	83
	10	360	C2	73
	11	360	C4	89
	12	360	C1	67
	13	380	C4	212
	14	380	C3	147
	15	380	C2	127
	16	380	C1	155
	17	380	C4	153
All rows	18	380	C2	100
Selected	19	380	C1	108
Excluded	20	380	C3	90
Hidden	21	360	C3	46
Labelled	22	360	C1	33
	23	360	C4	54
	24	360	C2	8

## Analysis of Split-Plot Designs

The completely randomized designs seen earlier in this chapter have one level of experimental units to which treatments are applied. As such, there is one error source, the residual error, to test for significance of model effects.

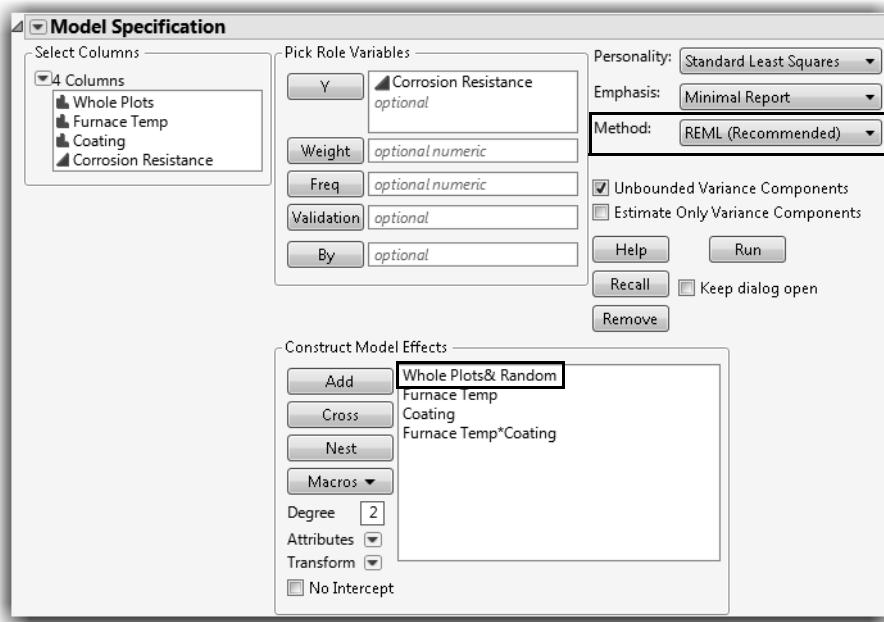
In split-plot experiments, there are two levels of experimental units: whole plots and split plots. Each of these levels contributes a unique source of variation.

The replication of the whole-plot factor, furnace temperature, provides an estimate of whole-plot random error due to resetting the furnace temperature. At the split-plot level, there is residual variation due to differences in temperature at different positions in the furnace, differences in coating thickness from bar to bar, and so on.

Each term in the model must be tested against the appropriate error term (whole plot or split plot). As a result, analysis of split-plot experiments is more complicated than that for completely randomized experiments (Jones and Nachtsheim, 2009). Fortunately, the correct analysis has been saved in the Tables panel of the data table.

- ❖ In the Box Corrosion Split-Plot sample data table, click the green triangle next to the Model script to run the script and see the completed model window in **Figure 15.35**.

**Figure 15.35** Split Plot Model Window



The model effects include Furnace Temp, Coating, and the two-way interaction, along with the random effect Whole Plots&Random. This extra term triggers the REML variance components analysis, and ensures that the model effects are tested against the correct error term.

**Note:** See the section “Optional Topic: Random Effects and Nested Effects” on page 406 in Chapter 14, “Fitting Linear Models,” for a discussion of the REML (Restricted Maximum Likelihood) Method.

- ❖ Click **Run** to produce the results in **Figure 15.36**. Note that the Effect Summary table is not shown.

The Fixed Effect Tests panel shows that Coating and the Furnace Temp\*Coating interaction are significant (with  $p$ -values of 0.0020 and 0.0241 respectively). Furnace Temp is not significant independent of the coating used.

The variance components estimates are given in the REML report. Furnace Temp is tested against the Whole-Plot variance, but Coating and the interaction are tested against the Residual variance. Note that the whole-plot variance is 9.41 times larger than the residual variance.

**Figure 15.36** Box Corrosion Split Plot Results

**Response Corrosion Resistance**

**Summary of Fit**

RSquare	0.977225
RSquare Adj	0.956347
Root Mean Square Error	11.15982
Mean of Response	101.125
Observations (or Sum Wgts)	24

**Parameter Estimates**

**Random Effect Predictions**

**REML Variance Component Estimates**

Random Effect	Var Ratio	Component	Std Error	95% Lower	95% Upper	Pct of Total
Whole Plots	9.4118434	1172.1667	982.60166	-753.6972	3098.0305	90.396
Residual		124.54167	58.709505	58.922814	415.07901	9.604
Total		1296.7083	983.47823	441.11562	13553.071	100.000

-2 LogLikelihood = 128.33807589

Note: Total is the sum of the positive variance components.  
Total including negative estimates = 1296.7083

**Covariance Matrix of Variance Component Estimates**

**Iterations**

**Fixed Effect Tests**

Source	Nparm	DF	DDFDen	F Ratio	Prob > F
Furnace Temp	2	2	3	2.7548	0.2093
Coating	3	3	9	11.4798	0.0020*
Furnace Temp*Coating	6	6	9	4.3757	0.0241*

But, what if we had ignored the split plot structure and analyzed this example as a completely randomized experiment as seen in **Figure 15.37**?

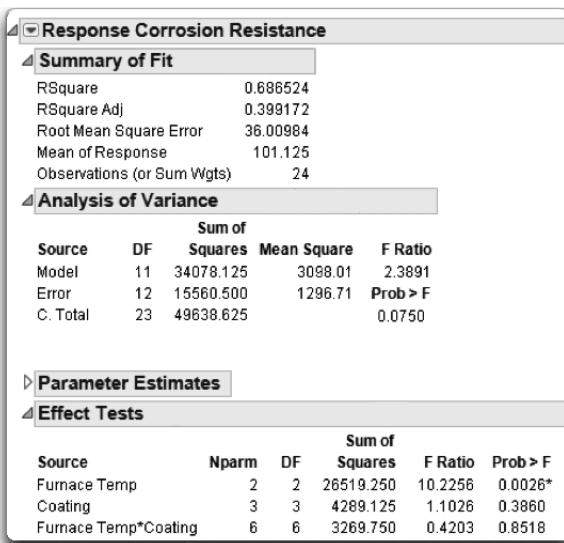
**Figure 15.37** Box Corrosion Incorrect Analysis

Table 15.1 summarizes some of the key differences.

**Table 15.1.** Compare *p*-values Between Correct Split Plot and Incorrect Analysis

	Correct Analysis (Split Plot)	Incorrect Analysis (Fully Randomized)
<b>RMSE</b>	11.16	36.01
<b>Furnace Temp</b>	0.2093	0.0026
<b>Coating</b>	0.0020	0.3860
<b>Furnace Temp*Coating</b>	0.0241	0.8518

By ignoring the split plot structure, we would have determined that only Furnace Temp is significant. Further, we would have concluded that there is no difference in coatings, and that the effect of temperature on corrosion resistance is independent of the coating applied. The incorrect analysis would have led to the opposite (and incorrect) conclusions!

## Design Strategies

So far, our discussion on generating designs has focused on the particulars of how to get certain types of designs. But why are these designs good? Here is some general advice.

**Multifactor designs yield information faster and better.** Experiments that examine only one factor, and therefore ask only one question at a time, are inefficient.

For example, suppose that you have one factor with two levels, and you need 16 runs to see a difference that meets your requirements. Increasing to seven factors (and running each as a separate experiment) would take  $7 \times 16 = 104$  runs. However, with a multifactor experimental design, all seven factors can be estimated in those 16 runs. The seven factors might have the same standard error of the difference, saving you  $7/8$ s of your runs.

But that is not all. With only 16 runs you can still check for two-way interactions among the seven factors. The effect of one factor might depend on the settings of another factor. This cannot be discovered with experiments that vary only one factor at a time.

Experimentation enables you to build a comprehensive model and understand more of the dynamics of the response's behavior.

**Do as many runs as you can afford.** The more information you have, the more you know. You should always try to perform an experiment with as many runs as possible. Since experimental runs are costly, you have to balance the limiting factor of the cost of trial runs with the amount of information that you expect to get from the experiment.

The literature of experimental design emphasizes orthogonal designs. They are great designs, but come only in certain sizes (for example, multiples of 4).

Suppose you have a perfectly designed (orthogonal) experiment with 16 runs. Given the opportunity to do another virtually free run, therefore using 17 runs, but making the design unbalanced and non-orthogonal, should you accept the run?

Yes. *Definitely.* Adding runs never subtracts information.

**Take values at the extremes of the factor range.** This has two advantages and one disadvantage:

- It maximizes the power. By separating the values, you are increasing the parameter for the difference, making it easy to distinguish it from zero.
- It keeps the applicable range of the experiment wide. When you use the prediction formula outside the range of the data, its variance is high, and it is unreliable for other reasons.
- The disadvantage is that the true response might not be approximated as well by a simple model over a larger range.

**Add some replicates.** Typing “1” into the **Number of Replicate Runs** field in the Design Generation panel adds one

additional run at the optimal design point. Replicating runs provides an estimate of pure error, and improves the estimates of terms in the model.

The screenshot shows the 'Design Generation' panel with the following settings:

- Group runs into random blocks of size
- Number of Center Points:
- Number of Replicate Runs:

You can also add center points to the design in the **Number of Center Points** field. Center points are runs set at the mid-points of continuous factors, and are commonly used to test for lack-of-fit (that is, potential curvature). However, the center of the design space is not the optimal location for replicates, and curvature can be modeled directly by adding quadratic terms.

**Randomize the assignment of runs.** Although we've presented designs that have been sorted (for illustration), randomization is critical to neutralize any inadvertent effects that might be present in a given sequence.

## DOE Glossary of Key Terms

**Augmentation.** If you want to estimate more effects, or learn more about the effects that you have already examined, adding runs to the experiment in an optimal way assures that you learn the most from the added runs.

**Balanced.** A design is balanced when there are the same number of runs for each level of a factor, and for various combinations of levels of factors.

**Coding.** For continuous factors, coding transforms the data from the natural scale to a -1 to 1 scale, so that effect sizes are relative to the range. For categorical factors, coding creates design columns that are numeric indicator columns for the categories.

**Definitive Screening Designs.** Definitive screening designs are a family of extremely efficient (small) screening designs for continuous or two-level categorical factors. The designs are particularly useful if you suspect active two-factor interactions or suspect that a continuous factor's effect on the response might exhibit strong curvature.

**Effect Precision.** Effect precision is the difference in the expected response attributed to a change in a factor. This precision is the factor's coefficient, or the estimated parameter in the prediction expression.

**Effect Size.** The standard error of the effect size measures how well the experimental design estimates the response. This involves the experimental error variance, but it can be determined relative to this variance before the experiment is performed.

**Folding.** If an effect is confounded with another, you can add runs to the experiment and flip certain high/low values of one or more factors to remove their confounding.

**Formulation.** Sometimes there is a model constraint. For example, factors that are mixture proportions must sum to 1.

**Fractional Factorial.** A fraction of a factorial design, where some factors are formed by interactions of other factors. They are frequently used for screening designs. A fractional factorial design also has a sample size that is a power of two. If  $k$  is the number of factors, the number of runs is  $2^{k-p}$  where  $p < k$ . Fractional factorial designs are orthogonal.

**Full Factorial.** All combinations of levels are run. If there are  $k$  factors and each has two levels, the number of factorial runs is  $2^k$ .

**Minimum-Aberration.** A *minimum-aberration design* has a minimum number of confoundings for a given resolution.

**Mixture Design.** Any design where some of the factors are mixture factors. Mixture factors express percentages of a compound and must therefore sum to one. Classical mixture designs include the *simplex centroid*, *simplex lattice*, and *ABCD*.

**D-Optimality.** Designs that are *D*-Optimal maximize a criterion so that you learn the most about the parameters (they minimize the generalized variance of the estimates). This is an excellent all-purpose criterion, especially in screening design situations.

**I-Optimality.** Designs that are *I*-Optimal maximize a criterion so that the model predicts best over the region of interest (they minimize the average variance of prediction over the experimental region). This is a very good criterion for response surface optimization situations.

**Optimization.** You want to find the factor settings that optimize a criterion. For example, you might want factor values that optimize quality, yield, or cost while minimizing bad side effects.

**Orthogonal.** The estimates of the effects are uncorrelated. If you remove an effect from the analysis, the values of the other estimates remain the same.

**Plackett-Burman.** *Plackett-Burman* designs are an alternative to fractional factorials for screening. Since there are no two-level fractional factorial designs with sample sizes between 16 and 32 runs, a useful characteristic of these designs is that the sample size is a multiple of four rather than a power of two. However, there are 20-run, 24-run, and 28-run Plackett-Burman designs.

The main effects are orthogonal, and two-factor interactions are only partially confounded with main effects. This is different from resolution-3 fractional factorials where two-factor interactions are indistinguishable from main effects.

In cases of *effect sparsity* (where most effects are assumed to have no measurable effect on the response), a stepwise regression approach can enable you to remove some insignificant main effects while adding highly significant and only somewhat correlated two-factor interactions.

**Prediction Designs.** Rather than looking at how factors contribute to a response, you instead want to develop the most accurate prediction.

**Robustness.** You want to determine operational settings that are least affected by variation in uncontrollable factors.

**Resolution.** The *resolution* number is a way to describe the degree of confounding, usually focusing on main effects and two-way interactions. Higher-order interactions are assumed to be zero.

In resolution-3 designs, main effects are not confounded with other main effects, but two-factor interactions are confounded with main effects. Only main effects are included in the model. For the main effects to be meaningful, two-factor interactions are assumed to be zero or negligible.

In resolution-4 designs, main effects are not confounded with each other or with two-factor interactions, but some two-factor interactions can be confounded with each other. Some two-factor interactions can be modeled without being confounded. Other two-factor interactions can be modeled with the understanding that they are confounded with two-factor interactions included in the model. Three-factor interactions are assumed to be negligible.

In resolution-5 designs, there is no confounding between main effects, between two-factor interactions, or between main effects and two-factor interactions. That is, all two-factor interactions are estimable.

**Response Surface.** Response Surface Methodology (RSM) is a technique that finds the optimal response within specified ranges of the factors. These designs can fit a second-order prediction equation for the response. The quadratic terms in these equations model the curvature in the true response function. If a maximum or minimum exists inside the factor region, RSM can find it.

In industrial applications, RSM designs usually involve a small number of factors, because the required number of runs increases dramatically with the number of factors.

**Screening Designs.** These designs sort through many factors to find those that are most important, that is, those that the response is most sensitive to. Such factors are the vital few that account for most of the variation in the response.

**Significance.** You want a statistical hypothesis test to show significance at meaningful levels. For example, you want to show statistically that a new drug is safe and effective as a regulatory agency. Perhaps you want to show a significant result to publish in a scientific journal. The  $p$ -value shows the significance. The  $p$ -value is a function of both the estimated effect size and the estimated variance of the experimental error. It shows how unlikely so extreme a statistic would be due only to random error.

**Space-Filling.** A design that seeks to distribute points evenly throughout a factor space.

**Split Plot.** In many design situations, some factors might be easy to change; others might be harder to change. Although it is possible to fully randomize designs involving only easy-to-change factors, randomization is limited with experiments involving factors that are hard to change. To accommodate the restrictions on randomization, split plot designs should be used.

In a split-plot experiment, one or more hard-to-change “whole plot” factors are held constant while the easy-to-change “split plot” factors are randomized within the whole plot. This is repeated for the different settings of the whole plot factor(s).

**Taguchi.** The goal of the Taguchi Method is to find control factor settings that generate acceptable responses despite natural environmental and process variability. In each experiment, Taguchi’s design approach uses two designs called the *inner* and *outer* array. The Taguchi experiment is the cross product of these two arrays. The control factors, used to modify the process, form the inner array. The noise factors, associated with process or environmental variability, form the outer array. Taguchi’s signal-to-noise ratios are functions of the observed responses over an outer array. The Taguchi designer in JMP supports all these features of the Taguchi method. The inner and outer array design lists use the traditional Taguchi orthogonal arrays such as L4, L8, L16, and so on.

## Exercises

1. The sample data table Reactor 32 Runs.jmp in the Sample Data library contains the factor settings and results for a 32 run  $2^5$  full factorial experiment. The response and factors are described in “An Interaction Model: The Reactor Data” on page 442.
  - (a) Go to **Analyze > Fit Model**. Assign Percent Reacted to **Y** and the five factors and all possible interactions as model effects. **Note:** Select the factors. Then under **Macros**, select **Full Factorial**. This fits a model with all main effects, two-way, three-way, four-way interactions, along with one five-way interaction. Click **Run**.
  - (b) Look at the Sorted Parameter Estimates table (select **Estimates > Sorted Estimates** from the red triangle next to Response). Are any of the higher order interactions significant? Which terms are significant? Compare these results to those found when we analyzed the more efficient Reactor 20 Custom.jmp experiment.
  - (c) Now, fit a model with only main effects and two-way interactions. **Hint:** Return to **Analyze > Fit Model**. The model effects panel should already be populated with all main effects and two-way interactions. To confirm that these terms are included in the model, select the factors, then under Macros

select **Factorial to Degree**. Make sure that Percent Reacted is assigned to **Y**, and click **Run**.

- (d) Use the Prediction Profiler to determine factor settings that maximize the predicted Percent Reacted. What are the optimal settings? What is the predicted response at these settings? Compare your findings to the prediction results from the Reactor 20 Custom.jmp experiment.
2. An industrial engineer would like to design an experiment to study the breaking strength of metal parts. Five two-level factors are under consideration, two continuous ( $X_1$  and  $X_2$ ) and three categorical ( $X_3$ ,  $X_4$ , and  $X_5$ ). Due to resource limitations and time constraints, only 15 trials can be conducted. It is known that factor  $X_1$  does not interact with the other factors. Use the custom designer to design a 15-trial screening experiment to estimate main effects and two-way interactions.
3. The sample data table Tiresread.jmp contains the results of a three-factor response surface experiment. The factors are Silica, Silane, and Sulfur. The experiment was conducted to optimize four responses: Abrasion, Modulus, Elongation, and Hardness. The goals for the four responses have been added as column properties. (Note: Click on the asterisk next to each response in the Columns panel, and select **Response Limits** to view the goals).
- Use the Distribution platform to explore the factor settings. How many settings were used for each factor?
  - Use **Graph > Graph Builder** to explore the relationship between the factors and the responses. Note: Drag a factor to the X zone and a response to the Y zone. To explore more than one factor at a time, drag a second factor to the X zone next to the first, and release. To explore more than one response at a time, drag a second response to the Y zone above the first, and release. The smooth lines (splines) describe the general relationship between the factors and each response.
  - Run the **RSM for 4 Responses** script in the Tables panel to fit a response surface model for all four responses. Scroll to the bottom of the Fit Least Squares window to see the Prediction Profiler. Note the prediction traces for each factor for the four responses. For example, the prediction traces for Sulfur are different for the four responses. An increase in sulfur results in an increase in abrasion, but causes a decrease in elongation. Also note the response goals in the last column of the profiler. The goals are maximized for the first two responses, and they match the target for the last two.

- (d) Select **Optimization and Desirability > Maximize Desirability** from the red triangle menu next to Prediction Profiler to simultaneously optimize all four responses. What are the optimal settings for the three factors? Were the response goals met (and met equally well) for all four responses?
- (e) Return to the data table. This experiment was a 20-run Central Composite response surface design. Use the Custom Designer to create a response surface design for the four responses and three factors. What is the minimum number of runs required? What is the default number of runs?



# 16

## Bivariate and Multivariate Relationships

### Overview

This chapter explores the relationship between two or more variables. You look for patterns and points that don't fit the patterns. You see where the data points are located, where the distribution is dense, and which way it is oriented. You explore which variables contribute the most information, and how observations and variables can be grouped or clustered.

Detective skills are built with the experience of looking at a variety of data, and learning to look at them in different ways. As you become a better detective, you also develop better intuition for understanding more advanced techniques.

It is not easy to look at lots of variables, but the increased range of the exploratory tools and techniques at your fingertips in JMP helps you make more interesting and valuable discoveries.

## Chapter Contents

Overview .....	479
Bivariate Distributions .....	481
Density Estimation .....	481
Bivariate Density Estimation .....	482
Mixtures, Modes, and Clusters .....	484
The Elliptical Contours of the Normal Distribution .....	485
Correlations and the Bivariate Normal .....	487
Simulating Bivariate Correlations .....	487
Correlations across Many Variables .....	490
Bivariate Outliers .....	491
Outliers in Three and More Dimensions .....	494
Identify Variation with Principal Components Analysis .....	496
Principal Components for Six Variables .....	499
How Many Principal Components? .....	501
Discriminant Analysis .....	502
Canonical Plot .....	503
Discriminant Scores .....	504
Stepwise Discriminant Variable Selection .....	507
Cluster Analysis .....	508
Hierarchical Clustering: How Does It Work? .....	508
A Real-World Example .....	511
Some Final Thoughts .....	514
Exercises .....	515

## Bivariate Distributions

Previous chapters covered how the distribution of a response can vary depending on factors and groupings. This chapter returns to explore distributions as simple unstructured batches of data. However, instead of a single variable, the focus is on the joint distribution of two or more responses.

## Density Estimation

As with univariate distributions, a central question is where are the data? What regions of the space are dense with data and what areas are relatively vacant? The histogram forms a simple estimate of the density of a univariate distribution. If you want a smoother estimate of the density, JMP has an option that takes a weighted count of a sliding neighborhood of points to produce a smooth curve. This idea can be extended to several variables.

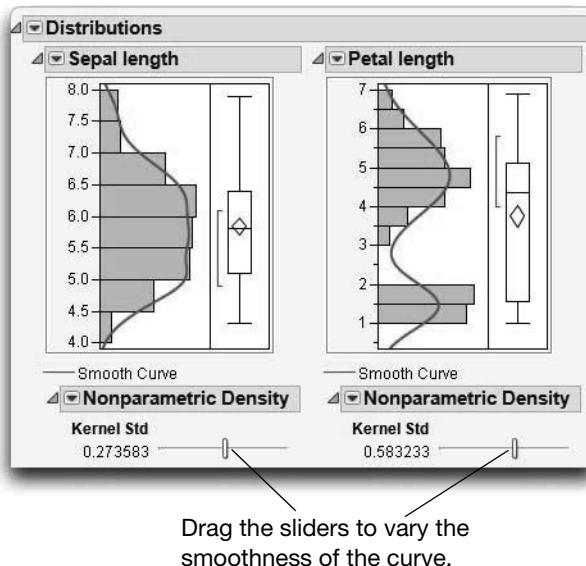
One of the most classic multivariate data sets in statistics contains the measurements of iris flowers that R. A. Fisher analyzed. Fisher's iris data are in the sample data table called Iris.jmp, with variables Sepal length, Sepal width, Petal length, and Petal width. First, look at the variables one at a time.

- ☛ Select **Help > Sample Data Library** and open Iris.jmp.
- ☛ Select **Analyze > Distribution**.
- ☛ Assign Sepal length and Petal length to **Y, Columns** and then click **OK**.
- ☛ In the Distribution report, hold down Ctrl and select **Continuous Fit > Smooth Curve** from the red triangle menu next to Sepal length.
- ☛ When the smooth curves appear, drag the density slider beneath the histogram to see the effect of using a wider or narrower smoothing distribution (**Figure 16.1**).

Notice in **Figure 16.1** that Petal length has an unusual distribution with two modes with no values between 2 and 3, which implies that there might be two distinct distributions.

**Note:** The bimodal distribution of Petal length could be explored using **Continuous Fit > Normal Mixtures > Normal 2 Mixture**, from the red triangle menu next to Petal length. But for now, we are interested in looking at how multiple variables behave together.

**Figure 16.1** Univariate Distribution with Smoothing Curve



## Bivariate Density Estimation

JMP has a smoother that works with two variables to show their bivariate densities. The goal is to draw lines around areas that are dense with points. Continue with the iris data and look at Petal length and Sepal length together:

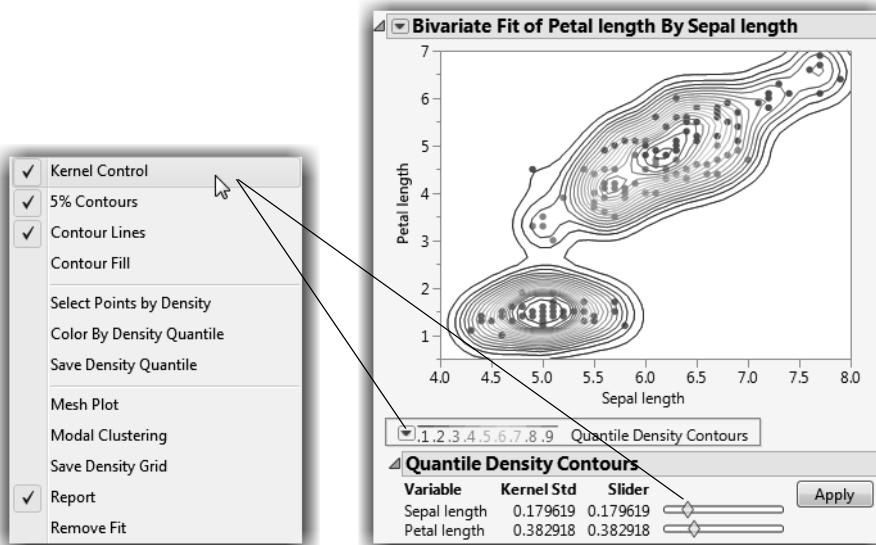
- ✓ Select **Analyze > Fit Y by X** and assign Petal length to **Y, Response** and Sepal length to **X, Factor**.
- ✓ Click **OK**.
- ✓ When the scatterplot appears, select **Nonpar Density** from the red triangle menu next to Bivariate Fit.

The result, shown in **Figure 16.2**, is a contour plot, where the various contour lines show paths of equal density. The density is estimated for each point on a grid by taking a weighted average of the points in the neighborhood, where the weights decline with distance. Estimates done in this way are called *kernel smoothers*.

The Nonparametric Bivariate Density table beneath the plot has slider controls available to control the vertical and horizontal width of the smoothing distribution.

- ☛ Select **Kernel Control** from the red triangle menu next to the Quantile Density Contours legend (below the plot).
- ☛ Adjust the slider bars. Calculating densities can take some time, so they are not re-estimated until you click the **Apply** button.

**Figure 16.2** Bivariate Density Estimation Curves



The density contours form a map showing where the data are most dense. The contours are calculated according to quantiles, where a certain percent of the data lie outside each contour curve. These quantile density contours show where each 5% and 10% of the data are. The innermost narrow contour line encloses the densest 5% of the data. The heavy line just outside surrounds the densest 10% of the data. It is colored as the 0.9 contour because 90% of the data lie outside it. Half the data distribution is inside the solid green lines, the 50% contours. Only about 5% of the data is outside the outermost 5% contour.

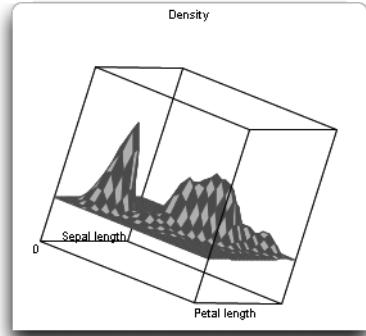
One of the features of the iris petal and sepal length data is that there seem to be several local peaks in the density. There are two "islands" of data, one in the lower left and one in the upper right of the scatterplot.

These groups of locally dense data are called *clusters*, and the peaks of the density are called *modes*.

- ☞ Select **Mesh Plot** from the menu on the legend of the Quantile Density Contours.

This produces a 3-D surface of the density, as shown here.

- ☞ Drag the mesh plot to rotate it.

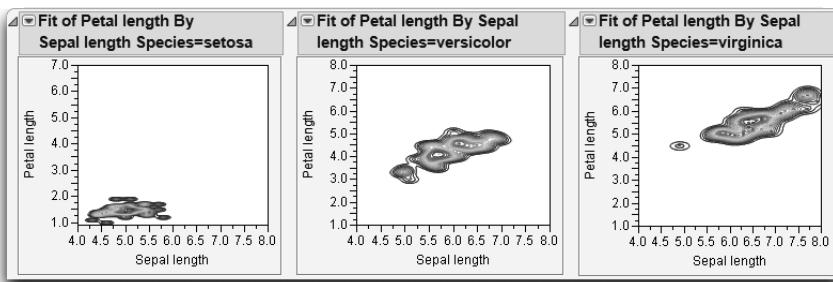


## Mixtures, Modes, and Clusters

Multimodal data often come from a mixture of several groups. Examining the data closely reveals that there is actually a collection of three species of iris: Virginica, Versicolor, and Setosa.

Conducting a bivariate density for each group results in the density plots in **Figure 16.3**. The axes are adjusted on these plots to show the same scales.

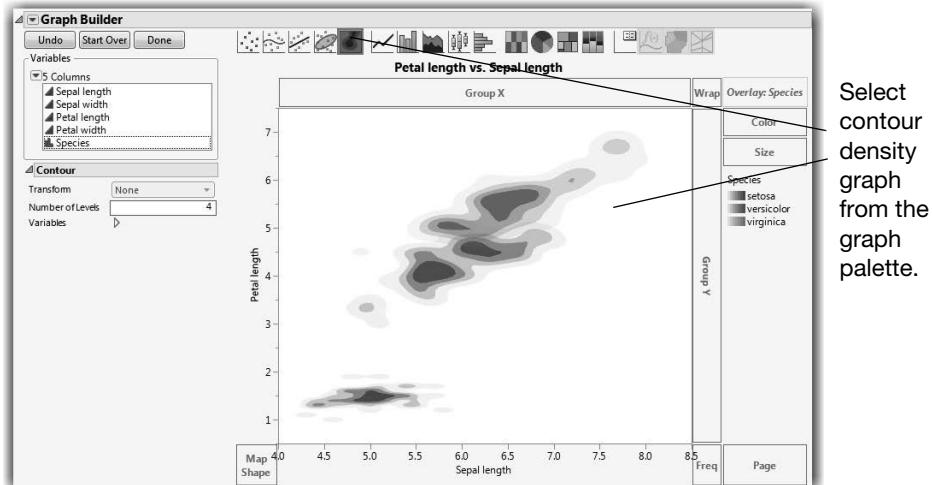
**Figure 16.3** Bivariate Density Curves



Another way to examine these groupings, on the same plot, is with the Graph Builder.

- ☞ Select **Graph > Graph Builder**
- ☞ Drag Petal length to the **Y** zone, Sepal length to the **X** zone, and Species to the **Overlay** zone.
- ☞ Click on the contour icon to see the graph in **Figure 16.4**.

**Note:** Enter a value in **Number of Levels** to change the number of contours.

**Figure 16.4** Graph Builder Contour Densities

You can again see that the three species fall in different regions for Petal length, while Virginica and Versicolor have some overlap of Sepal length.

**Note:** To classify an observation (an iris) into one of these three groups, a natural procedure is to compare the density estimates corresponding to the petal and sepal length of a specimen over the three groups, and assign it to the group where its point is enclosed by the highest density curve. This type of statistical method is also used in *discriminant analysis* and *cluster analysis*, shown later in this chapter.

## The Elliptical Contours of the Normal Distribution

Notice that the contours of the distributions on each species are elliptical in shape. It turns out that ellipses are the characteristic shape for a bivariate normal distribution. The Fit Y by X platform can show a graph of these normal contours.

Again select **Analyze > Fit Y by X** and assign Petal length to **Y, Response** and Sepal length to **X, Factor**.

Click **OK**.

When the scatterplot appears:

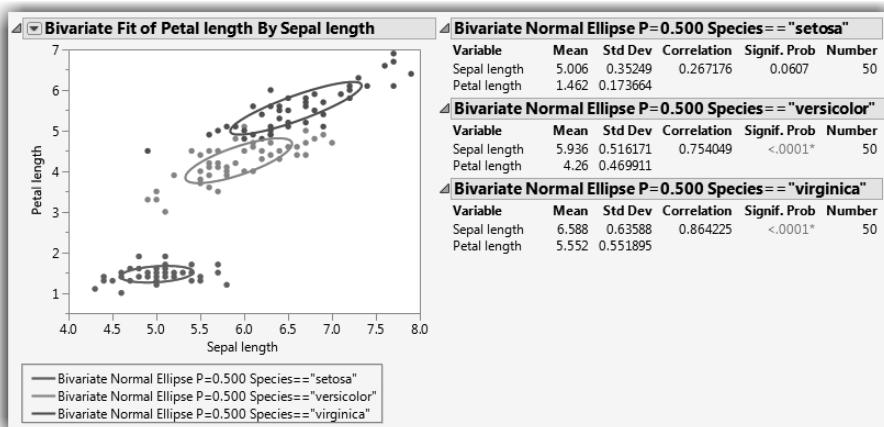
Select **Group By** from the red triangle menu next to Bivariate Fit and select Species as the grouping variable.

Click **OK**.

- >Select **Density Ellipse** from the red triangle menu next to Bivariate Fit with 0.50 as the probability level for the ellipse.

The result of these steps is shown in **Figure 16.5**. When there is a grouping variable in effect, there is a separate estimate of the bivariate normal density (or any fit you select) for each group. The normal density ellipse for each group encloses the densest 50% of the estimated distribution.

**Figure 16.5** Density Ellipses for Species Grouping Variable



Notice that the two ellipses toward the top of the plot are diagonally oriented, while the one at the bottom is not. The reports beneath the plot show the means, standard deviations, and correlation of Sepal length and Petal length for the distribution of each species. Note that the correlation is low for Setosa, and high for Versicolor and Virginica. The diagonal flattening of the elliptical contours is a sign of strong correlation. If variables are uncorrelated, their normal density contours appear to have a more non-diagonal and rounder shape.

One of the main uses of correlation is to see whether variables are related. You want to know if the distribution of one variable is a function of the other. That is, if you know the value of one variable can you predict the value of the other? When the variables are normally distributed and uncorrelated, the univariate distribution of one variable is the same no matter what the value of the other variable is. When the density contours have no diagonal aspect, the density across any slice is the same no matter where you take that slice (after you normalize the slice to have an area of one so that it becomes a univariate density).

The **Density Ellipse** command in the Fit Y by X platform gives the correlation, which is a measure, on a scale of -1 to 1, of how close two variables are to being linearly related. A significance test on the correlation shows the *p*-value for the hypothesis that there is no correlation. A low *p*-value indicates that there is a significant correlation.

The next sections uses simulations to cover the bivariate normal distribution in more detail.

**Note:** Density ellipses can also be shown by using the Graph Builder.



## Correlations and the Bivariate Normal

Describing normally distributed bivariate data is easy because you need only the means, standard deviations, and the correlation of the two variables to completely characterize the joint distribution. If the individual distributions are not normal, you might need a good deal more information to describe the bivariate relationship.

Correlation is a measure of the linear association between two variables. If you can draw a straight line through all the points of a scatterplot, then there is a perfect correlation. Perfectly correlated variables measure the same thing. The sign of the correlation reflects the slope of the regression line—a perfect positive correlation has a value of 1, and a perfect negative correlation has a value of -1.

### Simulating Bivariate Correlations

As we saw in earlier chapters, using simulated data created with formulas provides a reference point when you move on to analyze real data.

☞ Select **Help > Sample Data Library** and open Corrsim.jmp.

This table has no rows, but contains formulas to generate correlated data.

☞ Select **Rows > Add Rows** and enter 1000 when prompted for the number of rows.

☞ Click **OK**.

The formulas evaluate to give simulated correlations (**Figure 16.6**). There are two independent standard normal random columns, labeled **X** and **y**. The remaining columns (**y.50**, **y.90**, **y.99**, and **y.100**) have formulas constructed to produce the level of correlation indicated in the column names (0.5, 0.9, 0.99, and 1.00). For example, the column formula to produce the correlation of 1 is:

$$r = 1;$$

$$\sqrt{r} \cdot X + \sqrt{1 - r} \cdot y;$$

**Figure 16.6** Partial Listing of Simulated Values

	X	y	y.50	y.90	y.99	y.100
1	1.316354	-0.69515	0.056157	0.881709	1.205127	1.316354
2	0.964078	-0.96446	-0.35321	0.44727	0.818383	0.964078
3	1.022198	0.584535	1.017321	1.174771	1.094435	1.022198
4	-0.83937	2.660953	1.984767	0.584449	-0.2576	-0.63937
5	-1.52217	-0.38538	-1.09483	-1.53793	-1.56131	-1.52217
6	0.081642	1.725379	1.535043	0.825553	0.32422	0.081642
7	0.336023	1.310936	1.303315	0.873844	0.517593	0.336023
8	0.054181	1.003179	0.895869	0.486038	0.195155	0.054181
9	-0.52785	-0.22078	-0.45512	-0.5713	-0.55372	-0.52785
10	0.996136	-0.27862	0.256779	0.775077	0.946871	0.996136
All rows	1,000					

You can use the Fit Y by X platform to examine the correlations:

- ✓ Select **Analyze > Fit Y by X** and assign **X** to **X, Factor**, and all the **Y** columns to **Y, Response**.
- ✓ Hold down the Ctrl (or ⌘) key and select **Density Ellipse** from any red triangle menu. Select **0.90** as the density level.

Holding down the Ctrl (or ⌘) key applies the command to all the open plots in the Fit Y by X window simultaneously.

- ✓ Do the previous step twice more with **0.95** and **0.99** as density parameters.

These steps make normal density ellipses (**Figure 16.7**) containing 90%, 95%, and 99% of the bivariate normal density, using the means, standard deviations, and correlation from the data.

As an exercise, create the same plot for generated data with a correlation of -1, which is the last plot shown in **Figure 16.7**. To do this:

- ❖ Create a new column and call it Y(-1.00) or whatever you want.
- ❖ Right-click the column header and select **Formula**.
- ❖ Enter the following formula.

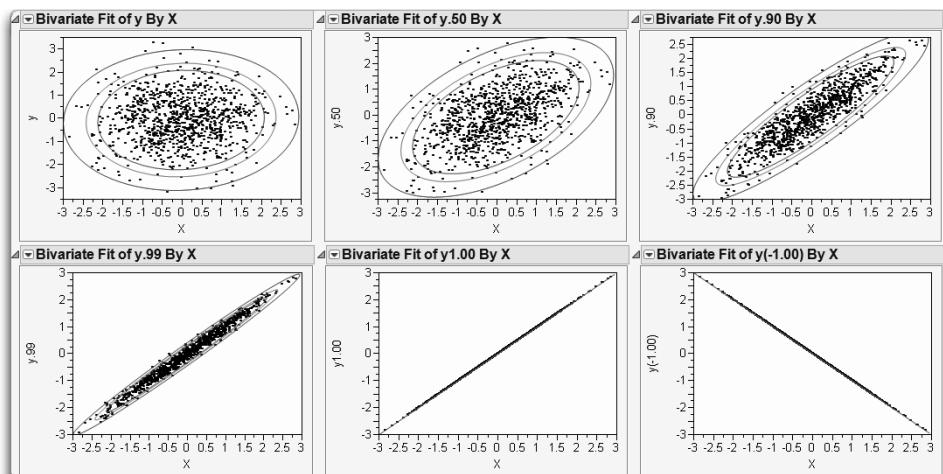
$$\begin{aligned} & \left( r = 1 ; \right. \\ & \left. \sqrt{r} \cdot X + \sqrt{1-r} \cdot y ; \right) \end{aligned}$$

**Note:** Open the Formula Editor window for the variable called Y1.00. Select its formula and drag it to the Formula Editor window for the new column that you are creating. To place a minus sign in front of the formula, select the whole formula and click the unary sign (the +/-) button on the Formula Editor keypad.

- ❖ Select **Analyze > Fit Y by X** to create the same plots for the new column as shown above.

Note in **Figure 16.7** that as the correlation grows from 0 to 1, the relationship between the variables gets stronger and stronger. The normal density contours are circular at correlation 0 (if the axes are scaled by the standard deviations) and collapse to the line at correlation 1.

**Figure 16.7** Density Ellipses for Various Correlation Coefficients



**Note:** The demoCorr teaching script provides an interactive tool for exploring correlations between two variables. For example, the script enables you to add

new points and drag points to see how the correlation changes for different data. To run this script, select **Help > Sample Data > Teaching Scripts > Teaching Demonstrations > demoCorr.**

## Correlations across Many Variables

The next example involves six variables. To characterize the distribution of a six-variate normal distribution, the means, the standard deviations, and the bivariate correlations of all the pairs of variables are needed.

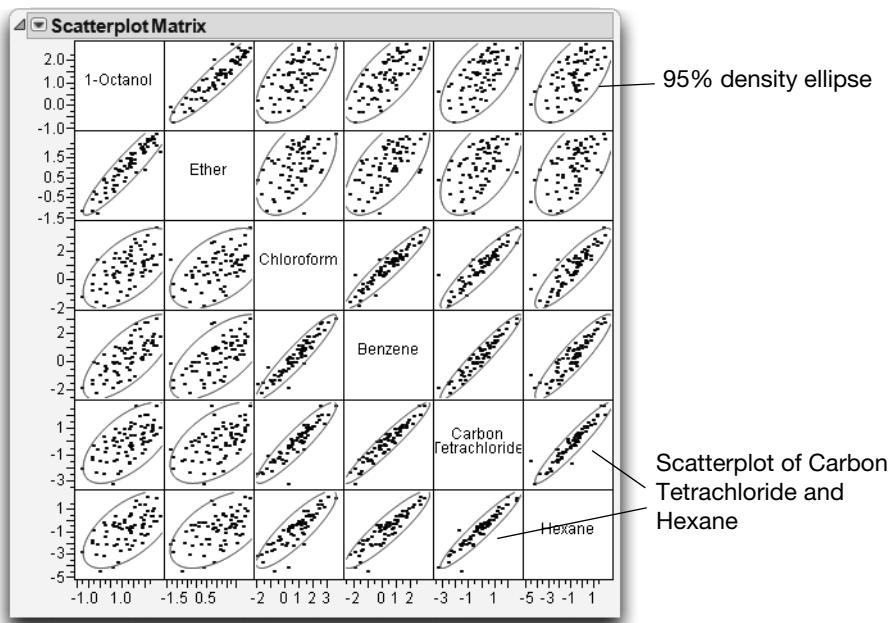
In a chemistry study, the solubility of 72 chemical compounds was measured with respect to six solvents (Koehler and Dunn, 1988). One purpose of the study was to see whether any of the solvents were correlated—that is, to identify any pairs of solvents that acted on the chemical compounds in a similar way.

- ☞ Select **Help > Sample Data Library** and open Solubility.jmp to see the Label (solvents) and the variables 1-Octanol, Ether, Chloroform, Benzene, Carbon Tetrachloride, and Hexane.

Note that the tag icon appears beside the Label column in the Columns panel to signify that JMP uses the values in this column to label points in plots.

- ☞ Select **Analyze > Multivariate Methods > Multivariate** and assign all six continuous solvent variables to **Y, Columns**.
- ☞ Click **OK**.

You see a correlations table and a scatterplot matrix like the one shown in **Figure 16.8**. Each small scatterplot is identified by the name cells of its row and column.

**Figure 16.8** Scatterplot Matrix for Six Variables

1-Octanol and Ether are highly correlated, but have weak correlations with other variables. The other four variables have strong correlations with one another. For example, note the correlation between Carbon Tetrachloride and Hexane.

You can resize the whole matrix by resizing any one of its small scatterplots.

- ❖ Move your mouse over the corner of any scatterplot until the cursor changes into a resize arrow. Click and drag to resize the plots.

Also, you can change the row and column location of a variable in the matrix by dragging its name on the diagonal.

Keep this report open to use again later in this chapter.

## Bivariate Outliers

Let's switch platforms to get a closer look at the relationship between Carbon Tetrachloride and Hexane using a set of density contours.

- ❖ Select **Analyze > Fit Y by X**.
- ❖ Assign Carbon Tetrachloride to **Y, Response** and Hexane to **X, Factor**, and then click **OK**.

- ☞ Select **Density Ellipse** from the red triangle menu next to Bivariate Fit to add four density contours to the plot (0.50, 0.90, 0.95, and 0.99), as shown in **Figure 16.9**.

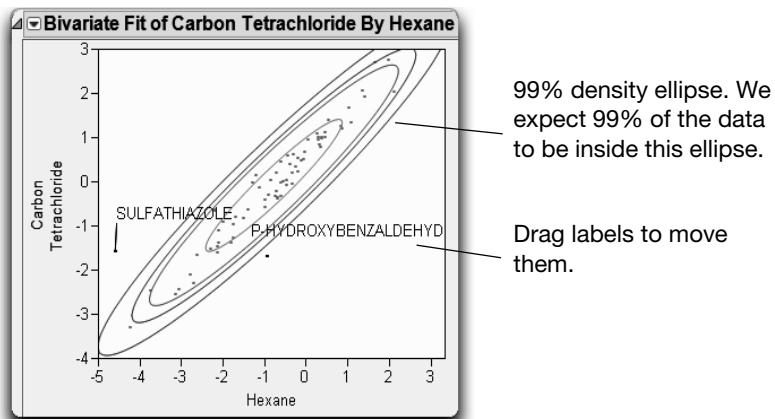
Under the assumption that the data are distributed bivariate normal, the inside ellipse contains half the points, the next ellipse 90%, then 95%, and the outside ellipse contains 99% of the points.

Note that there are two points that are outside even the 99% ellipse. To make your plot look like the one below:

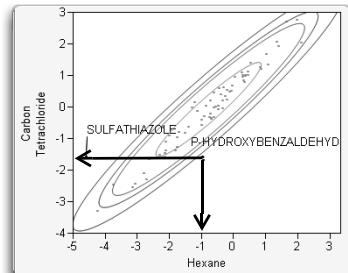
- ☞ Hold down the Shift key and select the two outside points.
- ☞ With the points highlighted, select **Rows > Label/Unlabel** to label them.
- ☞ To better read the labels, drag the label to an open area.

The labeled points are potential outliers. A point can be considered an *outlier* if its bivariate normal density contour is associated with a very low probability.

**Figure 16.9** Bivariate Plot With Ellipses and Outliers



Note that P-hydroxybenzaldehyde is not an outlier for either variable individually (that is, in a univariate sense). In the scatterplot, it is near the middle of the Hexane distribution, and is barely outside the 50% limit for the Carbon Tetrachloride distribution. However, it is a bivariate outlier because it falls outside the correlation pattern, which shows most of the points in a narrow diagonal elliptical area.

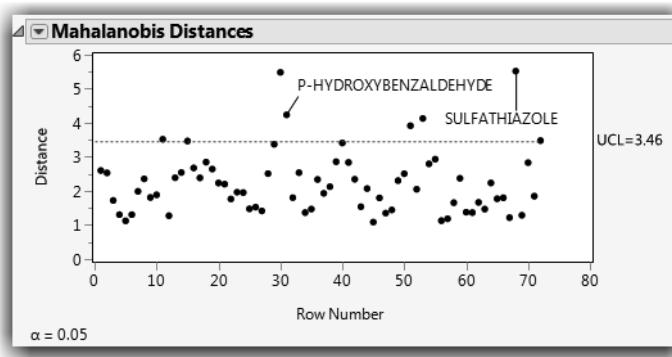


A common method for identifying outliers is the *Mahalanobis distance*. The Mahalanobis distance is computed with respect to the correlations as well as the means and standard deviations of both variables.

- ⇨ Click the Multivariate report that shows the scatterplot matrix to make it the active window.
- ⇨ Select **Outlier Analysis > Mahalanobis Distances** from the red triangle menu next to Multivariate.

This command gives the Mahalanobis Distance outlier plot shown in **Figure 16.10**.

**Figure 16.10** Outlier Analysis with Mahalanobis Outlier Distance Plot



The reference line is drawn using an *F*-quantile and shows the estimated distance that contains 95% of the points. In agreement with the ellipses in the bivariate plot, Sulfathiazole and P-hydroxybenzaldehyde show as prominent outliers (along with three other solvents). To explore these two outliers, follow these steps:

- ⌚ Make sure the two outliers are selected in the Mahalanobis distances plot. These two points are also selected in each scatterplot in the scatterplot matrix.
- ⌚ Look at the scatterplot for Carbon Tetrachloride and Hexane. Note that the two outliers are outside the density ellipse for these two variables.
- ⌚ Click on other points in the Mahalanobis distances plot. Note that points that are far below the reference line fall within all of the density ellipses.

## Outliers in Three and More Dimensions

In the previous section, we looked at two variables at a time—bivariate correlations for several pairs of variables. In this section, we discuss correlations that involve more than two variables at a time. First, let's do a little cleanup.

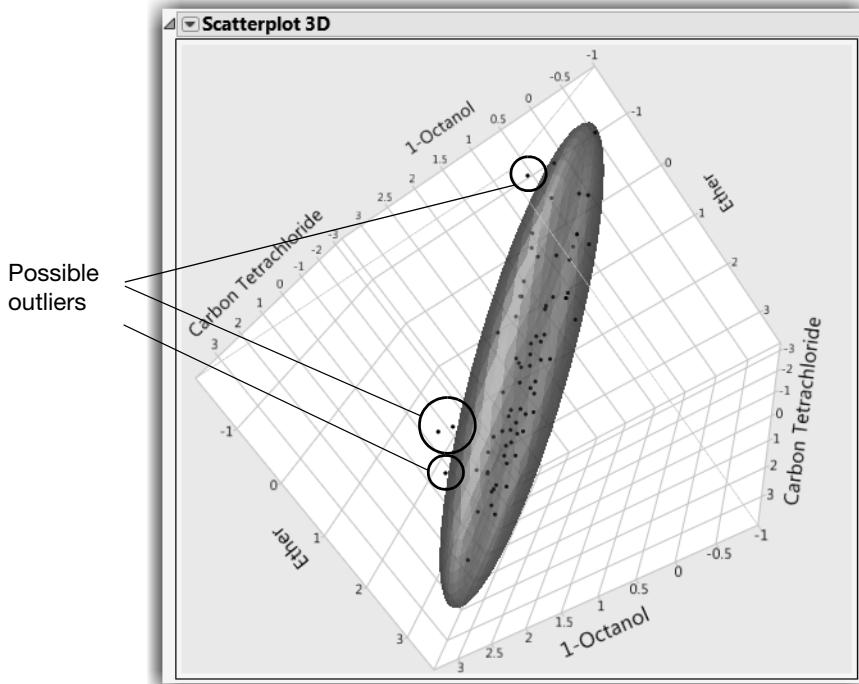
- ⌚ Select **Rows > Clear Row States** to unlabel the two data points.

To consider three variables at a time, look at three chemical variables, 1-Octanol, Ether, and Carbon Tetrachloride.

You can see the distribution of points with a 3-D scatterplot:

- ⌚ Select **Graph > Scatterplot 3D** and assign 1-Octanol, Ether, and Carbon Tetrachloride to **Y, Columns**.
- ⌚ Click **OK**.
- ⌚ Select **Normal Contour Ellipsoids** from the red triangle menu next to Scatterplot 3D, change the coverage to 0.9, and then click **OK** to see the three-dimensional ellipsoid in **Figure 16.11**.
- ⌚ Click in the 3-D plot and drag to rotate the plot and look for three-variate outliers.

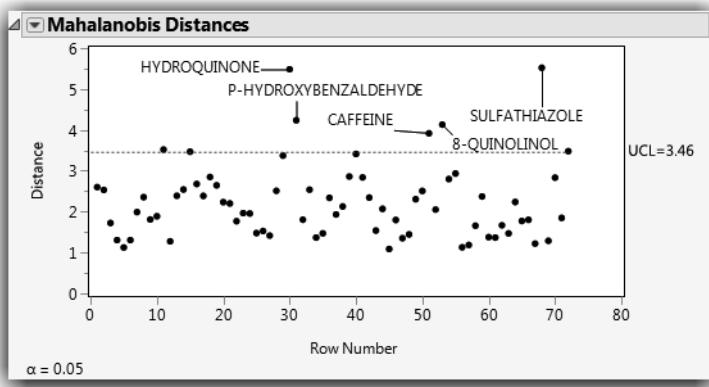
The orientation in **Figure 16.11** shows four points that appear to be outlying from the rest of the points with respect to the ellipsoid-shaped distribution.

**Figure 16.11** Outliers as Seen in Scatterplot 3D

The Fit Y by X and the Scatterplot 3D platforms revealed extreme values in two and three dimensions. How do we measure outliers in six dimensions? All outliers, from two dimensions to six dimensions (and beyond), should show up on the Mahalanobis distances outlier plot shown earlier. This plot measures distance with respect to all six variables.

If you closed the scatterplot window, select **Analyze > Multivariate Methods > Multivariate** with all six responses as Y variables. Then select **Outlier Analysis > Mahalanobis Distances** from the red triangle menu next to Multivariate.

Potential outliers across all six variables are labeled in **Figure 16.12**.

**Figure 16.12** Outlier Distance Plot for Six Variables

There is a refinement to the outlier distance that can help to further isolate outliers. When you estimate the means, standard deviations, and correlations, all points—including outliers—are included in the calculations and affect these estimates, causing an outlier to disguise itself.

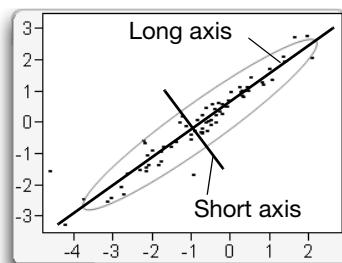
Suppose that as you measure the outlier distance for a point, you exclude that point from all mean, standard deviation, and correlation estimates.

This technique is called *jackknifing*. The jackknifed distances often make outliers stand out better.

- ❖ To see the jackknifed distance plot, select **Outlier Analysis > Jackknife Distances** from the red triangle menu next to Multivariate.

## Identify Variation with Principal Components Analysis

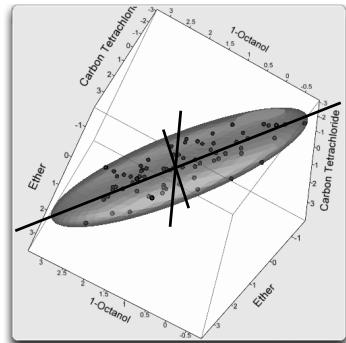
Lets return to the 3-D scatterplot. First, select **Rows > Clear Row States** to unlabel the outliers. As you rotate the plot, points in one direction in the space show a lot more variation than points in other directions. This is true in two dimensions when variables are highly



correlated. The long axis of the normal ellipse shows more variance; the short axis shows the least (as shown here).

Now, in three dimensions, there are three axes with a three-dimensional ellipsoid for the trivariate normal (shown here). The solubility data seem to have the ellipsoidal contours characteristic of normal densities, except for a few outliers (as noted earlier).

The directions of the axes of the normal ellipsoids are called the *principal components*. They were mentioned in the Galton example in “Why It’s Called Regression” on page 269 of Chapter 10, “Fitting Curves through Points: Regression.”



The *first principal component* is defined as the direction of the *linear combination of the variables* that has maximum variance. In a 3-D scatterplot, it is easy to rotate the plot and see which direction this is.

The *second principal component* is defined as the direction of the linear combination of the variables that has maximum variance. The second principal component is subject to it being at a right angle (orthogonal) to the first principal component. Higher principal components are defined in the same way. There are as many principal components as there are variables. The last principal component has little or no variance if there is substantial correlation among the variables. This means that there is a direction for which the normal density hyper-ellipsoid is very thin.

The Scatterplot 3D platform can show principal components.

- ⇨ Click the Scatterplot 3D report shown in **Figure 16.11** to make it active.
- ⇨ Select **Principal Components** from the red triangle menu next to Scatterplot 3D.

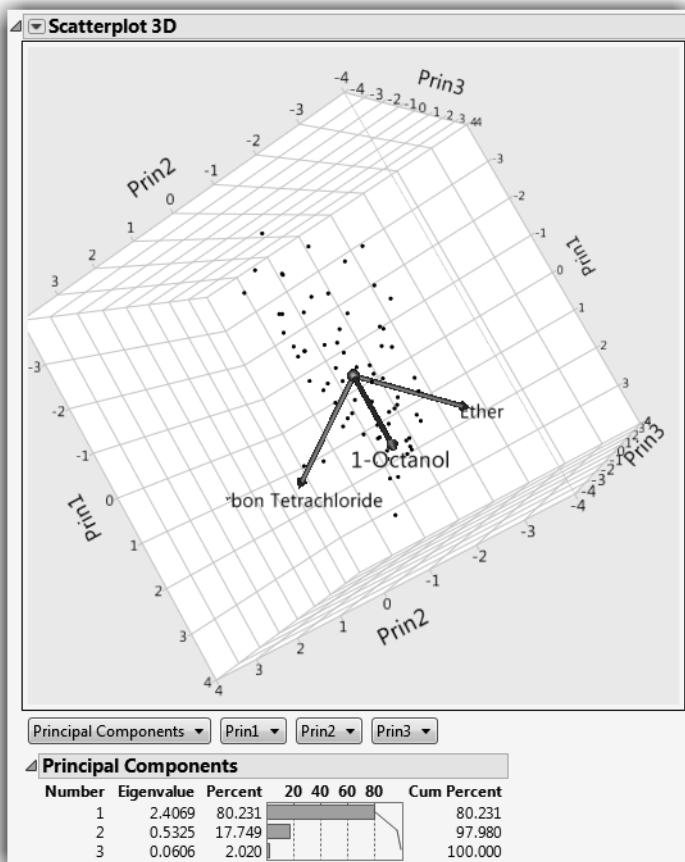
This adds three principal components to the variables list and creates a biplot with three rays corresponding to each of the variables, as shown in **Figure 16.13**.

The directions of the principal components are shown as rays from the origin of the point cloud. As you rotate the plot, you see that the principal component rays correspond to the directions in the data in decreasing order of variance. You can

also see that the Principal Components form right angles in three-dimensional space.

The Principal Components report in **Figure 16.13** shows what portion of the variance among the variables is explained by each principal component. In this example, 80% of the variance is carried by the first principal component, 18% by the second, and 2% by the third. It is the correlations in the data that make the principal components interesting and useful. If the variables are not correlated, then the principal components all carry the same variance.

**Figure 16.13** Biplot and Report of Principal Components



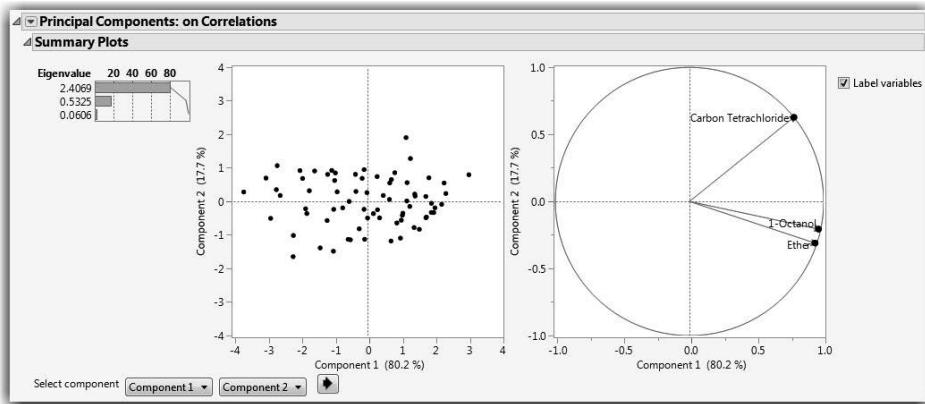
Principal Components are also available from the Principal Components platform. Let's repeat this analysis using this platform:

⇨ Select **Analyze > Multivariate Methods > Principal Components**.

- Again, assign 1-Octanol, Ether, and Carbon Tetrachloride to **Y, Columns**, and then click **OK**.

JMP displays the principal components and three plots, as shown in **Figure 16.14**.

**Figure 16.14** Principal Components for Three Variables



The Pareto plot (left) shows the percent of the variation accounted for by each principal component. As we saw earlier, the first principal component explains 80% of the variation in the data, while the second component accounts for 18%.

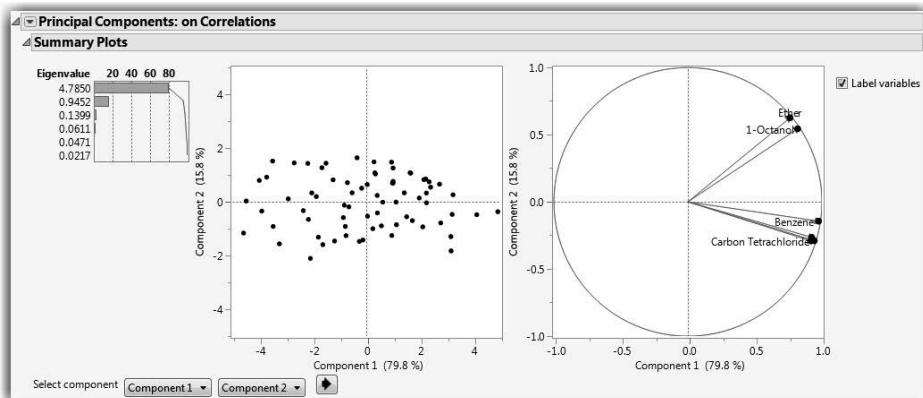
The middle score plot shows a scatterplot of the first two principal components. The first principal component accounts for far more variation (spread in the data) than the second component.

The loading plot (right) shows the correlations between the variables and the principal components. Carbon Tetrachloride is plotted by itself, while the other two variables are plotted together, indicating that they are correlated with one another. (We'll talk more about this in the next section.)

## Principal Components for Six Variables

Now let's move up to more dimensions than humans can visualize. We return to the Solubility.jmp sample data, but this time we look at principal components for all six variables in the data table.

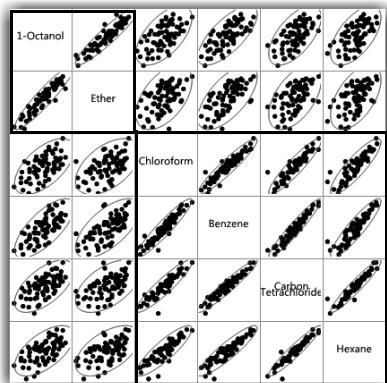
- Select **Analyze > Multivariate Methods > Principal Components**.
- Assign the six solvent variables to **Y, Columns** and then click **OK** to see the results in **Figure 16.15**.

**Figure 16.15** Principal Components for Six Variables

Examine the Pareto plot in **Figure 16.15**. Note that the first two principal components work together to account for 95.5% of the variation in six dimensions. To better see this, select **Eigenvalues** from the red triangle menu next to Principal Components.

Note the very narrow angle between rays for 1-Octanol and Ether in the loading plot (on the left). These two rays are at near right angles to the other four rays, which are clustered together. Narrow angles between principal component rays are a sign of correlation. Principal components try to squish the data from six dimensions down to two dimensions. To represent the points most accurately in this squish, the rays for correlated variables are close together because they represent most of the same information. Thus, the loading plot shows the correlation structure of high-dimensional data.

Recall the scatterplot matrix for all six variables (shown right). This plot confirms the fact that Ether and 1-Octanol are highly correlated, and the other four variables are also highly correlated with one another. However, there is little correlation between these two sets of variables.



## How Many Principal Components?

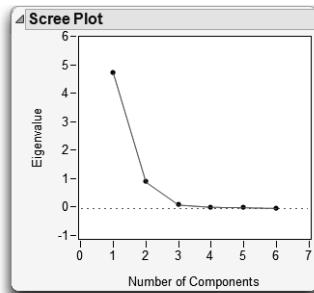
In the solubility example, more than 95% of the variability in the data is explained by two principal components. In most cases, only two or three principal components are needed to explain the majority (80-90%) of the variation in the data.

A scree plot can offer some guidance in determining the number of principal components required. Scree is a term for the rubble that accumulates at the bottom of steep cliffs, which this plot resembles. The place where the Scree Plot changes from a sharp downward slope to a more level slope (which is not always obvious) indicates the number of principal components to include.

- ⇨ From the Principal Components Analysis report, select **Scree Plot** from the red triangle menu next to Principal Components.

In the plot shown here, we can see that the plot levels out slightly after the second component, and is flat from the third component on.

Once you decide on the number of principal components needed, you can save the principal components to the data table.



- ⇨ Select **Save Principal Components** from the red triangle menu next to Principal Components.
- ⇨ Enter the desired number of principal components. The default is two.
- ⇨ Click **OK**.

This creates new columns in the data table, which might be useful in subsequent analyses and graphs.

**Note:** You might also consider a simple form of factor analysis, in which the components are rotated to positions so that they point in directions that correspond to clusters of variables. In JMP, **Factor Analysis** is an option in the Principal Components platform. Factor Analysis is also available in the Consumer Research menu.

## Discriminant Analysis

Both discriminant analysis and cluster analysis classify observations into groups. The difference is that discriminant analysis has known groups to predict; cluster analysis forms groups of points that are close together, but there is no known grouping of the points.

Discriminant analysis is appropriate for situations where you want to classify a categorical variable based on values of continuous variables. For example, you might be interested in the voting preferences (Democrat, Republican, or Other) of people of various ages and income levels. Or, you might want to classify animals into different species based on physical measurements of the animal.

There is a strong similarity between discriminant analysis and logistic regression.

- In logistic regression, the classification variable is random and predicted by the continuous variables.
- In discriminant analysis, the classifications are fixed, and the continuous factors are random variables.

However, in both cases, a categorical value is predicted by continuous variables.

Discriminant analysis is most effective when there are large differences among the mean values of the different groups. Larger separations of the means make it easier to determine the classifications.

The group classification is done by assigning a point to the group whose multivariate mean (centroid) is the closest, where closeness is with respect to the within-group covariance.

The example in this section deals with a trace chemical analysis of *cherts*. Cherts are rocks formed mainly of silicon, and are useful to archaeologists in determining the history of a region. By determining the original location of cherts, inferences can be drawn about the peoples that used them in tool making. Klawiter (2000) was interested in finding a model that predicted the location of a chert sample based on a trace element analysis. A subset of his data is found in the sample data table Cherts.jmp.

☞ Select **Help > Sample Data Library** and open Cherts.jmp.

☞ Select **Analyze > Multivariate Methods > Discriminant**.

- ✓ Assign all the chemical names to **Y, Covariates**.
- ✓ Assign location name to **X, Categories**, and then click **OK**.

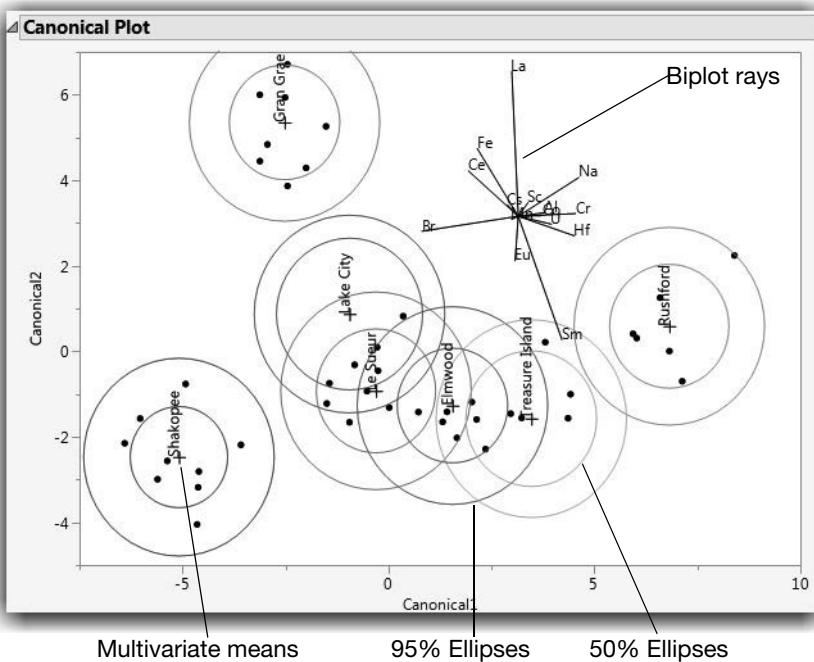
The discriminant analysis report consists of three outlines: the canonical plot, Discriminant Scores, and Score Summaries.

**Note:** The default discriminant method is Linear, which assumes a common covariance structure for all groups. Three other methods, Quadratic, Regularized, and Wide Linear, are also available. For details, select **Help > JMP Help** and refer to the *Multivariate Methods* book.

## Canonical Plot

The canonical plot in **Figure 16.16** shows the points and multivariate means in the two dimensions that best separate the groups. The term *canonical* refers to functions that discriminate between groupings. The first function, Canonical1, provides the most discrimination or separation between groups, and Canonical2 provides the second most separation.

Note that the biplot rays, which show the directions of the original variables in the canonical space, have been moved to better show the canonical graph. Click in the center of the biplot rays and drag them to move them around the report.

**Figure 16.16** Canonical Plot of the Cherts Data

The canonical plot also displays the following items:

- the multivariate mean for each group
- a 95% confidence ellipse for the mean of each group (the inner ellipse)
- a 50% density ellipse for each group, encompassing approximately 50% of the observations

In this example, the multivariate means for Shakopee, Gran Grae, and Rushford are more separated from the cluster of locations near the center of the graph.

## Discriminant Scores

The scores report shows how well each point is classified. A portion of the discriminant scores report for this example is shown in **Figure 16.17**. The report provides the actual classification, the distance to the mean of that classification, and the probability that the observation is in the actual classification. The histogram shows  $-\log(\text{prob})$ , the loss in log-likelihood when a point is predicted poorly. When the histogram bar is large, the point is being poorly predicted. The last columns show the predicted group, the associated probability, and other predictions.

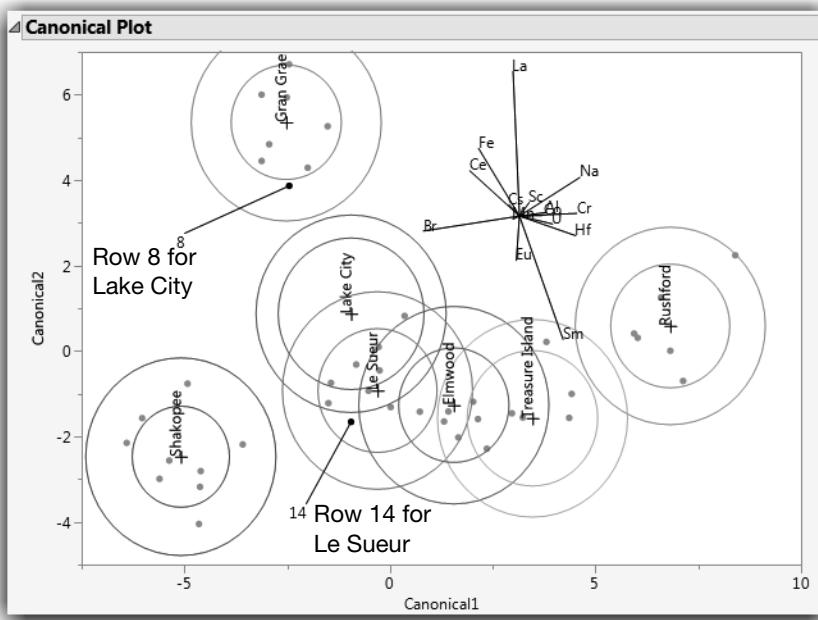
**Figure 16.17** Portion of Discriminant Scores Report

Row	Actual	SqDist(Actual)	Prob(Actual)	-Log(Prob)		Predicted	Prob(Pred)	Others
1	Gran Grae	15.52602	0.9998	0.000		Gran Grae	0.9998	
2	Gran Grae	9.85080	1.0000	0.000		Gran Grae	1.0000	
3	Gran Grae	16.71123	0.9998	0.000		Gran Grae	0.9998	
4	Gran Grae	16.58494	1.0000	0.000		Gran Grae	1.0000	
5	Gran Grae	31.52995	1.0000	0.000		Gran Grae	1.0000	
6	Gran Grae	12.07035	1.0000	0.000		Gran Grae	1.0000	
7	Gran Grae	7.57424	1.0000	0.000		Gran Grae	1.0000	
8	Lake City	20.11112	0.2069	1.576		*	Gran Grae	0.7931
9	Lake City	13.37920	0.6913	0.369		Lake City	0.6913	Le Sueur 0.30
10	Lake City	12.18366	0.9998	0.000		Lake City	0.9998	
11	Lake City	17.01456	0.9938	0.006		Lake City	0.9938	
12	Le Sueur	8.33448	0.9601	0.041		Le Sueur	0.9601	
13	Le Sueur	17.39611	0.9997	0.000		Le Sueur	0.9997	
14	Le Sueur	16.52688	0.3018	1.198		*	Elmwood	0.6975
15	Le Sueur	29.56057	1.0000	0.000		Le Sueur	1.0000	
16	Le Sueur	13.80015	1.0000	0.000		Le Sueur	1.0000	
17	Le Sueur	8.28931	0.8840	0.123		Le Sueur	0.8840	Elmwood 0.11
18	Rushford	9.81253	0.9987	0.001		Rushford	0.9987	

The predictions for rows 8 and 14 are incorrect, noted by an asterisk to the right of the bar plot. To see why these rows were misclassified, examine them in the canonical plot.

- ⇨ From the red triangle menu next to Discriminant Analysis, select **Score Options > Select Misclassified Rows.**

The result of this selection is shown in **Figure 16.18**.

**Figure 16.18** Misclassified Rows

Row 8, although actually from Lake City, is very close to Gran Grae in canonical space. This closeness is the likely reason it was misclassified. Row 14, on the other hand, is close to Le Sueur, its actual value. It was misclassified because it was closer to another center, though this is not apparent in this two-dimensional projection of a seven-dimensional space.

Another quick way to examine misclassifications is to look at the Score Summaries report (**Figure 16.19**) found below the discrimination scores. The overall misclassification rate is summarized, and actual versus predicted classifications are provided for each category. Zeros on the non-diagonal entries indicate perfect classification. The misclassified rows 8 and 14 are represented by the 1s in the non-diagonal entries.

**Figure 16.19** Counts Report

Actual location name	Predicted Count						Treasure Island
	Elmwood	Gran Grae	Lake City	Le Sueur	Rushford	Shakopee	
Elmwood	7	0	0	0	0	0	0
Gran Grae	0	7	0	0	0	0	0
Lake City	0	1	3	0	0	0	0
Le Sueur	1	0	0	5	0	0	0
Rushford	0	0	0	0	6	0	0
Shakopee	0	0	0	0	0	9	0
Treasure Island	0	0	0	0	0	0	5

Misclassified rows

## Stepwise Discriminant Variable Selection

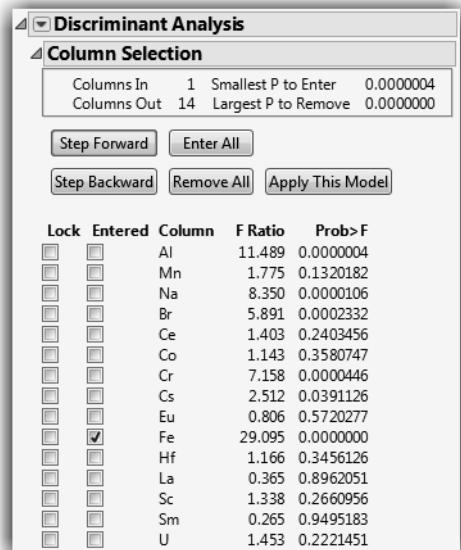
In discriminant analysis, we are building a model to predict which group a row belongs in. But, all variables in the data set might not aid in classification.

Stepwise variable selection determines which variables are useful, and allows us to build a model including only those variables.

JMP provides forward and backward variable selection. Select **Stepwise**

**Variable Selection** from the red triangle menu next to Discriminant Analysis to see the figure shown here:

- In forward selection (**Step Forward**), each variable is reviewed to determine which one provides the most discrimination between the groups. The variable with the lowest  $p$ -value is added to the model. The model-building process continues step by step, until you choose to stop.
- In backward selection (**Step Backward**), all variables are added to the model, and variables that contribute the least to discrimination are removed one at a time.



For the model that you choose (click **Apply this Model**), the discriminant analysis is performed and misclassification rates can be compared to the full model. Note that you can save probabilities and predicted classifications for the selected model to the data table. Select **Score Options > Save Formulas** from the red triangle menu next to Discriminant Analysis.

The stepwise option is available from the Discriminant launch window, or from the red triangle menu in the Discriminant report.

## Cluster Analysis

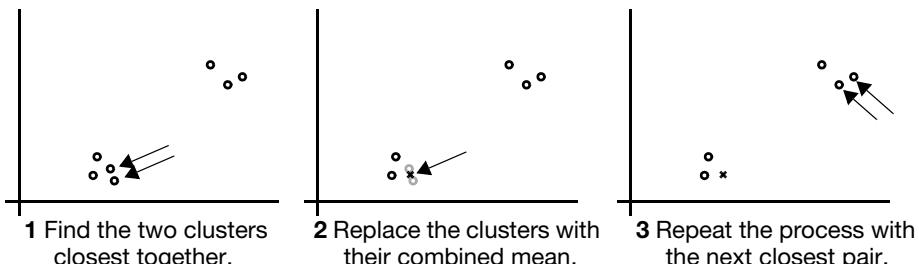
Cluster analysis is the process of dividing a set of observations into a number of groups where points inside each group are close to each other. JMP provides several methods of clustering, but here, we show only hierarchical clustering. Grouping observations into clusters is based on some measure of distance. JMP measures distance in the simple Euclidean way. There are many ways of measuring the proximity of observations. The essential purpose is to identify observations that are “close” to each other and place them into the same group. Each clustering method has the objective of minimizing within-cluster variation and maximizing between-cluster variation.

### Hierarchical Clustering: How Does It Work?

Hierarchical clustering works like this:

- Start with each point in its own cluster.
- Find the two clusters that are closest together in multivariate space.
- Combine these two clusters into a single group centered at their combined mean.
- Repeat this process until all clusters are combined into one group.

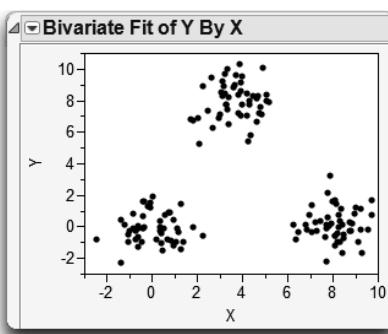
This process is illustrated in **Figure 16.20**.

**Figure 16.20** Illustration of Clustering Process

As a simple example, select **Help > Sample Data Library** and open **SimulatedClusters.jmp**.

- ✓ Select **Analyze > Fit Y By X**.
- ✓ Assign X to **X, Factor** and Y to **Y, Response**, and then click **OK**.

The results are shown in **Figure 16.21**.

**Figure 16.21** Scatterplot of Simulated Data

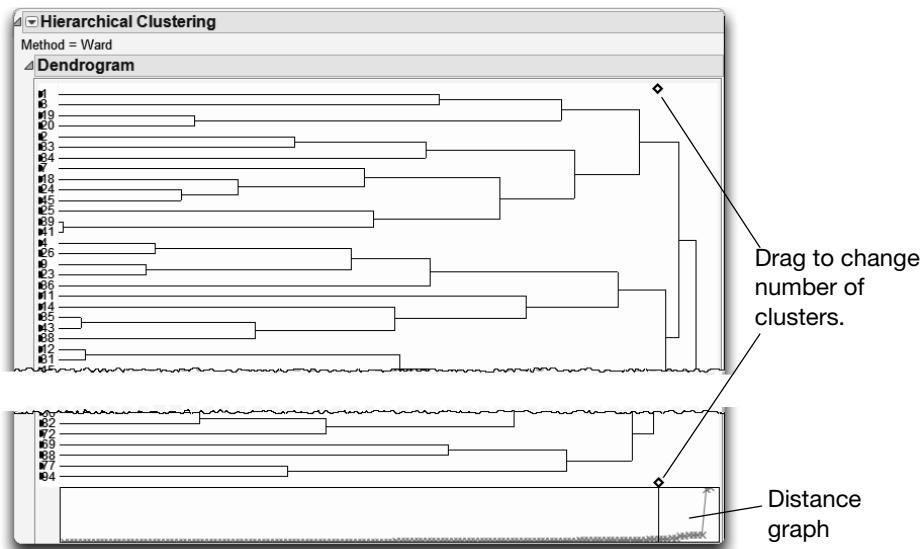
In this example, data clearly clump together into three clusters. In reality, clusters are rarely this clearly defined. We'll see a real life example of clustering later.

Let's see how the Cluster platform handles clustering with this simulated data.

- ✓ Select **Analyze > Clustering > Hierarchical Clustering**.
- ✓ Assign both X and Y to **Y, Response**, and then click **OK**.
- ✓ For illustration purposes, select **Dendrogram Scale > Even Spacing** from the red triangle menu next to Hierarchical Clustering.

The report appears as in **Figure 16.22**.

**Figure 16.22** Dendrogram and Distance Graph



The top portion of the report shows a dendrogram, a visual tree-like representation of the clustering process. Branches that merge on the left join together iteratively to form larger and larger clusters, until there is a single cluster on the right.

Note the small diamonds at the bottom and the top of the dendrogram. These dragable diamonds adjust the number of clusters in the model.

The distance graph, shown beneath the dendrogram, is similar to the scree plot in principal components. This plot offers some guidance regarding the number of clusters to include. The place where the distance graph changes from a level slope to a sharp slope is an indication of the number of clusters to include. In this example, the distance graph is very flat up to the point of three simulated clusters, where it rises steeply.

To better see the clusters:

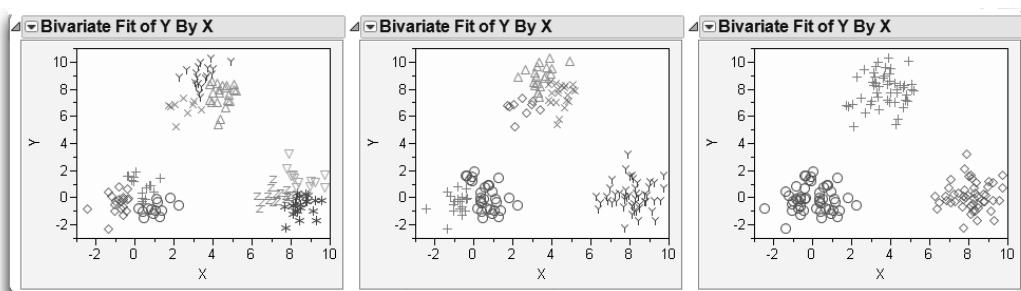
- ❖ From the red triangle menu next to Hierarchical Clustering, select both **Color Clusters** and **Mark Clusters**.

This assigns a special color and marker to each observation, which changes dynamically as you change the number of clusters in the model. To see this:

- ✓ Drag the windows so that you can see both the Fit Y By X scatterplot and the dendrogram at the same time.
- ✓ Drag the number of clusters diamond to the right. Drag it to nine clusters, and then to six clusters. Observe the changes in colors and markers in the scatterplot.
- ✓ Move the diamond to three clusters. This should correspond to the point of the sharp rise in the distance graph.

The scatterplot for nine, six, and three clusters should look similar to the ones in **Figure 16.23**.

**Figure 16.23** Nine, Six, and Three Clusters



Once you decide that you have an appropriate number of clusters, you can save a column in the data table that holds the cluster number for each point.

- ✓ From the red triangle menu next to Hierarchical Clustering, select **Save Clusters**.

The cluster numbers are often useful in subsequent analyses and graphs, as we'll see in the example to follow.

## A Real-World Example

The sample data table Cereal.jmp contains nutritional information about a number of popular breakfast cereals. A cluster analysis can show which of these cereals are similar in nutritional content and which are different.

- ✓ Select **Help > Sample Data Library** and open Cereal.jmp.

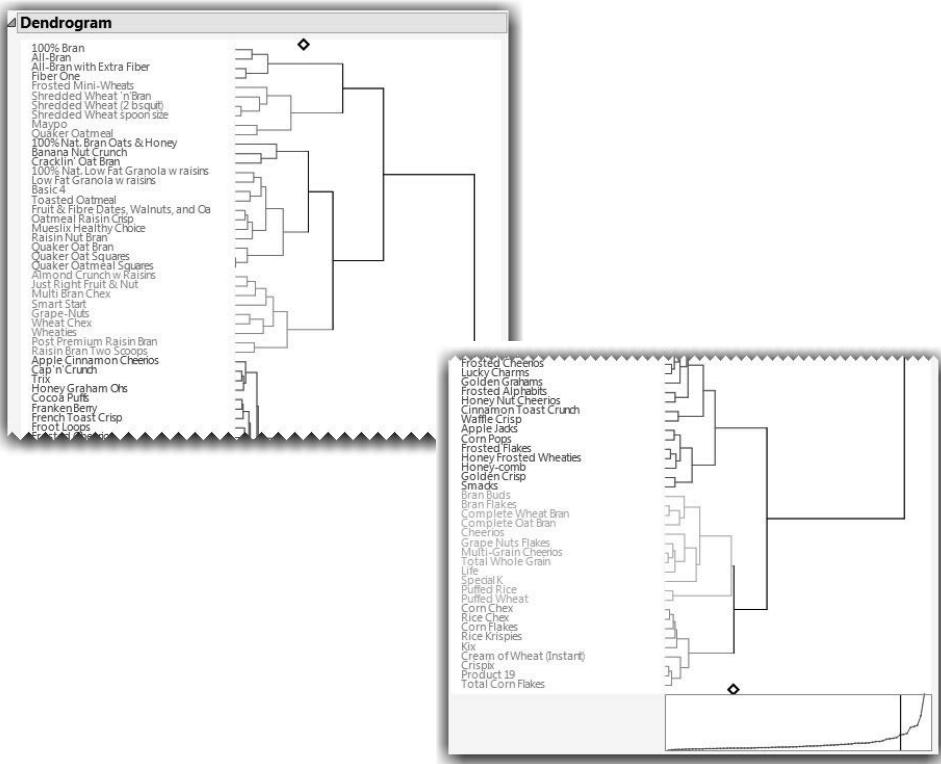
- ✓ Select **Rows > Clear Row States** to remove the labels and markers.
- ✓ Select **Analyze > Clustering > Hierarchical Cluster**.
- ✓ Select the Name column and assign it to **Label**.
- ✓ Assign all of the nutritional variables, from Calories through Potassium, to **Y, Columns**, and then click **OK**.

When the dendrogram appears:

- ✓ Select **Color Clusters** and **Mark Clusters** from the red triangle menu next to Hierarchical Clustering.

By default, eight clusters are displayed. The distance graph (bottom of **Figure 16.24**) does not show a clear number of clusters for the model. However, there seems to be some change in the steepness at around six clusters.

- ✓ Drag the number of clusters diamond to the point corresponding to six clusters, as in **Figure 16.24**.

**Figure 16.24** Dendrogram of Cereal Data

Examine the cereals classified in each cluster. Some conclusions based on these clusters are as follows:

- The all-bran cereals and shredded-wheat cereals form the top two clusters.
- Oat-based cereals, for the most part, form the third cluster.
- Cereals with raisins and nuts fall in the fourth cluster.
- Sugary cereals and light, puffed cereals form the last two clusters.

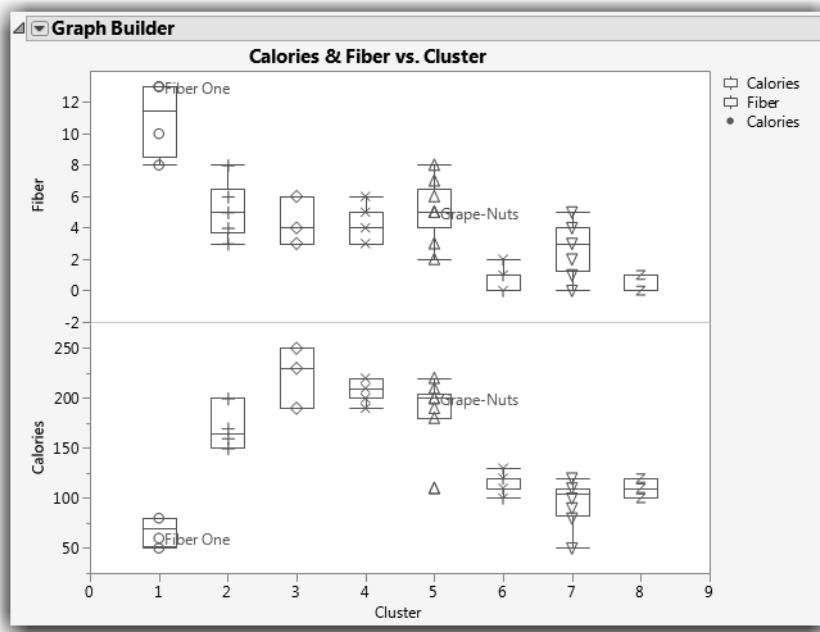
There are several built-in tools for exploring the clusters in the Hierarchical platform, including the cluster summaries, scatterplot matrix, and parallel coordinates plot. We encourage you to explore these options on your own.

You can also save cluster numbers to a column in the data table to help you gain additional insights into variables driving the clustering.

Select **Save Clusters** from the red triangle menu next to Hierarchical Clustering.

The Graph Builder is a handy tool for graphically exploring the clustering (from **Graph > Graph Builder**). For example, in **Figure 16.25**, we can see that cereals in Cluster 1 tend to be low in calories and high in fiber. However, cereals in Cluster 5 tend to be low in both calories and fiber.

**Figure 16.25** Exploring Clusters Using Graph Builder



## Some Final Thoughts

When you have more than three variables, the relationships among them can get very complicated. Many things can be easily found in one, two, or three dimensions, but it is difficult to visualize a space of more than three dimensions.

The histogram provides a good one-dimensional look at the distribution, with a smooth curve option for estimating the density. The scatterplot provides a good two-dimensional look at the distribution, with normal ellipses or bivariate smoothers to study it. In three dimensions, Scatterplot 3D provides the third dimension. To look at more than three dimensions, you must be creative and imaginative.

One good basic strategy for high-dimensional exploration is to take advantage of correlations and reduce the number of dimensions. The technique for this is principal components, and the aim is to reduce a large number of variables to a smaller more manageable set that still contains most of the information content in the data.

Discriminant analysis and cluster analysis are exploratory tools for classifying observations into groups. Discriminant analysis predicts classifications based on known groups, whereas cluster analysis forms groups of points that are close together.

You can also use highlighting tools to brush across one distribution and see how the points highlight in another view.

The hope is that you either find patterns that help you understand the data, or points that don't fit patterns. In both cases, you can make valuable discoveries.

## Exercises

1. In the sample data table Crime.jmp, data are given for each of the 50 U.S. states concerning crime rates per 100,000 people for seven classes of crimes.
  - (a) Use the Multivariate platform to produce a scatterplot matrix of all seven variables. What pattern do you see in the correlations?
  - (b) Conduct an outlier analysis (using Mahalanobis distance) of the seven variables, and note the states that seem to be outliers. Do the outliers seem to be states with similar crime characteristics?
  - (c) Conduct a principal components analysis on the correlations of these variables. Which crimes seem to group together?
  - (d) Using the eigenvalues from the principal components report, how many dimensions would you retain to accurately summarize these crime statistics?
2. The sample data table Socioeconomic.jmp (SAS Institute, 1988) consists of five socioeconomic variables for twelve census tracts in the Los Angeles Standard Metropolitan Statistical Area.

- (a) Use the Multivariate platform to produce a scatterplot matrix of all five variables.
  - (b) Conduct a principal components analysis (on the correlations) of all five variables. How many principal components would you retain for subsequent analysis?
  - (c) Using the loading plot, determine which variables load on each principal component.
3. A classic example of discriminant analysis uses Fisher's Iris data, stored in the sample data table Iris.jmp. Three species of irises (setosa, virginica, and versicolor) were measured on four variables (sepal length, sepal width, petal length, and petal width). Use the discriminant analysis platform to make a model that classifies the flowers into their respective species using the four measurements.
4. The sample data table World Demographics.jmp contains mortality (for example, birth and death) rates for several countries. Use the cluster analysis platform to determine which countries share similar mortality characteristics. What do you notice that is similar among the countries that group together?

Save the clusters and then use the Graph Builder to explore the clustering. Summarize what you observe.



# 17

## Exploratory Modeling

### Overview

*Exploratory modeling* (sometimes known as *data mining*) is the process of exploring large amounts of data, usually using an automated method, to find patterns and make discoveries. JMP has two platforms especially designed for exploratory modeling: the Partition platform and the Neural platform.

The Partition platform recursively partitions data, automatically splitting the data at optimum points. The result is a decision tree that classifies each observation into a group. The classic example is turning a table of symptoms and diagnoses of a certain illness into a hierarchy of assessments to be evaluated on new patients.

The Neural platform implements neural networks. Neural networks are used to predict one or more response variables from a flexible network of functions of input variables. They can be very good predictors, and are useful when the underlying functional form of the response surface is not important.

## Chapter Contents

Overview .....	517
Recursive Partitioning (Decision Trees) .....	519
Growing Trees .....	521
Exploratory Modeling with Partition .....	528
Saving Columns and Formulas .....	530
Neural Nets .....	531
A Simple Example .....	532
Modeling with Neural Networks .....	535
Saving Columns .....	535
Profiles in Neural .....	537
Exercises .....	541

## Recursive Partitioning (Decision Trees)

The Partition platform is used to form decision tree models. The platform recursively partitions a data set in ways similar to CART, CHAID, and C4.5. Recursively partitioning data is often taught as a data mining technique, for these reasons:

- It is good for exploring relationships without having a good prior model.
- It handles large problems easily.
- It works well with nominal variable and messy or unruly data.
- The results are very interpretable.

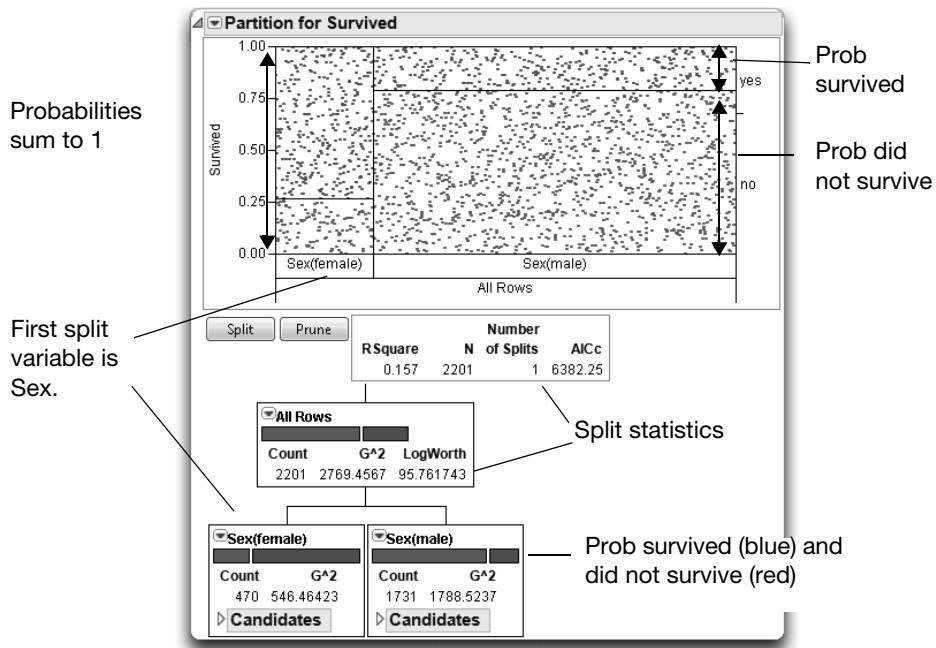
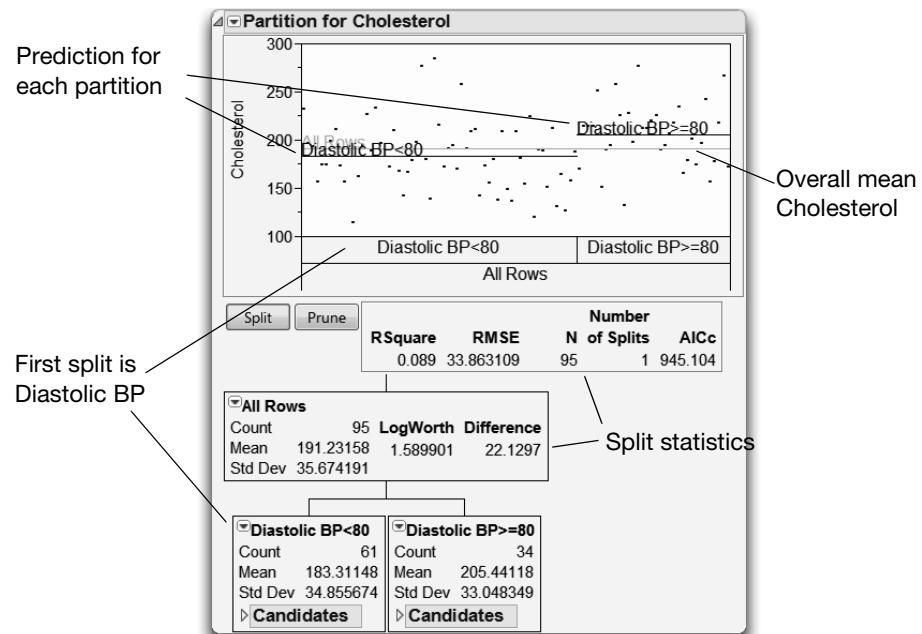
The factor columns ( $X$ s) can be either continuous or categorical. If an  $X$  is continuous, then the splits (partitions) are created by a *cutting value*. The sample is divided into values below and above this cutting value. If the  $X$  is categorical, then the sample is divided into two groups of levels.

The response column ( $Y$ ) can also be either continuous or categorical. If  $Y$  is continuous, then a regression tree is fit. The platform creates splits that most significantly separate the means by examining the sums of squares due to the differences in the means. If  $Y$  is categorical, then a classification tree is fit. The response rates (the estimated probability for each response level) become the fitted value. The most significant split is determined by the largest likelihood-ratio chi-square statistic ( $G^2$ ). In either case, the split is chosen to maximize the difference in the responses between the two branches of the split.

The Partition platform displays slightly different outputs, depending on whether the  $Y$ -variable is categorical or continuous.

**Figure 17.1** shows the partition plot and tree after one split for a categorical response. Points have been colored according to their response category. Each point in the partition plot represents the response category and the partition that it falls in. The  $x$ - and  $y$ -positions for each point are random within their corresponding partition. The partition width is proportional to the number of observations in the category, and the partition height is the estimated probability for that group.

**Figure 17.2** shows the case for a continuous response. The points are positioned at the response value, above or below the mean of the partition that they are in.

**Figure 17.1** Output with Categorical Response (Titanic Passengers.jmp)**Figure 17.2** Output for Continuous Responses (Lipid Data.jmp)

## Growing Trees

As an example of a typical analysis for a continuous response, select **Help > Sample Data Library** and open Lipid Data.jmp. These data contain results from blood tests, physical measurements, and medical history for 95 subjects. For examples using a categorical response, see the exercises at the end.

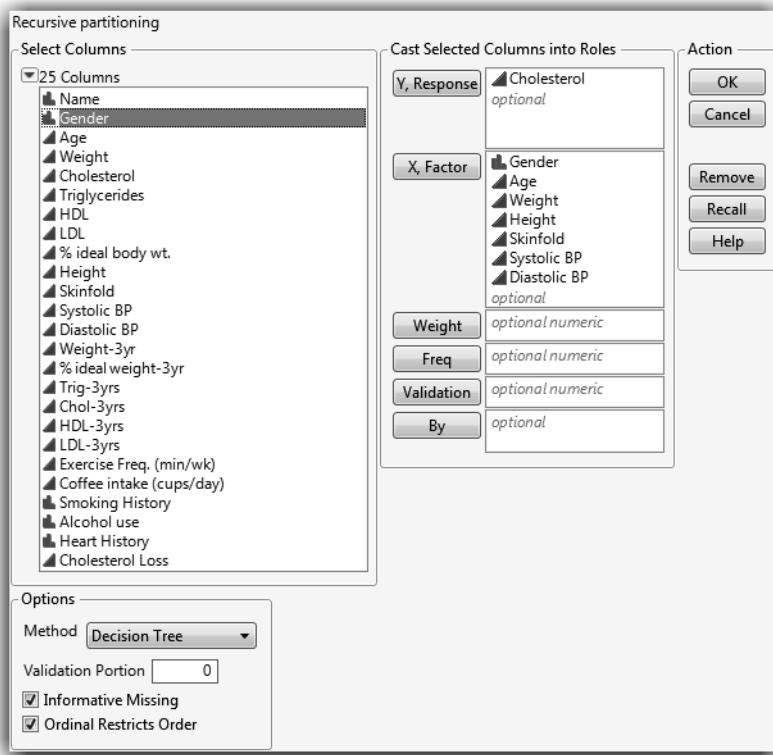
Cholesterol tests are invasive (requiring the extraction of blood) and require laboratory procedures to obtain results. Suppose these researchers are interested in using non-invasive, external measurements and also information from questionnaires to determine which patients are likely to have high cholesterol levels. Specifically, they want to predict the values stored in the Cholesterol column with information found in the Gender, Age, Height, Weight, Skinfold, Systolic BP, and Diastolic BP columns.

To begin the analysis:

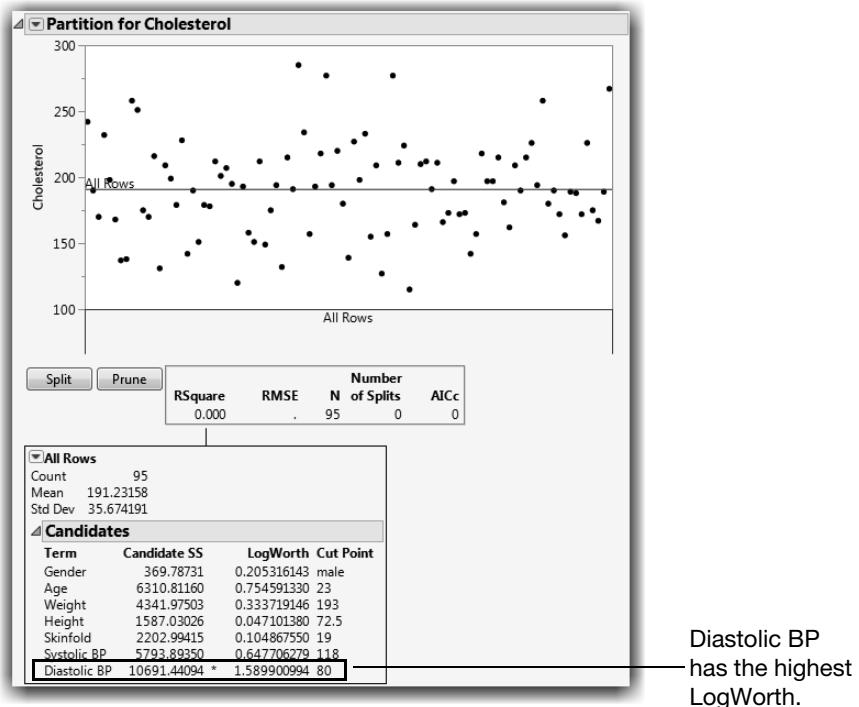
- ☞ Select **Analyze > Predictive Modeling > Partition**.
- ☞ Assign Cholesterol to **Y, Response**, and the **X, Factor** variables as shown in **Figure 17.3**.

**Note:** The default partitioning method is called the Decision Tree method. If you are using JMP Pro, a **Methods** menu appears at the lower left of the launch window, which offers four additional methods: **Bootstrap Forest**, **Boosted Tree**, **K Nearest Neighbors**, and **Naive Bayes**. JMP Pro also provides additional validation features.

- ☞ Click **OK** to see the results in **Figure 17.4**.

**Figure 17.3** Partition Launch Window

**Figure 17.4** shows the initial Partition report that appears. By default, the Candidates node of the report is closed, but is open here for illustration. Note that no partitioning has happened yet—all of the data are placed in a single group whose estimate is the mean cholesterol value (191.23).

**Figure 17.4** Initial Lipid Partition Report

The Partition platform looks at values of each X variable to find the optimum split. The variable that results in the highest reduction in total sum of squares is the optimum split, and is used to create a new branch of the tree.

As shown in **Figure 17.4**, the Candidate SS column for Diastolic BP results in a reduction of 10691.44094 in the total SS, so it is used as the splitting variable.

**Note:** The split chosen is based on the LogWorth statistic, which is the  $-\log_{10}(p\text{-value})$ . The LogWorth is based on an adjusted *p*-value, which takes into account the number of different ways splits can occur. The highest LogWorth statistic determines where the split actually occurs.

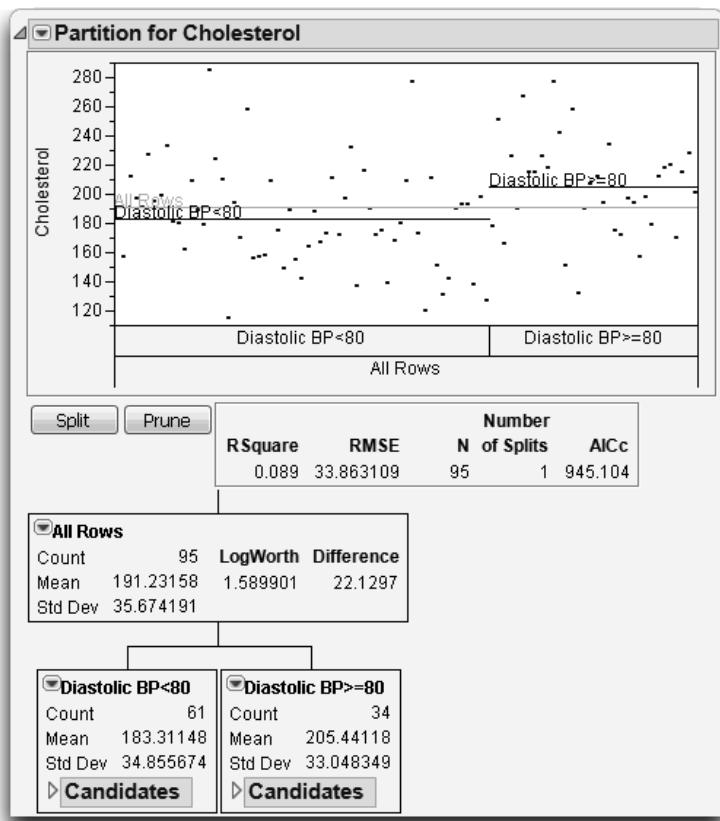
To begin the partitioning process, you interactively request splits.

- ☞ Click the **Split** button to split the data into two subsets.

As expected, the data are split on the Diastolic BP variable.

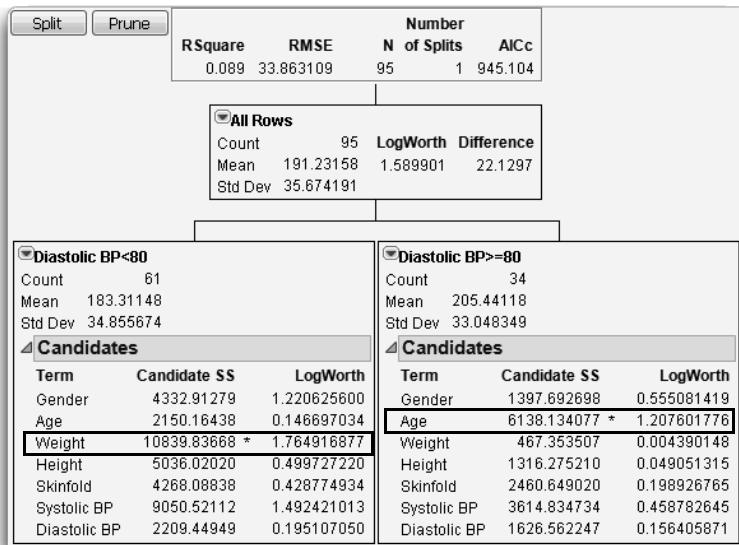
The complete report is shown in **Figure 17.5**. People with a diastolic blood pressure less than 80 tend to have lower cholesterol (a mean of 183.3) than those with blood pressure of 80 or more (with a mean of 205.4).

**Figure 17.5** First Split of Lipid Data



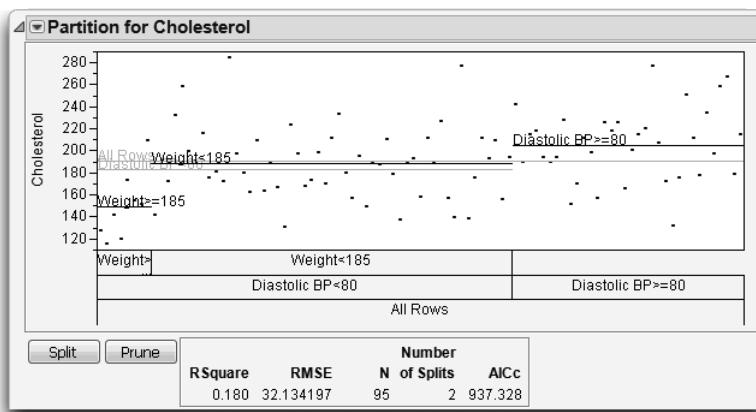
- To examine the candidates report, open the Candidates outline node in each of the Diastolic BP leaves.

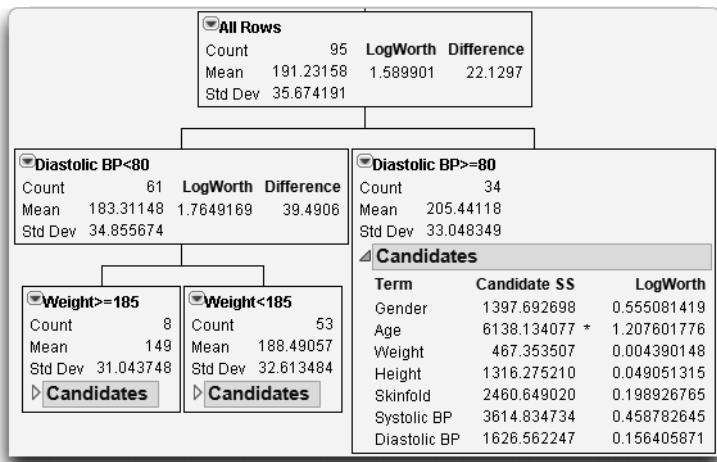
An examination of the candidates in **Figure 17.6** shows the possibilities for the second split. Under the Diastolic BP<80 leaf, a split using the Weight variable has a LogWorth of 1.76. The highest LogWorth under the Diastolic BP >=80 leaf is 1.21 for the Age variable. Therefore, you expect that pressing the **Split** button again will give two new weight leaves under the Diastolic<80 leaf, since Weight has the highest overall LogWorth.

**Figure 17.6** Candidates for Second Split

☞ Click the **Split** button to conduct the second split.

The resulting report is shown in **Figure 17.7**. Its corresponding tree is shown in **Figure 17.8**.

**Figure 17.7** Plot after Second Split

**Figure 17.8** Tree after Second Split

This second split shows that of the people with diastolic blood pressure less than 80, weight is the best predictor of cholesterol. For this group, the model predicts that those who weigh 185 pounds or more have an average cholesterol of 149. For those that weigh less than 185 pounds, the predicted average cholesterol is 188.5.

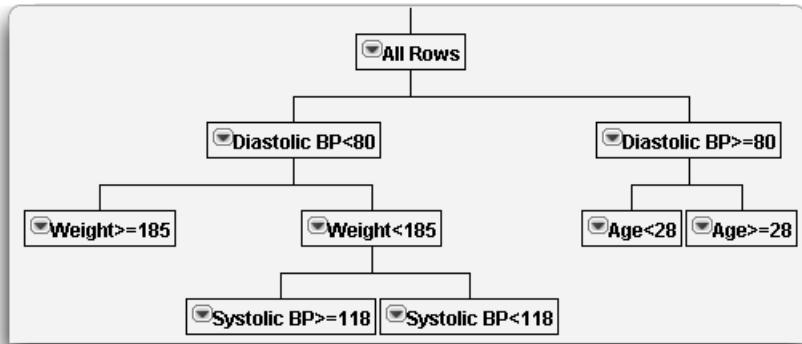
You can continue to split until you are satisfied with the predictive power of the model. As opposed to software that continues splitting until a criterion is met, JMP enables you to be the judge of the effectiveness of the model.

☞ Click the **Split** button two more times, to produce a total of four splits.

### Viewing Large Trees

With several levels of partitioning, tree reports can become quite large. JMP has several ways to ease the viewing of these large trees.

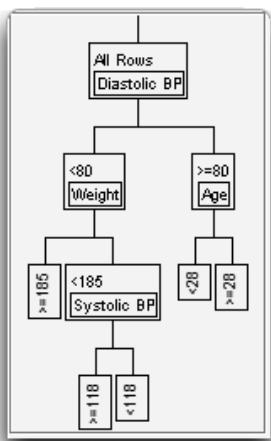
- Use the **Display Options** from the red triangle menu next to Partition to turn off parts of the report that are not needed. For example, **Figure 17.9** shows the current lipid results after four splits, with **Display Options > Show Split Stats** and **Display Options > Show Split Candidates** turned off.

**Figure 17.9** Lipid Data after Four Splits

You can also request a more compact version of the partition tree.

- ❖ Select **Small Tree View** from the red triangle menu next to Partition.

This option produces a compact view of the tree, appended to the right of the main partition graph. **Figure 17.10** shows the Small Tree View corresponding to **Figure 17.9**.

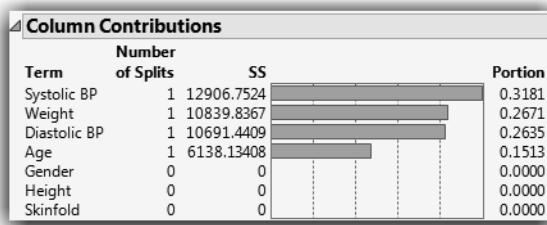
**Figure 17.10** Small Tree View

Additional options for exploring the tree, such as the Leaf Report, are available from the red triangle menu next to Partition.

### Viewing Column Contributions

You might want to see a summary of how many times a variable was split, along with the sums of squares attributed to that variable. This is particularly useful when building large trees involving many variables. The Column Contribution report provides this information. To see the report in **Figure 17.11** select **Column Contributions** from the red triangle menu next to Partition.

**Figure 17.11** Column Contributions after Four Splits



### Exploratory Modeling with Partition

You might wonder if there is an optimum number of splits to do for a given set of data. One way to explore recursive splitting is to use K Fold Crossvalidation. Lets do that now.

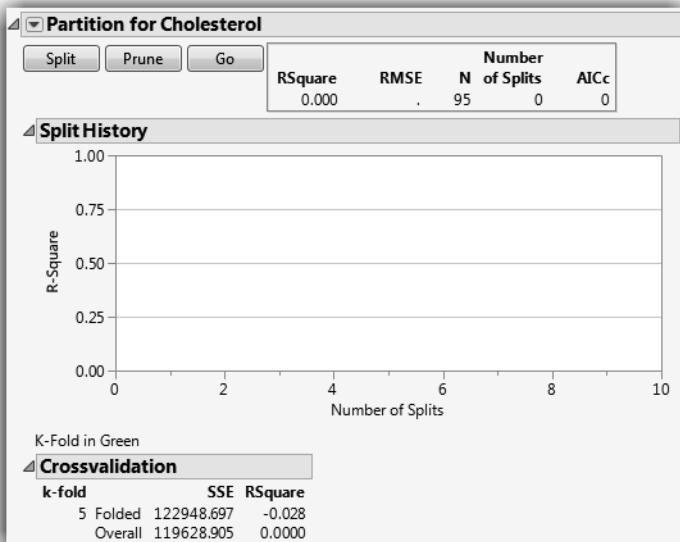
- ❖ Close the current partitioning of the Lipid Data to start over.
- ❖ Again select **Analyze > Predictive Modeling > Partition**. Click **Recall** to populate the window. If you haven't run the previous model, then assign variable roles as shown previously in **Figure 17.3**, and then click **OK**.
- ❖ When the Partition report appears, select the follow options from the red triangle menu next to Partition:
  - Select **Split History**.
  - Deselect **Display Options > Show Tree**.
  - Deselect **Display Options > Show Graph**.
  - Select **K Fold Crossvalidation**. When the cross validation window shows, use the default of 5 subgroups and click **OK**.

You have now customized the Partition platform to let you interactively split or prune the model step by step and observe the results. Furthermore, you are using a K Fold validation that randomly divides the original data into K subsets. Each of the K subsets is used to validate the model fit on the rest of the data, fitting a total

of K models. The model giving the best validation  $R^2$  statistic is considered the best model. The beginning partition platform should look like **Figure 17.12**.

**Figure 17.12** Initial Partition Platform with Cross Validation and History Plot

Click **Split**, and  
watch changes in  
the split statistics.

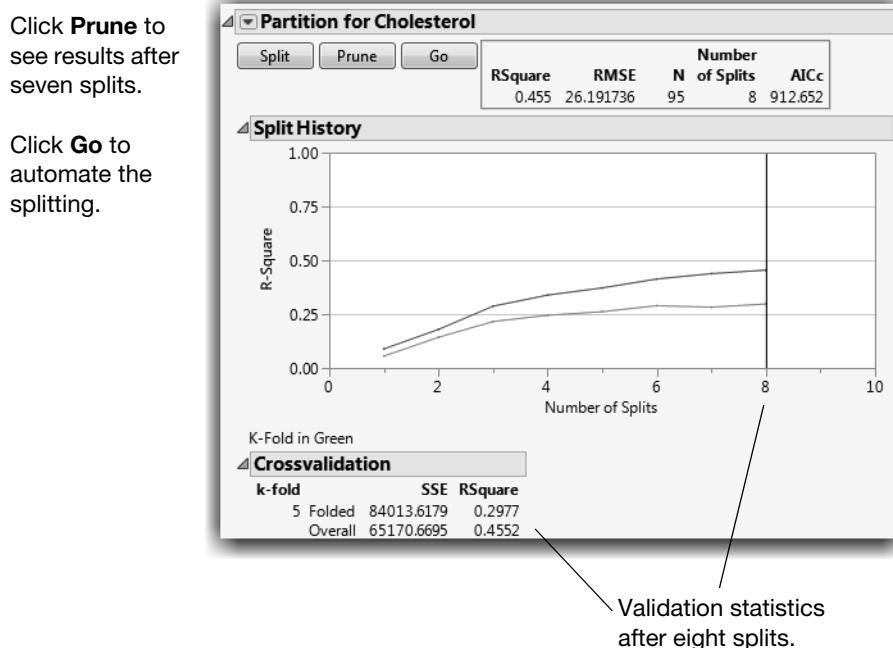


☞ Click **Split** eight times. Watch the change in the statistics at each split.

**Figure 17.13** shows the Split History after eight splits for this example. Your validation results might not be exactly the same as those shown because the validation subsets are randomly chosen.

If you continue to split until the data is completely partitioned, the model continues to improve fitting the data. However, this usually results in *overfitting*, which means the model predicts the fitted data well, but predicts future observations poorly, as seen by less desirable validation statistics.

We can see this in our example (**Figure 17.13**). After eight splits, the Overall  $R^2$  is 0.4552, but the K Fold  $R^2$  is 0.2977. From split seven to split eight the Overall  $R^2$  increased, but the validation  $R^2$  decreased. This could be an example of overfitting.

**Figure 17.13** Partition Platform with K-Fold Validation after Eight Splits

## Saving Columns and Formulas

Once you have partitioned the data to your satisfaction, you can save the partition model information.

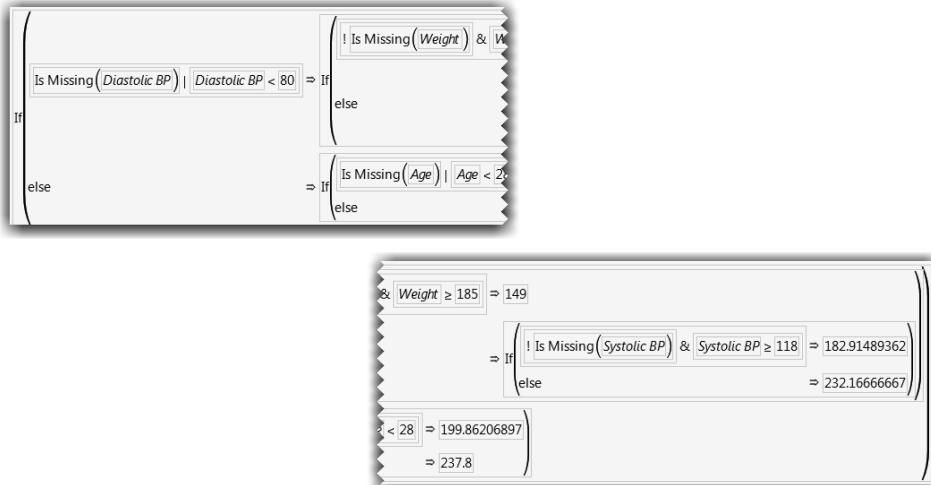
The **Save Columns** submenu provides options for saving results to the data table. For example, we return to the regression tree with four splits for Lipid Data.jmp.

- ☞ Select **Save Columns > Save Prediction Formula** from the red triangle menu next to Partition.

This adds a column to the report named Cholesterol Predictor that contains a formula using the estimates from the partitions of the tree.

To see the formula, return to the Lipid Data table and do the following:

- ☞ Right-click in the Cholesterol Predictor column header and select **Formula** from the menu that appears (**Figure 17.14**).

**Figure 17.14** Formula for Partition Model Example after Four Splits

You can see that the formula is a collection of nested If functions that use the partitioning results. You can copy and paste the formula to test its validity on other similar data.

**Note:** The Partition launch window in **Figure 17.13** includes the Informative Missing option, which is selected by default. This option tells JMP how to handle missing values. In **Figure 17.14**, we see that JMP includes missing values in the prediction formula. For more information, select **Help > JMP Help** and see the Partition chapter in the *Predictive and Specialized Modeling* book.

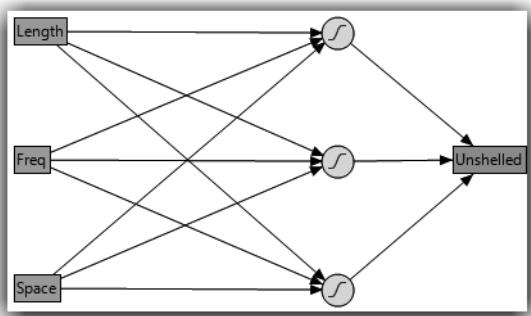
## Neural Nets

The Neural platform implements a connected network with one layer (or two layers in JMP Pro). Neural networks can predict one or more response variables using flexible functions of the input variables. This type of model can be a very useful when it is not necessary to do the following:

- describe the functional form of the response surface,
- to describe the relationship between the input variables and the response, or

- narrow down the list of important variables.

A neural network can be thought of as a function of a set of derived inputs called *hidden nodes*. Hidden nodes are nonlinear functions (called activation functions) of the original inputs. You can specify as many nodes as you want.



Inputs                  Nodes                  Prediction

At each node, the activation function transforms a linear combination of all the input variables using the S-shaped hyperbolic tangent function (TanH). The function then applied to the response is a linear combination of the nodes for continuous responses or a logistic transformation for categorical responses.

**Note:** Additional activation functions and other features are available in JMP Pro. This section continues with a simple example showing the abilities of the Neural platform in JMP. For more advanced examples that use JMP Pro features, and for technical details of the functions used in the neural network implementation, select **Help > JMP Help** and see the *Predictive and Specialized Modeling* book.

## A Simple Example

This section uses the sample data table Peanuts.jmp, from an experiment that tests a device for shelling peanuts. A reciprocating grid automatically shells the peanuts. The length and frequency of the reciprocating stroke, as well as the spacing of the peanuts, are factors in the experiment. Kernel damage, shelling time, and the number of unshelled peanuts need to be predicted. This example illustrates the procedure using only the number of unshelled peanuts as the response. A more involved neural modeling situation could have several factors, multiple responses, or both.

- ⌚ Select **Help > Sample Data Library** and open Peanuts.jmp.
- ⌚ Select **Analyze > Predictive Modeling > Neural**.
- ⌚ Assign Unshelled to **Y, Response** and Length, Freq, and Space to **X, Factor**.
- ⌚ Click **OK** to see the Neural control panel on the left in **Figure 17.15**.

**Note:** If you are using JMP Pro, or earlier versions of the Neural platform, your initial Control Panel has different options.

When the Control Panel appears, you have the option of selecting the validation method most suitable for your data.

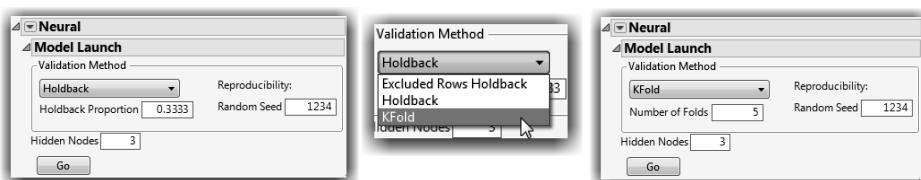
**Important:** Neural networks require some form of validation. Three validation methods are built into the Neural platform (additional validation options are available in JMP Pro). Validation designates that the original data be randomly divided into *training* and *validation* sets. The predictive ability of the response function derived on the training set is tested on the validation set. Validation statistics tell you how well the model fits data that were not used to fit the model.

The options on the Validation Method menu (middle of **Figure 17.15**) determine how the Neural fitting machinery subsets your data to test and decide on a final model:

- **Excluded Rows Holdback** uses row states to subset the data. Unexcluded rows are used as the training set, and excluded rows are used as the validation set.
- **Holdback**, the default, randomly divides the original data into the training and validation sets. You can specify the proportion of the original data to use as the validation set (holdback). The holdback proportion 0.333 is the default.
- **KFold** divides the original data into K subsets. Each of the K sets is used to validate the model fit on the rest of the data, fitting a total of K models. The model giving the best validation statistic is chosen as the final model.

Holdback, the default option, is the best used for larger samples (hundreds or thousands of observations). The **KFold** option is better used for smaller samples such as the peanuts data.

**Figure 17.15** Model Launch Control Panel



- For this example, select **KFold** from the Validation Method menu, and click **Go** to see Neural results similar to those shown in **Figure 17.16**.

**Note:** Your results are different because the Neural fitting process begins with a random seed. The random seed determines the starting point for the search algorithm. To produce the results shown in **Figure 17.16**, enter 1234 in the Random Seed field of the Model Launch control panel.

**Figure 17.16** Example Results from Neural Platform

The screenshot shows a JMP interface for a Neural network. At the top, it says "Neural" and "Validation: Random KFold". Below that is a red triangle menu labeled "Model Launch". Under "Model Launch" is a red triangle menu labeled "Model NTanH(3)". Inside "Model NTanH(3)" are two sections: "Training" and "Validation", each with a red triangle menu labeled "Unshelled".

Training		Validation	
Measures	Value	Measures	Value
RSquare	0.9522211	RSquare	0.898432
RMSE	8.6101243	RMSE	21.154849
Mean Abs Dev	6.2289721	Mean Abs Dev	19.679696
-LogLikelihood	57.150037	-LogLikelihood	17.883231
SSE	1186.1479	SSE	1790.1106
Sum Freq	16	Sum Freq	4

The reports in **Figure 17.16** give straightforward information for both the training and validation samples. You can see that the KFold number is reflected by 16 subjects in the training group and four in the validation group. The RSquare for the training group of 0.724 is very good, and the high RSquare for the validation group of 0.900 gives confidence that the model fits well. A Validation RSquare that is substantially lower than the Training RSquare indicates that the model is fitting noise rather than structure.

Neural networks are very flexible functions that have a tendency to *overfit* the data. When that happens, the model predicts the fitted data well, but predicts future observations poorly. However, the penalty system included in the Neural platform helps prevent the consequences of overfitting. If you are running JMP Pro, there is an option to select a specific penalty function.

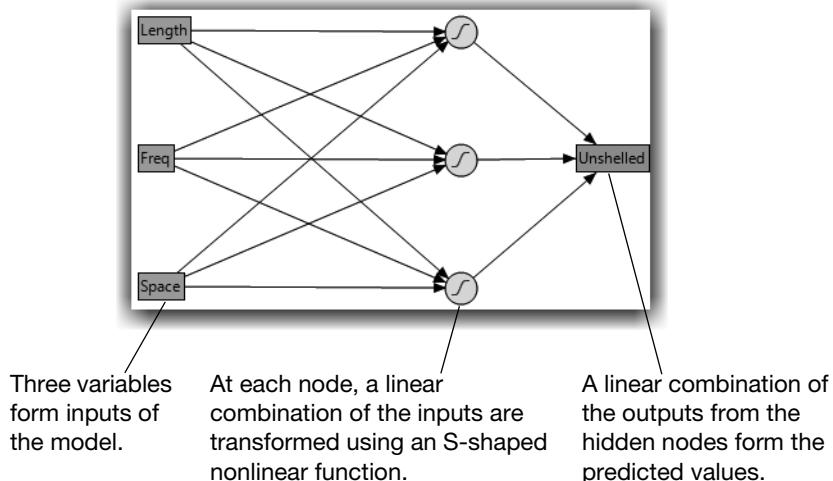
You can use Model Launch, shown open in **Figure 17.16**, and click **Go** to run the neural net fitting as many times as you want, or to run models with different numbers of nodes.

You can also request a diagram of the example showing how the factor columns are transformed through three hidden nodes, whose outputs are then combined to form the predicted values.

☞ Select **Diagram** from the red triangle menu next to Model NTanH(3).

You should see the diagram shown in **Figure 17.17**.

**Figure 17.17** Neural Net Diagram



## Modeling with Neural Networks

In the Neural Net diagram it's easy to see the inputs and the output, but the circle (hidden) nodes might seem more like black boxes. However, red triangle menu options for the Model let you see what is in the nodes and how they are used to produce the predicted output.

## Saving Columns

Like most analysis platforms, results from the Neural analyses can be saved to columns in the data table. The following save options are available.

**Save Formulas** creates new columns in the data table that contain formulas for the predicted response and for the hidden layer nodes. This option is useful if rows are added to the data table, because predicted values are automatically calculated.

**Save Profile Formulas** creates new columns in the data table that contain formulas for the predicted response. Formulas for the hidden nodes are embedded in this formula. This option produces formulas that can be used by the Flash version of the Profiler.

**Save Fast Formulas** creates new columns in the data table for the response variable, with embedded formulas for the hidden nodes. This formula evaluates

faster than the **Save Profile Formula** results but cannot be used in the Flash version of the Profiler.

- ☞ Select **Save Profile Formulas** from the red triangle menu next to Model NTanH(3) to create a new column in the Peanuts.jmp sample data table.

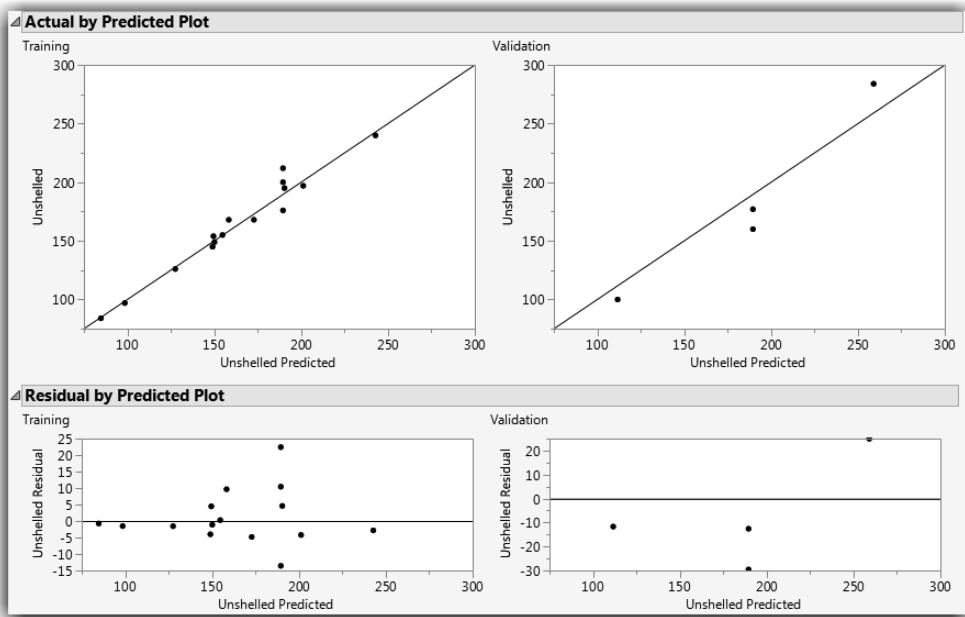
For this example, the new column, called Predicted Unshelled, contains the formula shown in **Figure 17.18**. You can see three TanH formulas for the three hidden nodes embedded in this prediction formula.

**Figure 17.18** Saved Profile Formula

$$\begin{aligned}
 & 75.091003497 \\
 & + 84.387188865 \cdot \text{Tanh} \left( 0.5 \cdot \left( 0.3458815859 + 0.3265717551 \cdot \text{Length} + -0.02287477 \cdot \text{Freq} + 6.4408973773 \cdot \text{Space} \right) \right) \\
 & + 57.331292214 \cdot \text{Tanh} \left( 0.5 \cdot \left( 18.391858994 + -5.033286182 \cdot \text{Length} + -0.043556885 \cdot \text{Freq} + -3.203189381 \cdot \text{Space} \right) \right) \\
 & + 78.245758171 \cdot \text{Tanh} \left( 0.5 \cdot \left( -11.41048396 + 0.0661458508 \cdot \text{Length} + 0.0556472107 \cdot \text{Freq} + 4.5832412823 \cdot \text{Space} \right) \right)
 \end{aligned}$$

- ☞ Next, return to the Neural analysis and select **Plot Actual by Predicted** and **Plot Residual by Predicted** from the red triangle menu next to Model NTanH(3) to see the plots in **Figure 17.19**.

The Actual by Predicted and the Residual by Predicted plots are similar to their counterparts in linear regression. Use them to help judge the predictive ability of the model. In this example, the model fits fairly well, and there are no glaring problems in the residual plots.

**Figure 17.19** Neural Net Plots

## Profiles in Neural

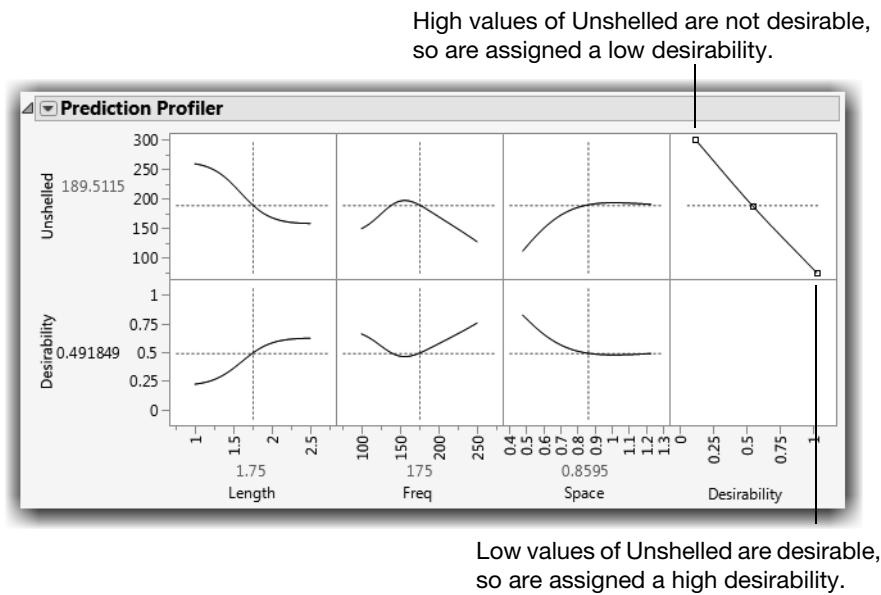
The slices through the response surface can be interesting and informative.

☞ Select **Profiler** from the red triangle menu next to Model NTanH(3).

The Prediction Profiler (**Figure 17.20**) clearly shows the nonlinear nature of the model. Running the model with more hidden nodes increases the flexibility of these curves; running with fewer stiffens them.

The Profiler has all of the features used in analyzing linear models and response surfaces (discussed in “Analyze the Model” on page 433 and in the *Profilers* book). Select **Help > JMP Help** and refer to the *Profilers* book.

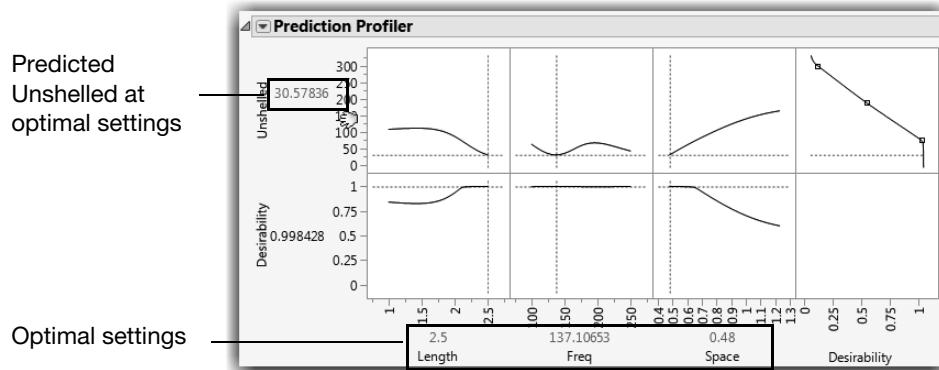
Since we are interested in minimizing the number of unshelled peanuts, we use the Profiler’s **Desirability Functions**, which are automatically shown (**Figure 17.20**).

**Figure 17.20** Prediction Profiler with Initial Settings

To have JMP automatically compute the optimum factor settings:

- ✓ Select **Optimization and Desirability > Maximize Desirability** from the red triangle menu next to Prediction Profiler.
- ✓ If needed, double-click the  $y$ -axis next to Unshelled in the Prediction Profiler and change the Minimum axis value to 0.

JMP computes the maximum desirability, which in this example is a low value of Unshelled. Results are shown in **Figure 17.21**.

**Figure 17.21** Prediction Profiler with Maximized Desirability Settings

Optimal settings for the factors show in red below the plots. In this case, optimal Unshelled values came from setting Length = 2.5, Freq = 137.1, and Space = 0.48. The predicted value of the response, Unshelled, dropped from 189.51 to 30.58.

**Note:** Recall that your results are different from the example shown here if you did not set the random seed 1234. A random seed is used to determine the starting point for the Neural fitting process.

In addition to seeing two-dimensional slices through the response surface, the Contour Profiler can be used to visualize contours (sometimes called level curves) and mesh plots of the response surface. The Surface Profilers also provide an interesting view of the response surface.

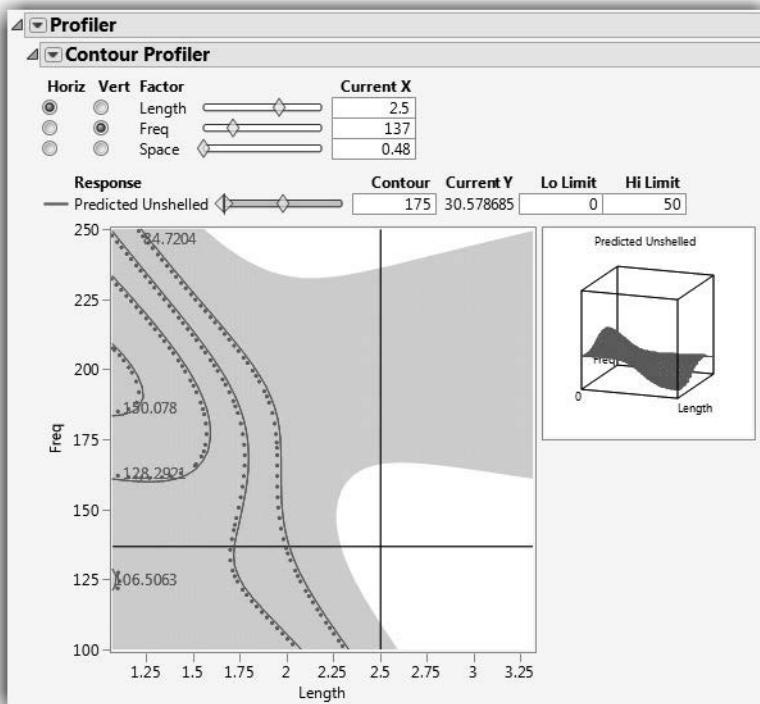
There are options on the red triangle menu next to Model in the Neural report for both of these plots. However, if you saved the prediction formula for the Neural model fit, you can replicate all the model plots at any time using commands in the **Graph** menu. Earlier, we showed how to save the profile formula in a single column (see **Figure 17.18**). You should have a column in the Peanuts table called Predicted Unshelled. Lets use **Graph** commands to see more views of the results.

- ⓐ Select **Graph > Contour Profiler**.
- ⓐ Select Predicted Unshelled and assign it to **Y, Prediction Formula**.
- ⓐ Click **OK** to see the initial contours and mesh plot for this example.
- ⓐ Select **Contour Grid** from the red triangle menu next to Contour Profiler and click **OK** to display more contours and contour values.

- ✓ Next, enter the optimal factor settings given by the Prediction Profiler in **Figure 17.21** (enter *your* settings).
- ✓ Finally, enter Lo Limit and Hi Limit values. Let's say we can tolerate up to 50 unshelled peanuts. The Lo Limit is 0 and the Hi Limit is 50, as shown.
- ✓ Adjust the *x*-axis and *y*-axis scaling if necessary.

In **Figure 17.21**, you can see that for these low and high limits and factor settings, the number of unshelled peanuts falls in an acceptable (unshaded) region. Click on the cross-hairs and drag to explore this region for the two variables selected. Note that, for this example, there is more than one acceptable region.

**Figure 17.22** Contour Profiler with Optimal Values



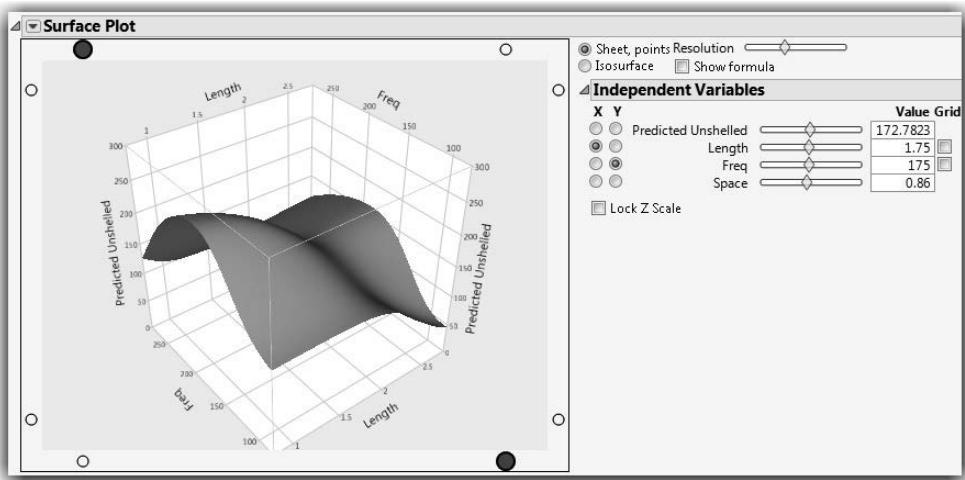
You can also see a response surface view of the number of unshelled peanuts as a function of two of the three factors, using a surface profiler (or surface plot, which has more features). The **Surface Profiler** is available in the Neural platform, or from any of the profilers available from the Graph menu. The **Surface Plot** command on the **Graph** menu can be used if you saved the prediction equation, as in this example.

- ✓ Select **Graph > Surface Plot.**
- ✓ On the launch window, assign Predicted Unshelled to **Columns** to assign it as the response.

The surface plot in **Figure 17.23** shows the surface when the optimal settings for the factors and the predicted value of unshelled peanuts are entered into the platform, as shown.

This plot shows just how nonlinear, or bendy, our fitted Neural model is.

**Figure 17.23** Surface Plot of Unshelled Peanuts



## Exercises

1. As shown in this chapter, the sample data table Lipid Data.jmp contains blood measurements, physical measurements, and questionnaire data from subjects in a California hospital. Repeat the Partition analysis of this chapter to explore models for these variables:
  - (a) HDL (good cholesterol - higher than 60 is considered protection against heart disease)
  - (b) LDL (bad cholesterol - less than 120 is optimal)
  - (c) Triglyceride levels (less than 150 is normal)

- Statistics provided by the American Heart Association (<http://www.heart.org>).
2. Use the sample data table Peanuts.jmp and the Neural platform to complete this exercise. The factors of Freq, Space, and Length are described earlier in this chapter. Use these factors for the following:
    - (a) Create a model for Time, the time to complete the shelling process. Use the Profiler and Desirability Functions to find the optimum settings for the factors that minimize the time of shelling.
    - (b) Create a model for Damaged, the number of damaged peanuts after shelling is complete. Find values of the factors that minimize the number of damaged peanuts.
    - (c) Compare the values found in the text of the chapter (**Figure 17.21**) and the values that you found in parts (a) and (b) of this question. What settings would you recommend to the manufacturer?
  3. The sample data table Mushroom.jmp contains 22 characteristics of 8,124 edible and poisonous mushroom. To see probabilities, select **Display Options > Show Split Prob** from the top red triangle menu.
    - (a) Use the Partition platform to build a seven split model to predict whether a mushroom is edible based on the 22 characteristics.
    - (b) Which characteristics are most important? **Note:** Use Column Contributions.
    - (c) What are the characteristics of edible mushrooms? **Note:** Use the Small Tree View and Leaf Report.
    - (d) Prune back to two splits. What is the predicted probability that a mushroom with no odor and a white spore print color is edible?
    - (e) Return to the Partition launch window, enter the value 0.3 in the Validation Portion field, and then rerun the model. We use a holdout validation set to determine the best number of splits. Do the validation statistics improve much after three splits? After four splits?

### Exploratory Modeling: A Case Study

The following exercise is a case study that examines characteristics of passengers on the Titanic. The response is whether an individual passenger survived or was lost. The case study uses many of the platforms introduced so far in this book.

1. Select **Help > Sample Data Library** and open Titanic Passengers.jmp, which describes the survival status of individual passengers on the Titanic. The response variable is Survived ("Yes" or "No"), and the variables of interest (factors or  $x$ -variables) are Passenger Class, Sex, Age, Siblings and Spouses, and Parents and Children.
  - (a) Use the Distribution platform and dynamic linking to explore the Survived variable and the variables listed above. Click on the bars for "Yes" and "No" in the Survived plot. Does Survived seem to be related to any of the other variables?
    - Did passengers who survived tend to fall in a particular Passenger Class?
    - Did they tend to be males or females?
    - Is Survived related to the other variables?
    - Do there appear to be any interactions? For example, does the relationship between Passenger Class and Survived depend on the Sex?
  - (b) Use Fit Y by X and Graph Builder platforms to further explore the relationship between Survived and the other variables.
    - Are any of the interactions significant?
    - Which main effects are significant?
    - Use the profiler (under the top red triangle menu) to explore the model. Drag the vertical red lines for each factor to see changes in the predicted survival rate. Since interactions were included, you also see changes in the profile traces for other factors.
    - For which group was the predicted survival rate the highest?
    - For which group was it the lowest? (Keep this window open).
  - (c) Use Fit Model to fit a logistic model for the response Survived and the five factors. Include the following interaction terms: Passenger Class\*Sex, Passenger Class\*Age, Sex\*Age.
    - Are any of the interactions significant?
    - Which main effects are significant?
    - Use the profiler (under the top red triangle menu) to explore the model. Drag the vertical red lines for each factor to see changes in the predicted survival rate. Since interactions were included, you also see changes in the profile traces for other factors.
    - For which group was the predicted survival rate the highest?
    - For which group was it the lowest? (Keep this window open).
  - (d) Use the Partition platform to build a classification tree for Survived, using the same five factors.
    - Select **Display Options > Show Split Prob** from the red triangle menu next to Partition, and then split the model several times.
    - What are the most important split variables?

- Do you see evidence of important interactions? For example, were the second and third splits on the same variable, or did it choose different split variables?
  - Compare these results to those found earlier using logistic regression. Do you come to similar conclusion, or are the conclusions very different? (Keep this window open as well).
- (e) Use the Neural platform to build a Neural Net for Survived, using the same factors. Click **Go** on the Neural control panel to accept the default model settings.
- Select the Profiler from the red triangle menu next to Model. Drag the vertical red lines for each factor to explore changes in the predicted survival rate.
  - How does this profiler compare to the one for the Logistic model found previously?
- (f) Using the three final models (logistic, partition, and neural), to determine the predicted survival rate for (1) a first-class female and (2) a 20-year-old man. Are the results comparable? **Hint:** Save the formulas for these models, and select **Profiler** from the **Graph** menu to compare the results.
- (g) Summarize your exploration of the Titanic data and conclusions in a form suitable for presentation. **Note:** Results can be saved in a variety of formats including PowerPoint and interactive HTML. Most JMP output can also be saved as an interactive Web report by selecting **View > Create Web Report**.



# 18

## Control Charts and Capability

### Overview

Some statistics are for proving things. Some statistics are for discovering things. And, some statistics are to keep an eye on things, watching to make sure something stays within specified limits.

The watching statistics are needed mostly in industry for production processes that sometimes stray from proper adjustment. These statistics monitor variation, and their job is to distinguish the usual random variation (called *common causes*) from abnormal changes (called *special causes*).

These statistics are usually from a time series, and the patterns that they exhibit over time are clues to what is happening to the production process. If they are to be useful, the data for these statistics need to be collected and analyzed promptly so that any problems they detect can be fixed.

The use of SQC techniques became popular in the 1980s as industry began to better understand the issues of quality, after the pioneering effort of Japanese industry and under the leadership of W. Edwards Deming and Joseph Juran.

This whole area of statistics is called *Statistical Process Control* (SPC) or *Statistical Quality Control* (SQC). The most basic tool is a graph called a *control chart* (or *Shewhart control chart*, named for the inventor, Walter Shewhart). In some industries, SQC techniques are taught to everyone—engineers, mechanics, shop floor operators, even managers.

In addition to control charts, JMP offers many quality and process tools, such as Pareto charts, measurement systems analysis, capability analysis, and cause and effect diagrams (also known as fishbone charts or Ishikawa diagrams).

This chapter provides an overview of control charts and process capability studies.

## Chapter Contents

Overview .....	545
What Does a Control Chart Look Like .....	548
Types of Control Charts .....	549
Variables Charts.....	550
Attributes Charts.....	551
Specialty Charts.....	551
Control Chart Basics .....	551
Control Charts for Variables Data .....	552
Variables Charts Using Control Chart Builder.....	553
The Control Chart Builder Work Space.....	553
Control Chart Builder Examples.....	554
Control Charts for Attributes Data .....	557
Specialty Charts .....	560
Presummarize Charts.....	560
Levey-Jennings Charts .....	561
Uniformly Weighted Moving Average (UWMA) Charts .....	561
Exponentially Weighted Moving Average (EWMA) Chart.....	563
Capability Analysis .....	564
What Is Process Capability? .....	564
Capability for One Process Measurement.....	567
Capability for Many Process Measurements .....	569
Capability for Time-Ordered Data .....	572
A Few Words about Measurement Systems .....	574
Exercises.....	574

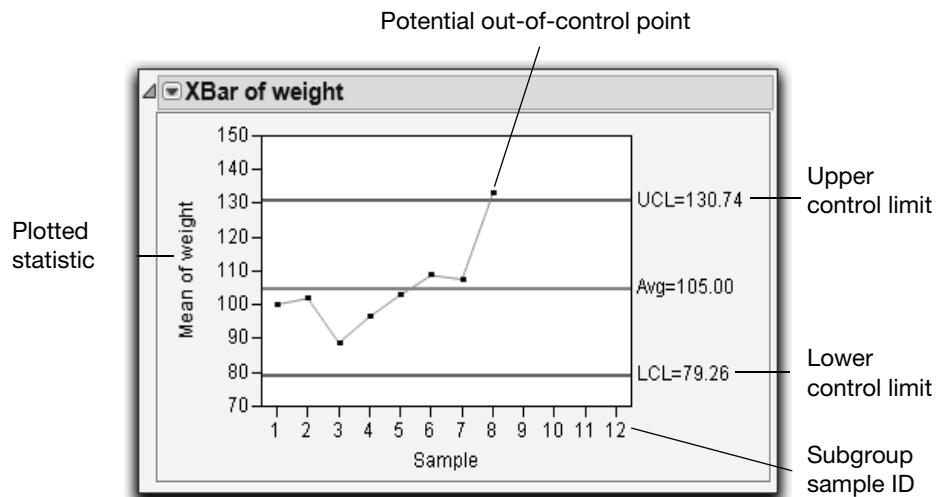
## What Does a Control Chart Look Like

Control charts are a graphical and analytical tool for deciding whether a process is in a state of statistical control. Control charts in JMP are automatically updated when rows are added to the current data table. In this way, control charts can be used to monitor an ongoing process.

The most common types of control charts can be created using the Control Chart Builder. Additional charts are available from the **Analyze > Quality and Process > Control Chart** menu commands.

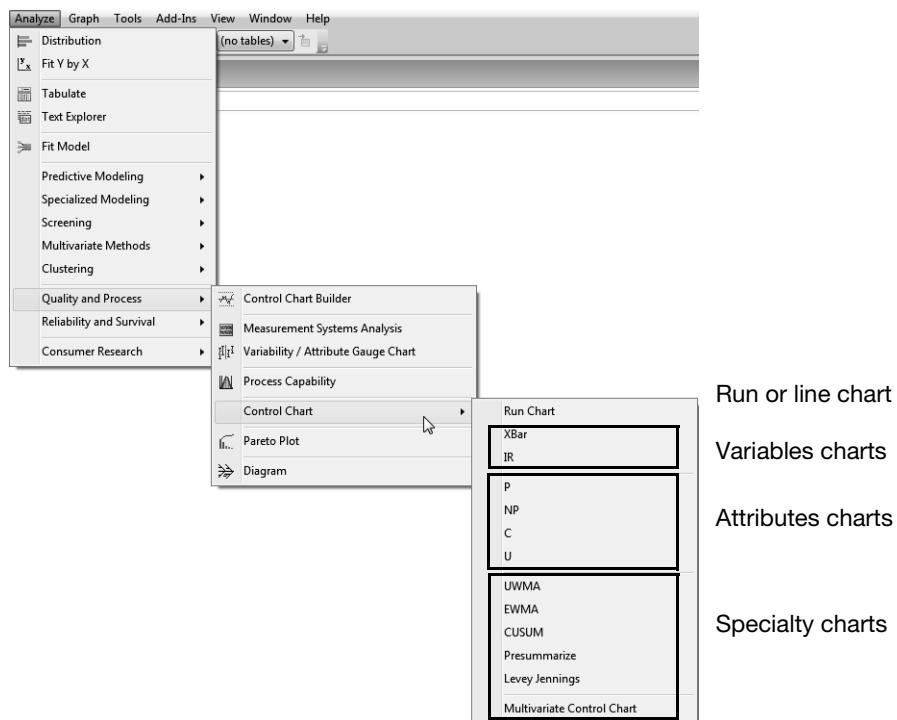
The example in **Figure 18.1** illustrates characteristics of most control charts:

- Each point represents a summary statistic computed from a subgroup sample of measurements of a quality characteristic.
- The vertical axis of a control chart is scaled in the same units as the summary statistics plotted on the chart.
- The horizontal axis of a control chart identifies the subgroup samples, which are sorted in time order.
- The center line on a control chart indicates the average (expected) value of the summary statistic when the process is in statistical control.
- The upper and lower control limits, labeled UCL and LCL, give the range of variation to be expected in the summary statistic when the process is in statistical control.
- A point outside the control limits signals that there might be a special cause of variation. (Note that there are other signals of special causes.)

**Figure 18.1** Control Chart Example

## Types of Control Charts

Control charts are broadly classified as *variables charts* and *attributes charts*. JMP also provides a number of specialty charts, as shown in **Figure 18.2**.

**Figure 18.2** The Control Chart Menu

The control charts in **Figure 18.2** are briefly described below.

## Variables Charts

Variables control charts are used when the quality characteristic to be monitored is measured on a continuous scale.

There are different types of variables control charts based on the subgroup sample summary statistic plotted on the chart. The plotted statistic can be the mean (average), the range, the standard deviation of a measurement, an individual measurement itself, or a moving range.

For quality characteristics measured on a continuous scale, it is typical to analyze both the process mean and its variability by showing an XBar (mean) chart aligned above its corresponding R- (range) or S- (standard deviation) chart. The center line on the XBar chart is the overall average, and each point is a subgroup mean.

If you are charting individual response measurements, the Individual Measurement chart is aligned above its corresponding Moving Range chart.

## Attributes Charts

Attributes control charts are used when the quality characteristic of a process is qualitative in nature. When quality is measured by counting the number of nonconformities (defects) in an item or batch of items, a C or U chart is used to monitor the process. When calculating the proportion of nonconforming (defective) items in a sample, a P or NP chart is used.

## Specialty Charts

The Control Chart menu also includes a command for run (or line) charts, along with a variety of specialty charts.

- UWMA and EWMA charts are used to plot moving averages.
- CUSUM charts, or cumulative sum charts, are an alternative to Shewhart charts for detecting small process shifts.
- Presummarize and Levey-Jennings provide additional options for construction of variables control charts.
- Multivariate charts are for monitoring more than one process characteristic on the same chart.

## Control Chart Basics

Control charts are used to monitor a process. They tell us where the process is centered and how much variation we can expect (assuming nothing “special” happens). They also signal when something in the process has changed.

Control limits show the range of variation that we can expect to see, assuming that the process is stable.

- A stable process has only common causes of variation. Common cause variation is generally small scale variation that is inherent to the process.
- An unstable process also has special causes of variation. Special cause variation is external to the process and is often large in scale. For example, there might be a shift in the mean, spikes, cycles, or trends.

Control limits are generally plotted at  $\pm 3$  standard errors of the plotted statistic. Why  $\pm 3$  standard errors? For a normal distribution, approximately 99.73% of all values fall within this interval if the process is stable. That is, a point only falls outside this interval, just by chance, approximately 0.27% of the time. If a point falls outside the limits, we have a pretty good idea that it is a “special” event.

To make control charts more sensitive to special causes of variation, a number of tests have been developed. In JMP, the available tests can be requested for each control chart.

Use the **Tests** command to request the *Western Electric Rules*. These 8 tests are based on points falling in zones positioned at one, two, or three standard errors from the center line. **Note:** These tests will not run when the subgroup sizes are not constant.

An alternative to the Western Electric Rules is **Westgard Rules**. These tests are based on standard deviations rather than zones, so they can be computed regardless of the subgroup size.

The tests and the rules that apply to them are described in detail in the *Quality and Process Methods* book. Select **Help > JMP Help** to find the book.

## Control Charts for Variables Data

Control charts for variables data are classified according to the subgroup summary statistic plotted on the chart.

The XBar selection produces an XBar (averages) chart with an option to produce one of two other charts, an R or and an S chart:

- **XBar** charts display subgroup means (averages).
- **R** charts display subgroup ranges (maximum–minimum).
- **S** charts display subgroup standard deviations.

The estimate of the standard error used to compute the control limits on the XBar chart is derived from the within subgroup variation from the R or S chart.

The **IR** selection produces the following chart types:

- **Individuals** (or **Individual Measurement**) charts display individual observations.

- **Moving Range** charts display moving ranges of two or more successive measurements. The default range span is 2, but this can be changed.

The control limits on the Individuals chart are based on the moving ranges. Because moving ranges are correlated, these charts should be interpreted with care.

Variables control charts can be produced using options in the Control Chart menu shown in **Figure 18.2**, or Control Chart Builder, an interactive way to quickly and easily generate charts.

Let's proceed with constructing variables control charts with Control Chart Builder.

## Variables Charts Using Control Chart Builder

Control Chart Builder provides a workspace where you can drag process variables and subgroup sample variables to investigate the stability of a process. Several basic types of variables control charts are available: XBar, Range, Standard Deviation, Individual Measurement, and Moving Range.

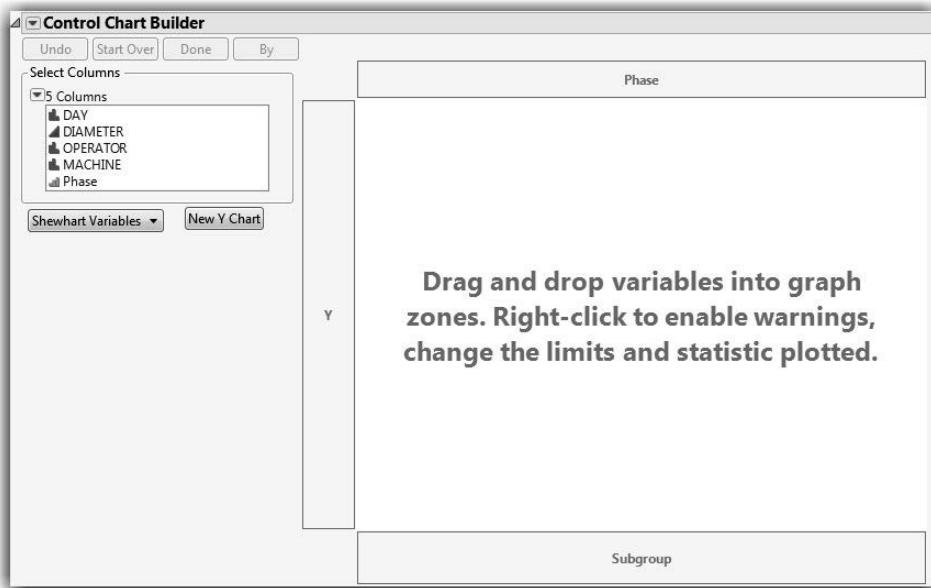
### The Control Chart Builder Work Space

Start Control Chart Builder from an open data table by selecting **Analyze > Quality and Process > Control Chart Builder**.

You first see a blank workspace like the one shown in **Figure 18.3**. The workspace has a central area to display the charts, and *drop zones* labeled Y, Phase, and Subgroup. The variable list is on the left, in the Control Chart Builder control panel. You create charts by dragging variables from the variable list to one (or more) of the drop zones. You can also drag a variable in the center of the workspace.

When you drop variables in the workspace, instant feedback encourages further exploration of the data. You can change your mind and quickly create another type of chart, or change the current settings by right-clicking on the existing chart.

The best way to start is to just jump in and try dragging variables to drop areas. However, the next section shows how to construct specific types of control charts with the Control Chart Builder.

**Figure 18.3** Control Chart Builder Workspace

## Control Chart Builder Examples

To create a control chart:

- ✓ Select **Help > Sample Data Library** and open Diameter.jmp.

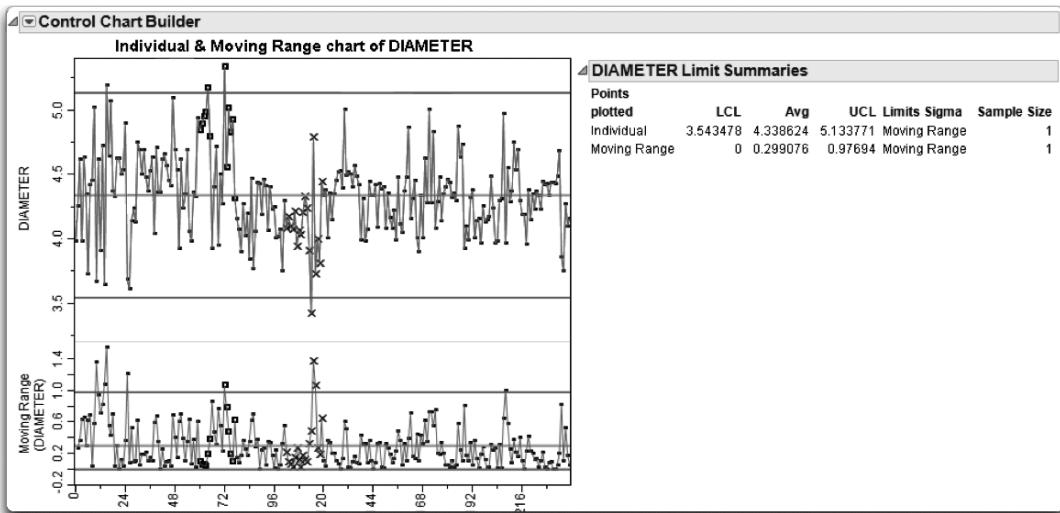
The measure of interest is the diameter of 4.4mm tubing used in medical applications. Samples of six tubes are measured each day over a 40-day period.

- ✓ Select **Analyze > Quality and Process > Control Chart Builder**.

- ✓ Drag DIAMETER to the **Y** drop zone.

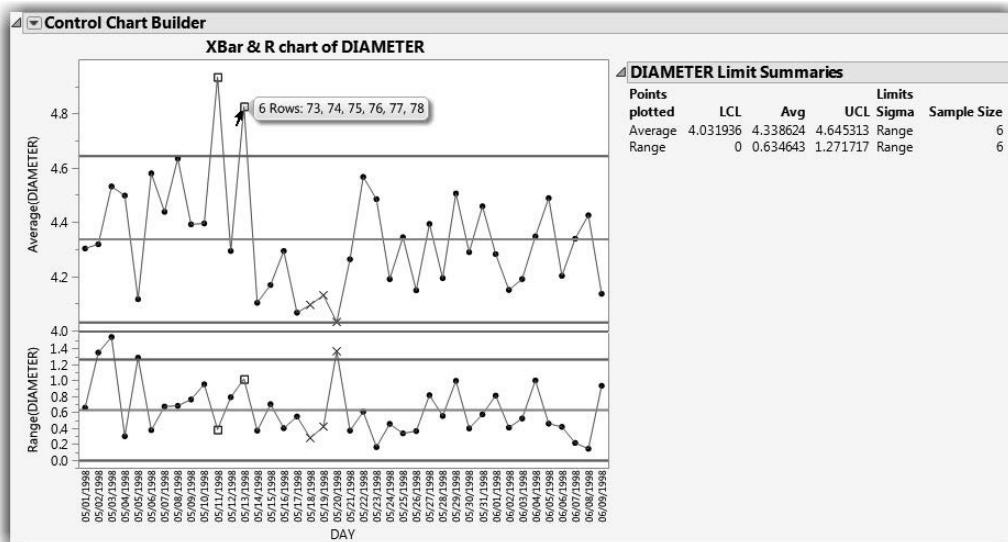
As simple as that, you can see the Individuals and Moving Range (IR) charts shown in **Figure 18.4**.

**Note:** In **Figure 18.4** through **Figure 18.6**, we clicked the **Done** button to close the control panel. To reopen the control panel, select **Show Control Panel** from the red triangle menu next to Control Chart Builder.

**Figure 18.4** Individuals and Moving Range Chart of DIAMETER

To create an XBar and R chart, as shown in **Figure 18.5**:

- ❖ Drag DAY to the middle of the graph.

**Figure 18.5** XBar and Range Chart of DIAMETER

If you move your mouse over a point (as shown), the corresponding row numbers are displayed. Click on a point to highlight it and select the corresponding rows in the data table. Each plotted point in the Averages (top) chart is the average of diameter values from a sample size of 6 rows.

Right-click anywhere in the chart to see a menu that lets you change or modify the control chart. For example, to request tests for special causes, right-click and select **Warnings > Tests** and select either individual tests or run all tests.

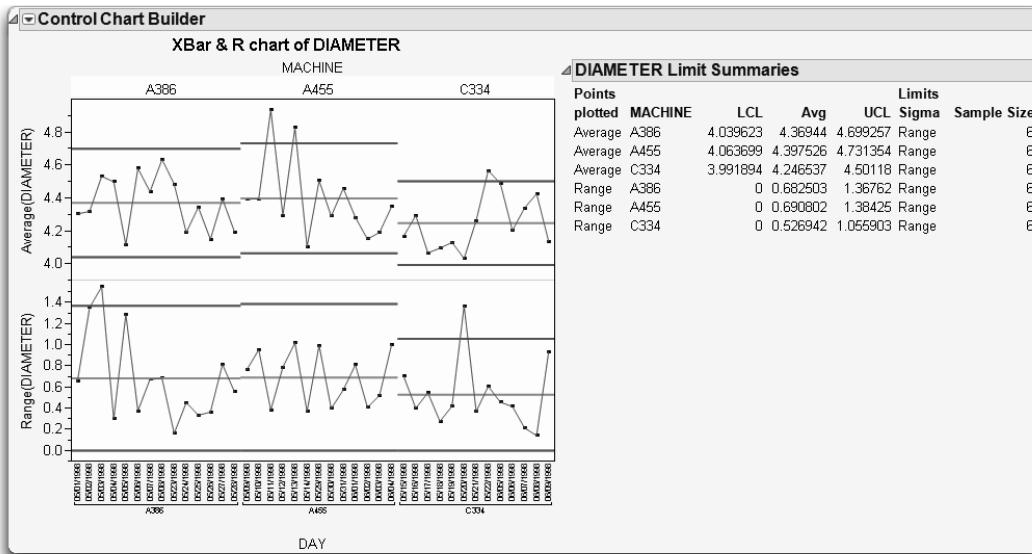
**Important:** The chart is completely interactive. Here are some examples:

- If you make mistakes, click the **Undo** button on the control panel. You can click **Undo** multiple times.
- If you want to create a new control chart from scratch, click **Start Over**.
- If you don't want to see the range chart, simply drag the title away from the chart, and it is gone.
- You can drag other variables to any of the drop zones in the chart.

This last bullet is important because that functionality is available only using the Control Chart Builder. For example, the Diameter.jmp sample data table contains information for three different machines.

To generate control limits for each machine:

☞ Drag Machine to the Phase drop zone. **Figure 18.6** shows the resulting graph.

**Figure 18.6** Phased Diameter Control Chart

**Note:** To produce variables control charts for several process variables at a time, use the Process Screening platform in the **Analyze > Screening** menu. This platform produces numerous metrics to assess stability.

## Control Charts for Attributes Data

Attributes charts, like variables charts, are classified according to the subgroup sample statistic plotted on the chart:

- P-charts display the proportion of nonconforming (defective) items in a subgroup sample.
- NP-charts display the number of nonconforming (defective) items in a subgroup sample.
- C-charts display the number of nonconformities (defects) in a subgroup sample that consists of a constant number of inspection units.
- U-charts display the average number of nonconformities (defects) per unit in a subgroup sample with an arbitrary number of inspection units.

**Note:** You can generate control charts for attributes data in Control Chart Builder. However, for the sake of illustration, we use the options in the Control Chart menu in this section.

### P- and NP-Charts

The Washers.jmp sample data table contains the number of defective washers in 15 lots of galvanized washers. The washers were inspected for finish defects such as rough galvanization and exposed steel. A defective washer had one or more finish defects.

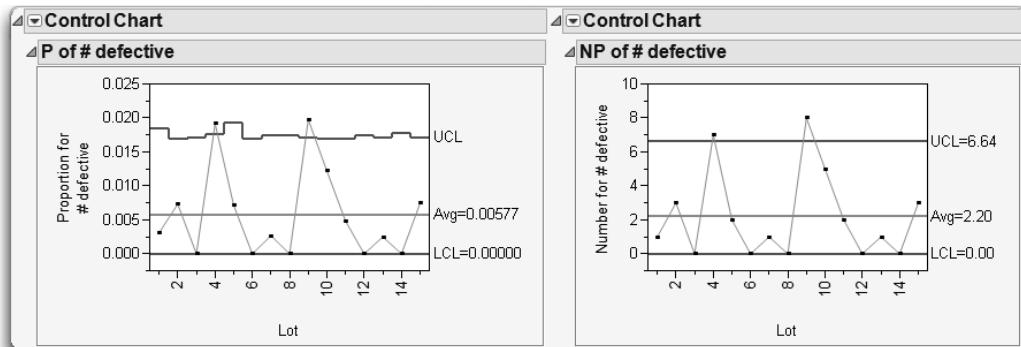
Let's say that we're interested in monitoring the proportion of defective washers in a lot, and that the lot size can vary. The chart to the left in **Figure 18.7** illustrates a *P*-chart for the proportion of defective washers per lot. To generate this *P*-chart:

- ~ With Washers.jmp active, select **Analyze > Quality and Process > Control Chart > P**.
- ~ Use # defective as the **Process** variable, Lot as the **Sample Label**, and Lot Size 2 as the **Sample Size**.

Like the other charts that we've discussed thus far, control limits are placed at  $\pm 3$  standard errors (of the proportion, in this case). Control limits vary according to the lot (or sample size). Larger lot sizes result in a smaller standard error, and therefore, tighter control limits. The lower control limit is bounded by zero.

Assume, instead, that the number of washers in a lot is constant, and that we're interested in monitoring the number of defective washers per lot. To generate the NP-chart in **Figure 18.7**:

- ~ Select **Help > Sample Data Library** and open Washers.jmp.
- ~ Select **Analyze > Quality and Process > Control Chart > NP**.
- ~ Use # defective as the **Process** variable, and Lot as the **Sample Label**.
- ~ The lot size is not actually used, but a value is required. Enter Lot Size as the **Sample Size**, or enter 400 (or any value, for that matter) under **Constant Size**.

**Figure 18.7** P and NP Charts for the Washers Sample Data

### C- and U-Charts

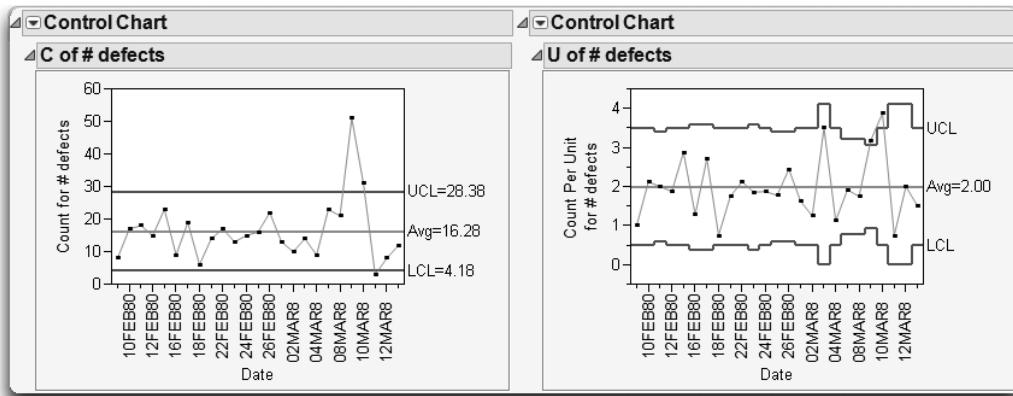
The Braces.jmp sample data records the defect count in boxes of automobile support braces. A box of braces is one inspection unit. The number of brace defects found in a day is the process variable.

Lets assume that a constant number of units (or boxes) are inspected per day. The C-chart on the left in **Figure 18.8** shows the number of defects found per day. To generate this chart:

- ☞ Select **Help > Sample Data Library** and open Braces.jmp.
- ☞ Select **Analyze > Quality and Process > Control Chart > C**.
- ☞ Use # defects as the **Process** variable and Date as the **Sample Label**.

The U-chart (also known as a DPU chart) shown to the right in **Figure 18.8** is monitoring the number of brace defects per box. The subgroup sample size, the number of boxes inspected in a day, is not constant. As we saw with the P-chart, the upper and lower control limits vary according to the sample size (in this case, the number of boxes inspected in a day).

- ☞ With Braces.jmp active, select **Analyze > Quality and Process > Control Chart > U**.
- ☞ Use # defects as the **Process** variable, Date as the **Sample Label** and Unit Size as the **Unit Size**.

**Figure 18.8** C and U Charts for the Braces Sample Data

## Specialty Charts

This section discusses the other types of control charts available from the Control Chart menu. The previous control charts plot each point based on information from a single subgroup sample.

More advanced charts such as CUSUM charts and Multivariate Charts are also available, but are not discussed here. For more information, select **Help > JMP Help** and refer to the *Quality and Process Methods* book.

### Presummarize Charts

Presummarize charts summarize the process column before charting. If you select

**Presummarize** from the **Control Chart** menu, the launch window includes the additional set of checkbox options shown here for plotting variables data. These options give a wider range of control chart types and options within each type.

<input checked="" type="checkbox"/> Individual on Group Means
<input type="checkbox"/> Individual on Group Std Devs
<input checked="" type="checkbox"/> Moving Range on Group Means
<input type="checkbox"/> Moving Range on Group Std Devs
<input type="checkbox"/> Median Moving Range on Group Means
<input type="checkbox"/> Median Moving Range on Group Std Devs
Range Span <input type="text" value="2"/>

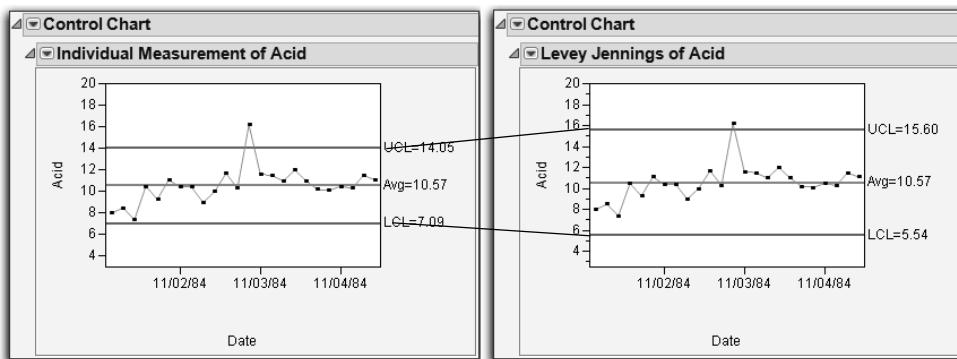
**Note:** Many of these options are available in the Control Chart Builder.

## Levey-Jennings Charts

Levey-Jennings charts show a process mean with control limits based on a long-term sigma, placed at 3 standard deviations from the center line. The standard deviation for the Levey-Jennings chart is the overall standard deviation of the process variable. **Figure 18.9** shows a comparison of the same data plotted on an IR chart and on a Levey-Jennings chart.

The overall standard deviation is usually larger than the estimate of the standard deviation based on subgroups, resulting in wider control limits.

**Figure 18.9** Compare Limits between IR and Levey-Jennings Control Charts



In the previous control charts, each point plotted is based on information from a single subgroup sample. Moving average charts are different because each point combines information from the current and past samples. As a result, moving average charts are more sensitive to small shifts in the process average. Unfortunately, it is more difficult to interpret patterns of points on a moving average chart because consecutive moving averages can be highly correlated (Nelson 1984).

## Uniformly Weighted Moving Average (UWMA) Charts

Each point on a Uniformly Weighted Moving Average (UWMA) chart is the average of the  $w$  most recent subgroup means, including the present subgroup mean. When you obtain a new subgroup sample, the next moving average is computed by dropping the oldest of the previous  $w$  subgroup means and including the newest subgroup mean. The constant  $w$  is called the *span* of the moving average. There is an inverse relationship between  $w$  and the magnitude of the shift that can be detected. Thus, larger values of  $w$  allow the detection of smaller shifts.

To see an example, follow these steps:

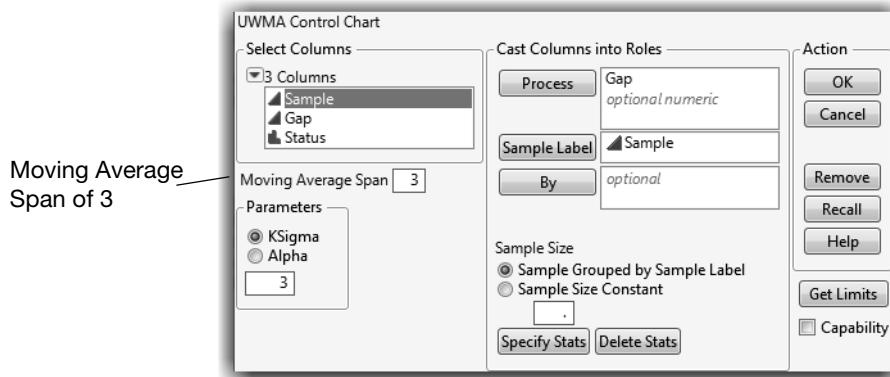
- ☞ Select **Help > Sample Data Library** and open Clips1.jmp.

The measure of interest is the gap between the ends of manufactured metal clips. To monitor the process for a change in average gap, subgroup samples of five clips are selected daily, and a UWMA chart with a moving average span of three samples is examined. To see the UWMA chart, complete the Control Chart window.

- ☞ Select **Analyze > Quality and Process > Control Chart > UWMA**.
- ☞ Use Gap as the **Process** variable and Sample as the **Sample Label**.
- ☞ Enter 3 as the Moving Average Span.

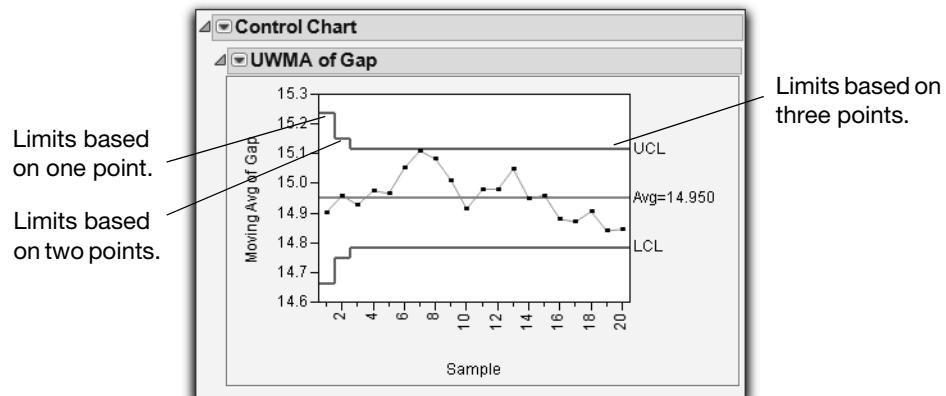
The Control Chart launch window should look like the one shown in **Figure 18.10**.

**Figure 18.10** Specification for UWMA Charts of Clips1 Data



- ☞ Click **OK** to see the chart shown in **Figure 18.11**.

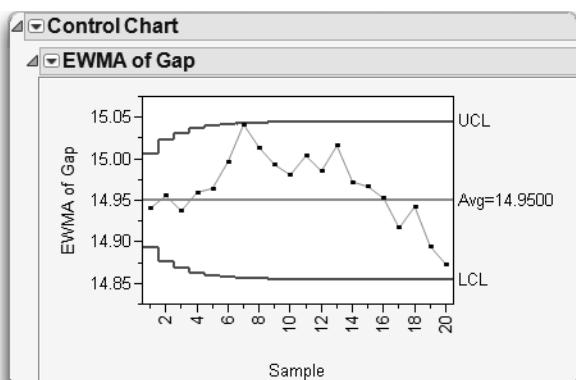
The point for the first day is the mean of the first subsample only, which consists of the five sample values taken on the first day. The plotted point for the second day is the average of subsample means for the first and second day. The points for the remaining days are the average of subsample means for each day and the two previous days.

**Figure 18.11** UWMA Charts for the Clips1 Data

## Exponentially Weighted Moving Average (EWMA) Chart

Each point on an Exponentially Weighted Moving Average (EWMA) chart is the weighted average of all the previous subgroup means, including the mean of the present subgroup sample. The EWMA chart is also referred to as a Geometric Moving Average (GMA) chart.

The weights decrease exponentially going backward in time. The weight ( $0 < r < 1$ ) assigned to the present subgroup sample mean is a parameter of the EWMA chart. Small values of  $r$  are used to guard against small process shifts. If  $r = 1$ , the EWMA chart reduces to a Mean control (Shewhart) chart, previously discussed.



The default value of  $r$  is 0.2, which makes the EWMA chart sensitive to relatively small shifts in the process mean. The figure shown here is an EWMA chart for the same data used for **Figure 18.11** (Clips1.jmp).

**Note:** To explore the values plotted on the UWMA and EWMA charts, you can create a new series in the data table with uniformly or exponentially weighted averages and graph these values in Graph Builder. An easy way to calculate these weighted values is to right-click the column header for the process variable in the data table and select **New Formula Column > Row > Moving Average**.

## Capability Analysis

As we have seen, control charts are used to assess and monitor the stability of a process. The terms “stable” and “in control” are used to describe a process with only common causes of variation. However, a stable process might not be capable of producing the desired quality relative to specifications or tolerances. In other words, a stable process is not necessarily a capable process.

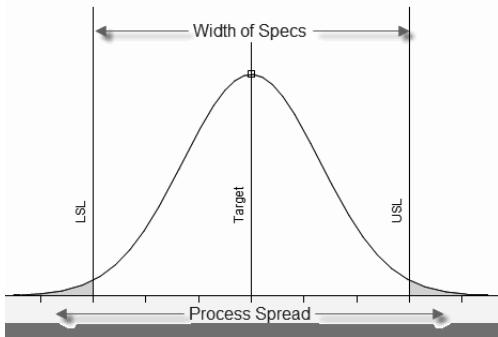
A capability analysis can be performed in three JMP platforms. The platform that you use depends on the number of variables under study and whether the data are time ordered, described below.

- Use the Distribution platform for one continuous variable.
- Use the Process Capability platform for several continuous variables.
- Use control charts for time-ordered variables.

### What Is Process Capability?

Capability is a useful measure of how well a process keeps within the specification limits. For normally distributed measurements, capability is purely a function of the mean and variance of the process. You can have a process go wrong because its mean is off target, or because there is too much variation, or both.

Process capability studies compare the variability (or spread) of a stable process to the width of the specification limits, as shown in **Figure 18.12**.

**Figure 18.12** Calculating Capability Indices

A variety of indices are available to quantify capability. The most common measures are  $C_p$ ,  $C_{pk}$ ,  $C_{pl}$ , and  $C_{pu}$ .

The index that you use depends on the situation:

- If it is important that you hit a target, then  $C_{pk}$  is the best choice.
- If you're concerned about staying within a one-sided lower or upper specification limit, then  $C_{pl}$  (lower) or  $C_{pu}$  (upper) should be used.
- If you need to make sure your variation is under control, given it is relatively easy to steer the mean to the target,  $C_p$  is the best measure.

Formulas for computing these indices are shown below.

$$C_p = (USL - LSL)/(6s)$$

$$C_{pl} = (\bar{x} - LSL)/(3s)$$

$$C_{pu} = (USL - \bar{x})/(3s)$$

$$C_{pk} = \min(C_{pl}, C_{pu})$$

The  $C_{pk}$  measure is the minimum of  $C_{pl}$  and  $C_{pu}$ . As a result, if the process is off target,  $C_{pk}$  is lower than  $C_p$ —a penalty for being off target.

Lets take a closer look at these formulas. The LSL and USL are the lower and upper specification limits,  $s$  is the standard deviation of the process and  $\bar{x}$  is the mean.

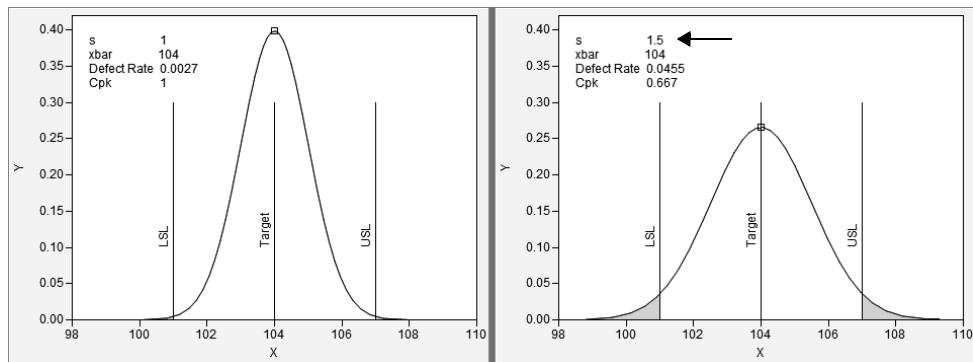
Why is  $3s$  built into the formulas for  $C_{pl}$  and  $C_{pu}$  (and, thus,  $C_{pk}$ )? For a normal distribution, it turns out that at 3 standard deviations, we have a tail probability of 0.00134. Let's say that our process is centered at the target, and that the distance

between the mean and each spec limit is 3 standard deviations.  $C_{PL}$  and  $C_{PU}$  are both 1.0.  $C_{PK}$  is then 1.0, and the probability that an observation falls either below the lower spec or above the upper spec is 0.0027. This is an acceptable rate of defects for some processes, and the process would be considered capable.

This scenario is illustrated to the left in **Figure 18.13**. The LSL = 101, the target = 104, the USL = 107, and the process is on target. The  $C_{PK}$  is 1.0, and the overall defect rate is 0.27%.

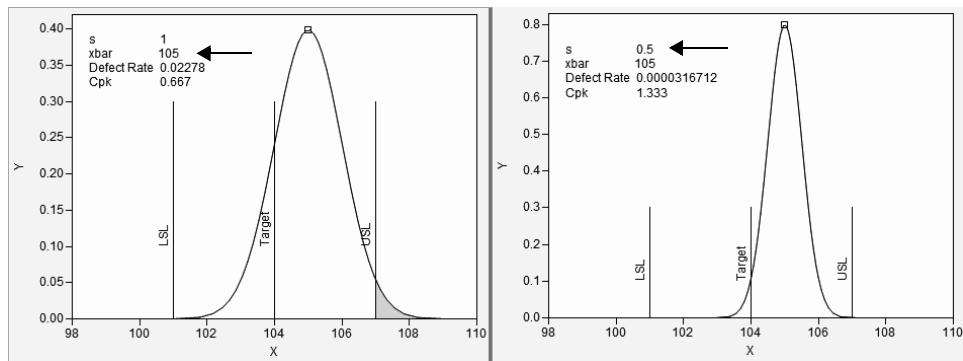
With the same process mean, if the standard deviation increased from 1.0 to 1.5, then the  $C_{PK}$  would drop to 0.667, producing 4.55% defects (on the right in **Figure 18.13**).

**Figure 18.13** Capability Examples



If the standard deviation remained at 1.0, but the process mean shifted off target to 105, then the process is also incapable, producing defects at around 2.2% as shown to the left in **Figure 18.14**.

However, if we cut the variability in half (from 1.0 to 0.5), even if the process remains off target, our capability improves and there are only rare defects as shown to the right.

**Figure 18.14** More Capability Examples

**Note:** A number of add-ins for exploring capability indices are available on the JMP Community at <http://community.jmp.com>. Search for “capability animation” or “capability add-in”.

## Capability for One Process Measurement

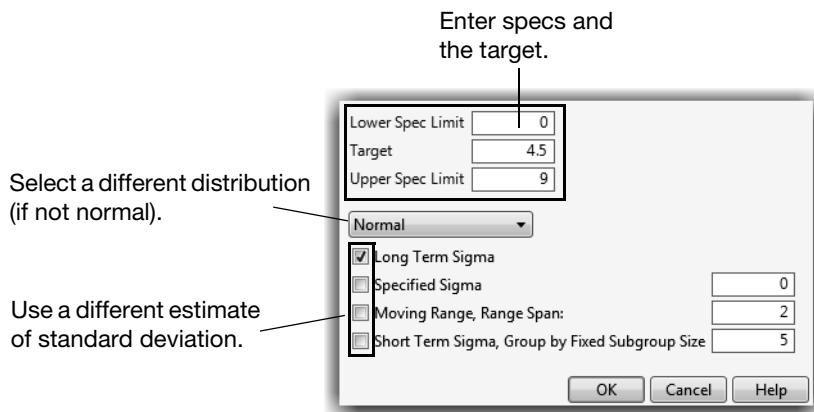
The Distribution platform provides a **Capability Analysis** option on the red triangle menu of each continuous variable plotted. Here is an example:

- ❖ Select **Help > Sample Data Library** and open Cities.jmp.

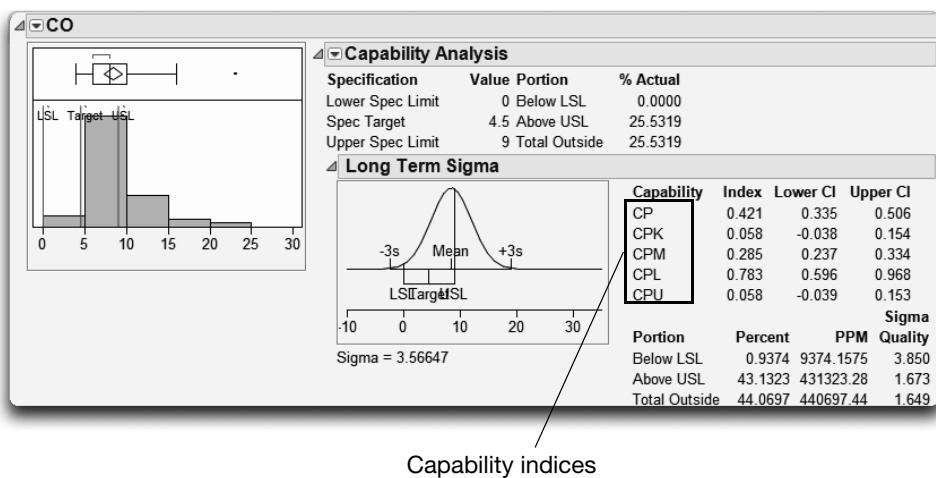
This file contains data on pollutants for 52 cities. The variable of interest is CO (Carbon Monoxide). In cities, high levels of carbon monoxide are primarily caused by vehicle exhaust. According to the United States EPA (<http://airnow.gov>), the “Good” (desired) range of CO is 0-4.5 ppm, while “Moderate” (acceptable) levels are between 4.5 and 9 ppm.

- ❖ Select **Analyze > Distribution**, assign CO to **Y, Columns**, and then click **OK**.
- ❖ Select **Capability Analysis** from the red triangle menu next to CO, and populate the capability analysis window as shown in **Figure 18.15**.
- ❖ Click **OK**.

Additional options are noted in **Figure 18.15**.

**Figure 18.15** Distribution Capability Specification Window

The specification limits and target appear on the histogram, and a Capability Analysis report is appended to the Distribution output, as shown in **Figure 18.16**. The Quantiles and Summary Statistics have been deselected.

**Figure 18.16** Distribution Capability Analysis for CO (Carbon Monoxide)

The  $C_{PK}$  of 0.058 indicates that the carbon monoxide levels in the cities are well above acceptable levels.

**Note:** Process capability can be calculated using long-term estimates of sigma (the overall standard deviation of the data) or short-term estimates (calculated from within subgroup variation). When the long term estimate of sigma is used to

calculate capability, the labels  $P_p$  and  $P_{PK}$  are generally used instead of  $C_p$  and  $C_{PK}$ . For the Long Term Sigma capability indices, select **File > Preferences** (or **JMP > Preferences** on Macintosh) > **Platforms > Distribution** and select  **$P_{PK}$  Capability Labeling**. This labeling is then used in future analyses (or if you redo the analysis).

## Capability for Many Process Measurements

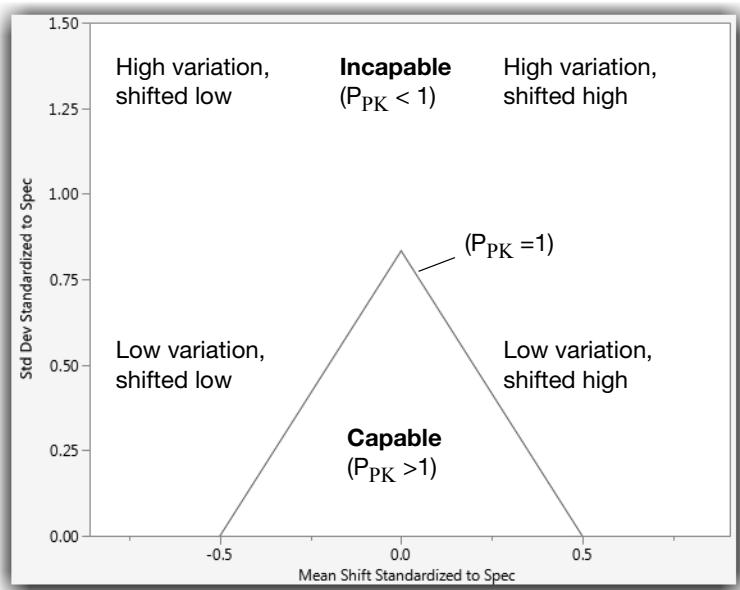
Assessing capability for one process, as we have seen, is easy. But, what if we have hundreds of processes, and we're only interested in finding the problem cases? To assess the capability of many processes, no one wants to look at a complex report and graph for each process variable. How do we show capability in just one picture, especially when the processes have different targets and different specification limits?

The solution is to normalize the mean and the standard deviation for each process relative to the specification range. These normalized values can then be plotted on the same graph, called a goal plot (**Figure 18.17**).

The goal plot displays a triangle corresponding to a  $P_{PK} = 1.0$ . Measures with a  $P_{PK} > 1.0$  are plotted inside the triangle, and measures with a  $P_{PK} < 1.0$  are plotted outside the triangle. Where they are plotted depends on the shift from the target and the variation.

- A perfect process is at the origin (0, 0), hitting the target and having no variation.
- An off-target process or one with too much variation might be acceptable, as long as it stays within the capability triangle.
- If the process has too much variation but is on-target, it likely falls outside the triangle.
- If the process has low variation but the mean falls outside the specification limits, the process is producing all defects.

To get a higher  $P_{PK}$ , you need to move the mean closer to the target, reduce the variation, or both.

**Figure 18.17** Interpretation of the Goal Plot

**Note:** The goal plot in JMP displays  $P_{PK}$  values by default. To display  $C_{PK}$  values instead, select **File > Preferences** (or **JMP > Preferences** on Macintosh) > **Platforms** > **Process Capability** and deselect **AIAG (Ppk) Labeling**.

For example, the sample data table Semiconductor Capability.jmp contains measurements on 128 quality characteristics. Each of these characteristics has different specification limits, which have been added to the column properties for each process variable in the data table. To view this information:

- ☞ Select **Help > Sample Data Library** and open Semiconductor Capability.jmp.

The process variables are grouped in a Processes group, which you see in the Columns panel.

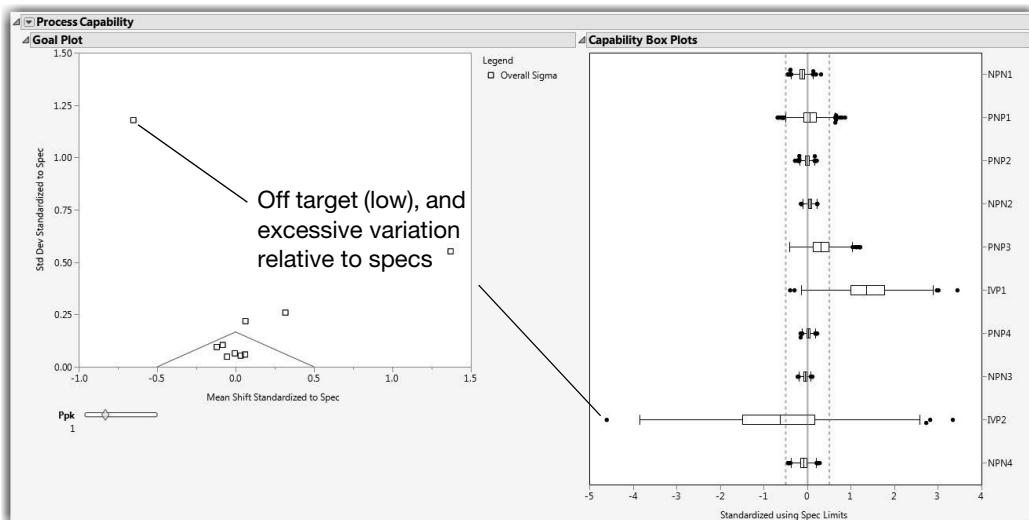
- ☞ To view the process variables, click the gray icon next to Processes in the Columns panel.
- ☞ To view the specification limits, right-click on any process variable in the Columns panel, and select **Column Info**.

Let's say that, for simplicity, we are interested in seeing a capability analysis for the first 10 quality characteristics, NPN1 through NPN4.

- ✓ Select **Analyze > Quality and Process > Process Capability**.
- ✓ Select the 10 columns as **Y, Process**, and then click **OK** to see the results in **Figure 18.18**.

By default, the red triangle in the goal plot corresponds to an acceptable  $P_{PK}$  of 1.0. However, this can be changed using the slider under the goal post.

**Figure 18.18** Capability Analysis Platform Results



The Capability Box Plots panel displays box plots for each variable standardized by individual specification limits. The standardized target is displayed as a solid green line, and the spec limits are displayed as green dashed lines. Box plots for off target characteristics are shifted away from the center line, and box plots for characteristics with too much variability are wider than the dashed spec limits.

Capability indices for each characteristic, and other options, can be requested from the top red triangle menu.

**Note:** If specification limits are not entered as column properties, as they were in this example, JMP opens a Spec Limit window. Specification limits can either be imported from a table or entered manually.

## Capability for Time-Ordered Data

When generating variables control charts from the Control Chart platform for time-ordered data, a capability analysis can be requested in addition to the control chart. To see an example:

- ✓ Select **Help > Sample Data Library** and open Clips1.jmp.

Recall that the measure of interest is the gap between the ends of manufactured metal clips. Samples of five clips are selected daily, and the gap is measured. The target and specifications for the process are  $15 \pm 0.5$  mm.

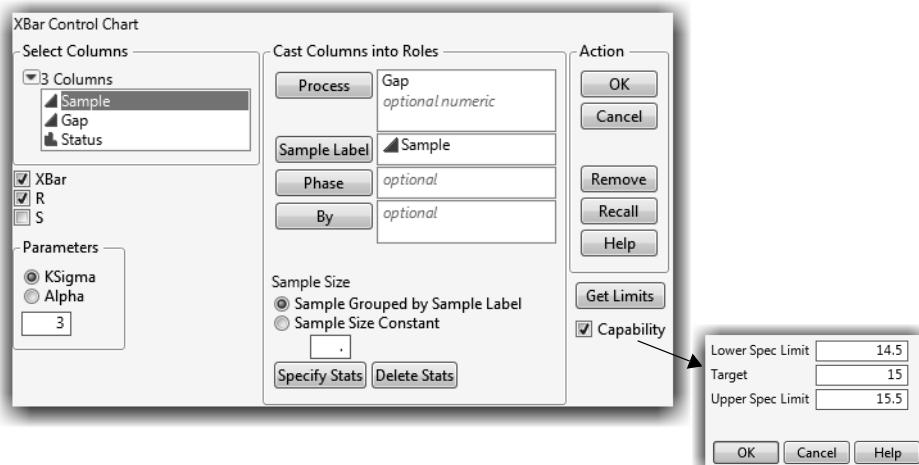
To generate an XBar and R chart, and request the capability analysis from the control chart launch window:

- ✓ Select **Analyze > Quality and Process > Control Chart > XBar**.
- ✓ Enter Gap as the **Process**, and Sample as the **Sample Label**.
- ✓ Select the **Capability** box, and then click **OK**.

The completed window is shown in **Figure 18.19**.

- ✓ In the Specification window, enter the specs and target as shown to the right in **Figure 18.19**, and click **OK**.

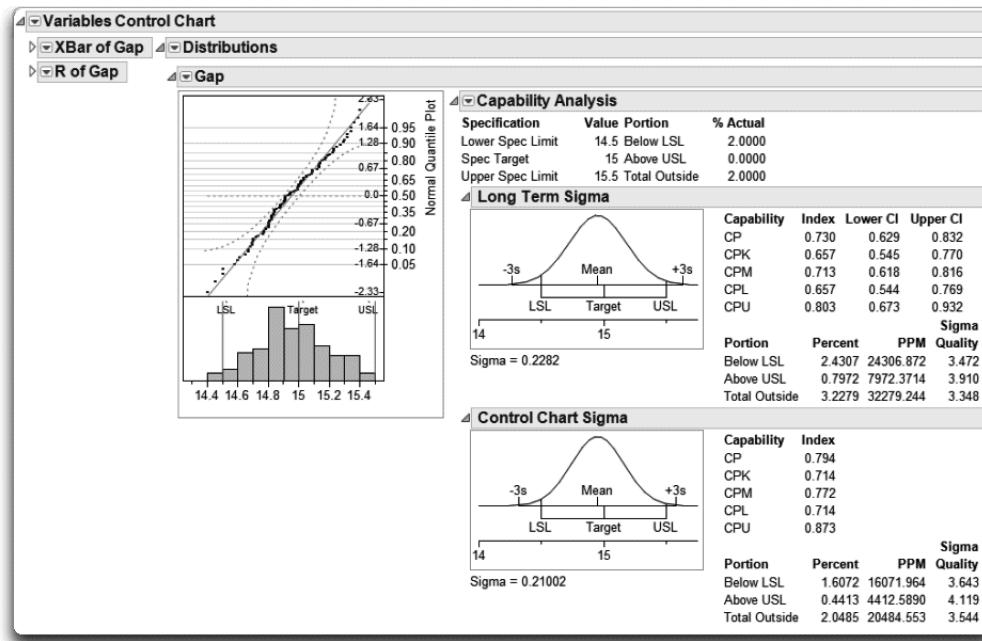
**Figure 18.19** Control Chart Launch Window and Capability Analysis



The resulting capability analysis is shown **Figure 18.20**.

**Note:** The XBar and R control charts are not displayed. Before conducting a capability analysis, the stability of the process should be verified. If a process is not stable, its future performance is unpredictable. So, by definition, an *unstable process is not capable*.

**Figure 18.20** Control Chart Capability Analysis Results



The output is identical to what we saw from the Distribution platform, with the addition of a normal quantile plot and a second set of capability results. The normal quantile plot should be used to assess the normality of the underlying distribution. The Long Term Sigma results are based on the overall (long term) estimate of the standard deviation. The Control Chart Sigma (short term) results are based on the within-subgroup estimate of the standard deviation computed from the Range chart.

### Additional Notes

- The capability analysis can also be requested from the red triangle menu next to the Variables control chart.
- To perform a capability analysis in Control Chart Builder, create a variables control chart, right-click on the top chart, select **Limits > Add Spec Limits**.

- $C_{pk}$  labeling is generally used when the short-term (within subgroup) estimate of sigma is used to estimate process capability. However, when the overall estimate of sigma is used to estimate the long-term capability,  $P_{pk}$  labeling is generally used. Long-term capability is also referred to as process performance.
- The examples in this section assume that the process characteristics follow a normal distribution. For nonnormal distributions, capability indices can (and should) be calculated instead. For information on nonnormal capability indices, search for “nonnormal capability” in the JMP Help. You can also select **Help > JMP Help** and refer to the *Quality and Process Methods* book.

## A Few Words about Measurement Systems

We generally assume that our measurements are representative of the true value of the characteristic being measured. However, this is often not the case. A measurement system analysis (MSA) is a study used to verify the integrity and quality of our measurement process.

JMP provides two platforms for analyzing measurement systems. Both are found under **Analyze > Quality and Process**.

- **Measurement Systems Analysis** is the EMP (Evaluating the Measurement Process) approach developed by Donald Wheeler (2006). A Gauge R&R analysis can also be performed from this platform.
- **Variability / Attribute Gauge Chart** performs a Variance Components Analysis, Gauge R&R, or Attribute measurement system analysis.

For details about using these methods, select **Help > JMP Help** and refer to the *Quality and Process Methods* book.

## Exercises

1. The sample data table Oil2 Cusum.jmp contains the distribution of weight measurements for a can-filling process. Four cans are measured per hour over a 12-hour period. We are interested in assessing the stability and capability of the process. The specs for the process are  $8.1 \text{ oz} \pm 0.1 \text{ oz}$ .

- (a) Open the data table, and look at the data. Which of the two control chart types would be more appropriate for monitoring the filling process: an IR or an XBar R or S? Why?
  - (b) Use the Control Chart Builder to create the control chart. What is the mean of the process? Does the process appear stable?
  - (c) Now, run the tests for special causes (the Western Electric Rules). Right-click on the graph and select **Warnings > Tests**, and select **All Tests**. Is the process stable?
2. Perform a capability study for the Oil2 Cusum.jmp sample data using the specs given in question 1. Use either Control Chart Builder or the specific chart from the Control Chart platform.
- (a) The assumptions to perform a capability study are that the process is stable and that the underlying distribution is normal. Are these assumptions met?
  - (b) What percent of the measurements fell outside the specification limits?
  - (c) What percent of measurements are predicted to fall outside the specification limits over the long term?
  - (d) Is the process capable? Recall that a  $C_{PK}$  of 1.0 is generally considered capable?
  - (e) Engineers have decided that the process needs to be improved. Should they focus on (1) centering the mean on the target, (2) reducing process variation, or (3) both?
3. The sample data table Fabric.jmp contains information about the number of defects found in upholstery fabric used in automobiles. An incoming inspection process randomly selects and inspects one bolt (100 yards long and 72 inches wide) per shipment and records the number of defects.
- (a) Based on the description above, what type of control chart is appropriate to monitor defects per bolt? P, NP, C, or U? Why?
  - (b) Open the file, and use the Control Chart platform to generate the control chart. Does the process appear to be stable?
  - (c) Run the tests for special causes. Note that for attribute control charts only the first four tests are available. Which tests signal that there are special causes?

- (d) Repeat steps b and c above using Control Chart Builder.
  - (e) Use the JMP Help to investigate the signals of the special causes found above.
4. Open Abrasion.jmp, and use the Control Chart Builder to create and XBar and R chart of Abrasion (Y), Date (Subgroup) and Shift (Phase).
- (a) Does there appear to be a difference in abrasion measurements between the two shifts?
  - (b) Use Graph Builder or Fit Y by X to further explore the potential difference between the two shifts.



# 19

## Mechanics of Statistics

### Overview

This chapter is an essay on fitting for those of you who are mechanically inclined. If you have any talent for imagining how springs and tire pumps work, you can put it to work here in a fantasy in which all the statistical methods are visualized in simple mechanical terms.

The goal is to not only remember how statistics works, but also train your intuition so that you are prepared for new statistical issues.

Here is an illuminating trick that helps you understand and remember how statistical fits really work. It involves pretending that statistical fitting is performed by machines. If we can figure out the right machines and visualize how they behave, we can reconstruct all of statistics by putting together these simple machines into arrangements appropriate to the situation. We need only two machines of fit, the spring for fitting continuous normal responses and the pressure cylinder for fitting categorical responses.

Readers interested in this approach should consult Farebrother (2002), who covers physical models of statistical concepts extensively.

## Chapter Contents

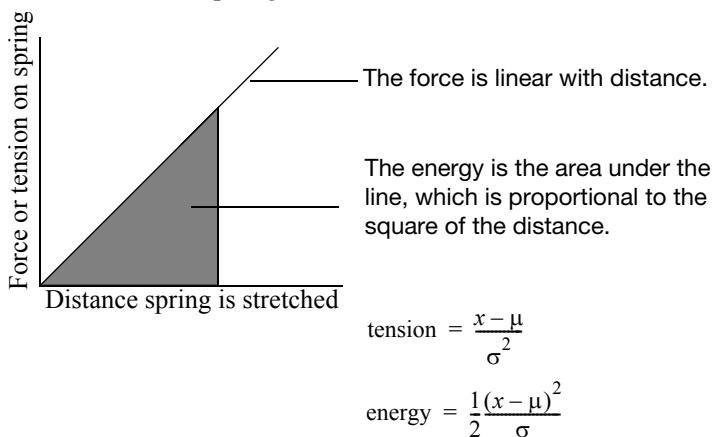
Overview .....	577
Springs for Continuous Responses .....	579
Fitting a Mean .....	579
Testing a Hypothesis.....	580
One-Way Layout .....	581
Effect of Sample Size Significance .....	581
Effect of Error Variance on Significance.....	582
Experimental Design's Effect on Significance.....	583
Simple Regression.....	584
Leverage .....	585
Multiple Regression .....	586
Summary: Significance and Power.....	586
Mechanics of Fit for Categorical Responses.....	586
How Do Pressure Cylinders Behave? .....	587
Estimating Probabilities .....	588
One-Way Layout for Categorical Data.....	589
Logistic Regression .....	591

## Springs for Continuous Responses

How does a spring behave? As you stretch the spring, the tension increases linearly with distance. The energy that you need to pull a spring a given distance is the integral of the force over the distance, which is proportional to the square of the distance.

Take  $1/\sigma^2$  as the measure of how stiff a spring is. Then the graph and equations for the spring are as shown in **Figure 19.1**.

**Figure 19.1** Behavior of Springs



In this way, springs help us visualize least squares fits. They also help us do maximum likelihood fits when the response has a normal distribution.

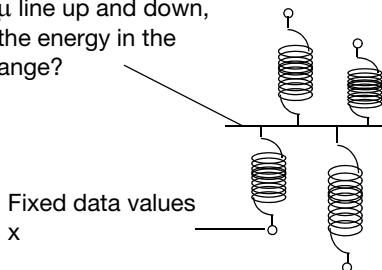
The formula for the log of the density of a normal distribution is identical to the formula for energy of a spring centered at the mean, with a spring constant equal to the reciprocal of the variance. A spring stores and yields energy in exactly the way that normal deviations get and give log-likelihood. So, maximum likelihood is equivalent to least squares, which is equivalent to minimizing energy in springs.

## Fitting a Mean

How do you fit a mean by least squares? Imagine stretching springs between the data points and the line of fit (see **Figure 19.2**). Then you move the line of fit around until the forces acting on it from the springs balance. That will be the point of minimum energy in the springs. For every minimization problem, there is an equivalent balancing (or orthogonality) problem, in which the forces (tensions, relative distances, residuals) add up to zero.

**Figure 19.2** Fitting a Mean by Springs

Movable line of fit. As you move this  $\mu$  line up and down, how does the energy in the springs change?



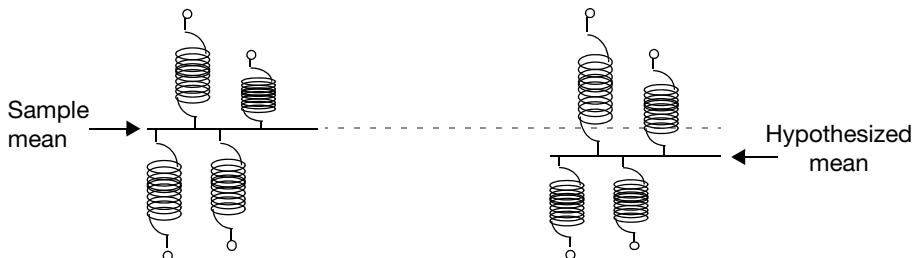
$$\text{tension} = \frac{x - \mu}{\sigma^2}$$

$$\text{energy} = \frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}$$

The least energy position is the least squares position, which is also the balancing position.

## Testing a Hypothesis

Suppose that you want to test a hypothesis that the mean is some value. You would force the line of fit to be that value and measure how much more energy you had to add to the springs (how much more the sum of squared residuals was) to constrain the line of fit. This is the sum of squares that is the main ingredient of the  $F$ -test. To test that the mean is (not) the same as a given value, find out how hard it is to move it there (see **Figure 19.3**).

**Figure 19.3** Compare a Mean to a Given Value

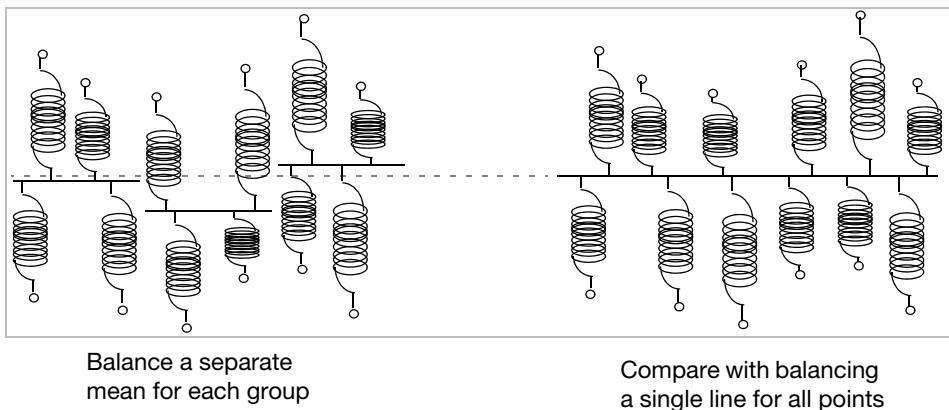
Here, the line of fit is the balance point of the springs; the energy in the springs is at a minimum when the system is balanced.

Here, the line of fit has been forced down to the hypothesized value for the mean, and it took a certain amount of energy to push it to this hypothesized value.

## One-Way Layout

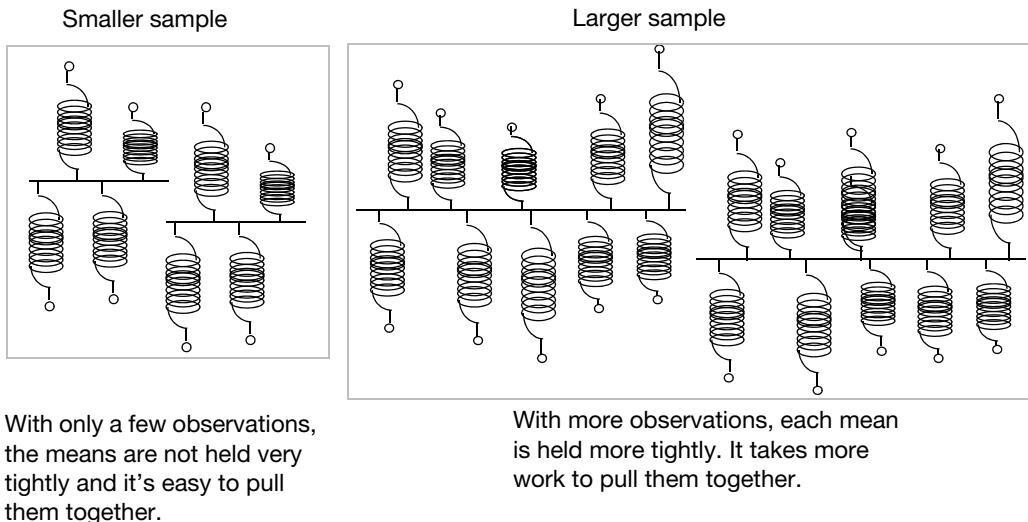
If you want to fit several means, you can do so by balancing a line of fit with springs for each group. To test that the means are the same, you force the lines of fit to be the same, so that they balance as a single line. You also measure how much energy you had to add to the springs to do this (how much greater the sum of squared residuals was). See **Figure 19.4**.

**Figure 19.4** Means and the One-Way Analysis of Variance



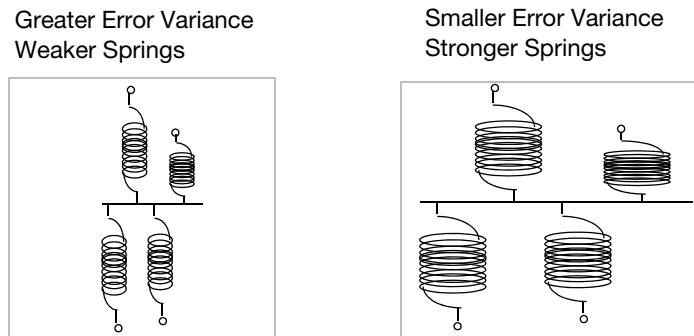
## Effect of Sample Size Significance

When you have a larger sample, there are more springs holding on to each mean estimate, and it is harder to pull them together. Larger samples lead to a greater energy expense (sum of squares) to test that the means are equal. The spring examples in **Figure 19.5** show how sample size affects the sensitivity of the hypothesis test.

**Figure 19.5** A Larger Sample Helps Make Hypothesis Tests More Sensitive

## Effect of Error Variance on Significance

The spring constant is the reciprocal of the variance. Thus, if the residual error variance is small, the spring constant is bigger, the springs are stronger, it takes more energy to bring the means together, and the test is therefore more significant. The springs in **Figure 19.6** illustrate the effect of variance size.

**Figure 19.6** Reduced Residual Error Variance Makes Hypothesis Tests More Sensitive

The spring constant is  $1/\sigma^2$ .  
So, greater error variance means weaker springs, less energy required to bring the means together, and nonsignificant tests.

Smaller error variance means stronger springs, more energy required to bring the means together, and significant tests.

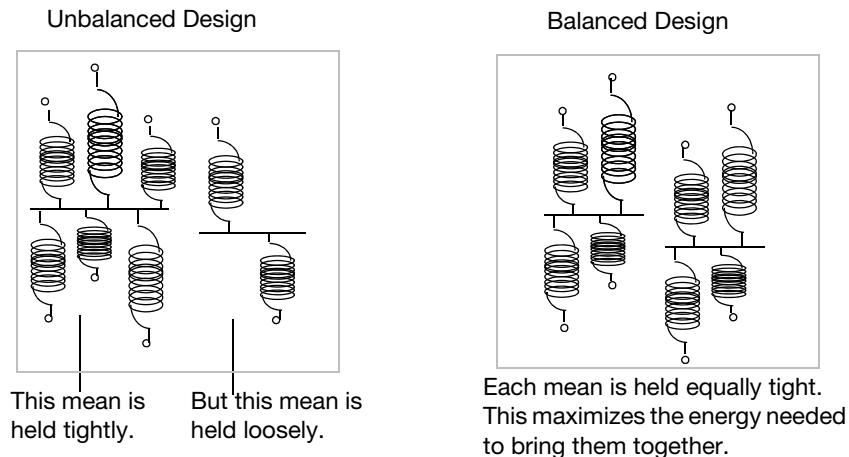
## Experimental Design's Effect on Significance

If you have two groups, how do you arrange the points between the two groups to maximize the sensitivity of the test that the means are equal? Suppose that you have two sets of points loading two lines of fit, as in the one-way layout shown previously in **Figure 19.4**. The test that the true means are equal is done by measuring how much energy it takes to force the two lines together.

Suppose that one line of fit is suspended by a lot more points than the other. The line of fit that is suspended by few points will be easily movable and can be stretched to the other mean without much energy expenditure. The lines of fit would be more strongly separated if you had more points on this loosely sprung side, even at the expense of having fewer points on the more tightly sprung side. It turns out that to maximize the sensitivity of the test for a given number of observations, it is best to allocate points in equal numbers between the two groups. In this way, both means are equally tight, and the effort to bring the two lines of fit together is maximized.

So the power of the test is maximized in a statistical sense by a balanced design, as illustrated in **Figure 19.7**.

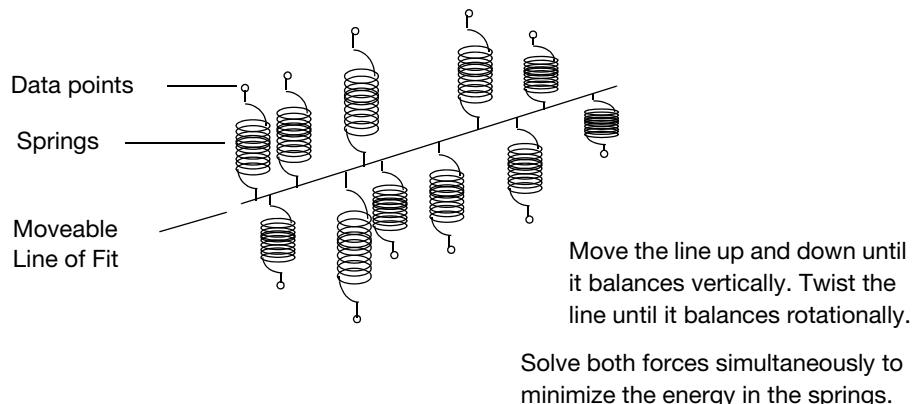
**Figure 19.7** Design of Experiments



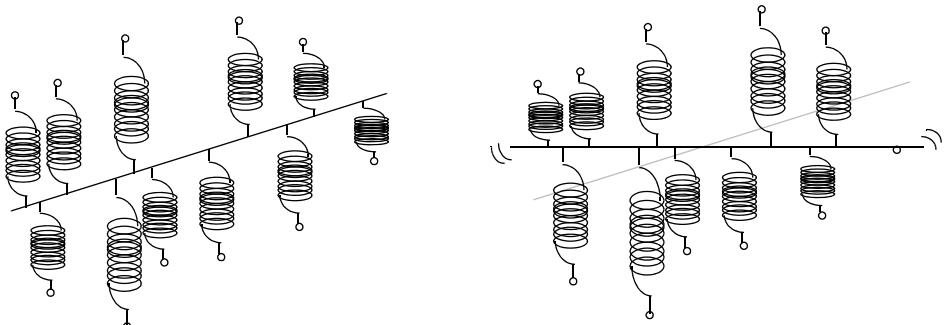
## Simple Regression

If you want to fit a regression line through a set of points, you fasten springs between the data points and the line of fit, such that the springs stay vertical. Then let the line be free so that the forces of the springs on the line balance, both vertically and rotationally (see **Figure 19.8**). This is the least squares regression fit.

**Figure 19.8** Fitting a Regression Line with Springs



If you want to test that the slope is zero, you force the line to be horizontal so that you're just fitting a mean and measure how much energy it took to constrain the line (the sum of squares due to regression) (see **Figure 19.9**).

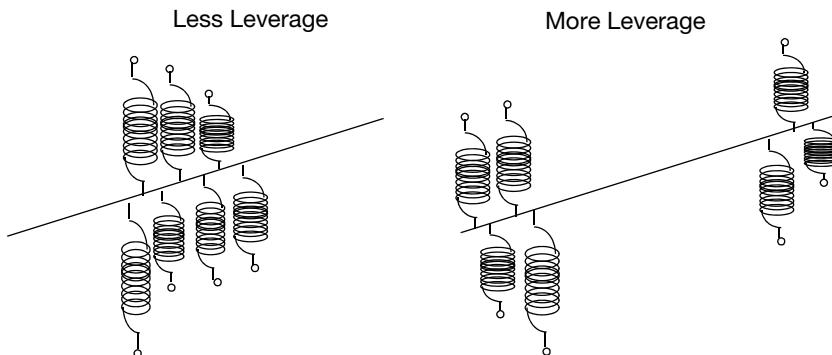
**Figure 19.9** Testing the Slope Parameter for the Regression Line

This line is where the forces governing the slope of the line balance. It is the minimum energy solution.

If you force the line to have a slope of zero, how much additional energy do you have to give to the springs? How much work is it to move the line to be horizontal?

## Leverage

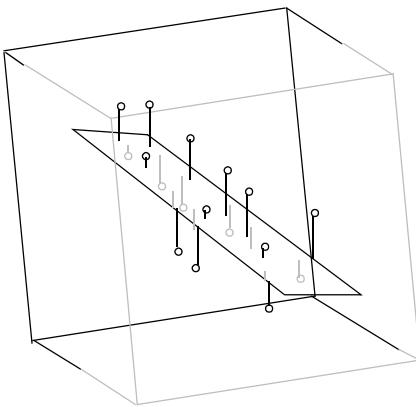
If most of the points that are suspending the line are near the middle, then the line can be rotated without much effort to change the slope within a given energy budget. If most of the points are near the end, the slope of the line of fit is pinned down with greatest resistance to force. That is the idea of leverage in a regression model. Imagine trying to force the line to have a different slope. Look at **Figure 19.10** and decide which line would be easier to twist.

**Figure 19.10** Leverage with Springs

## Multiple Regression

The same idea works for fitting a response to two regressors; the difference is that the springs are attached to a plane rather than a line. Estimation is done by adjusting the plane so that it balances in each way. Testing is done by constraining the plane.

**Figure 19.11** Three-Dimensional Plot of Two Regressors and Fitted Plane



### Summary: Significance and Power

Suppose that you want a stronger (more significant) fit, in which the line of fit is suspended more tightly. You must either have stiffer springs (have smaller variance in error), use more data (have more points to hang springs from), or move your points farther out on both ends of the  $x$ -axis (more leverage). The power of a test is how likely it is that you will be unable to move the line of fit given a certain energy budget (sum of squares) determined by the significance level.

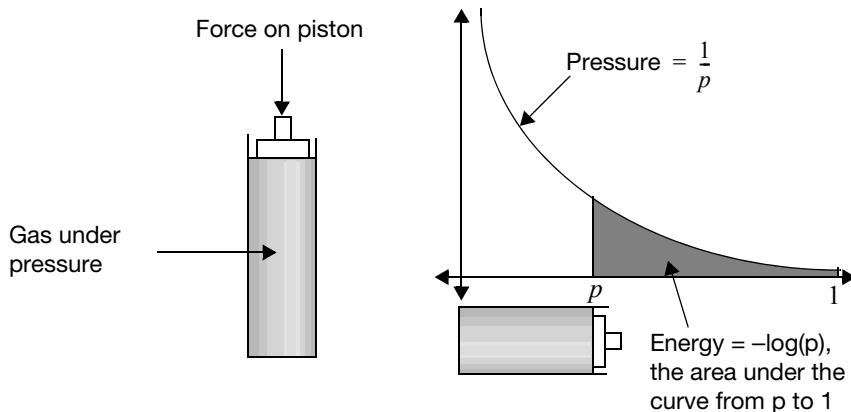
## Mechanics of Fit for Categorical Responses

Just as springs are analogous to least squares fits, gas pressure cylinders are analogous to maximum likelihood fits for categorical responses (see **Figure 19.12**).

## How Do Pressure Cylinders Behave?

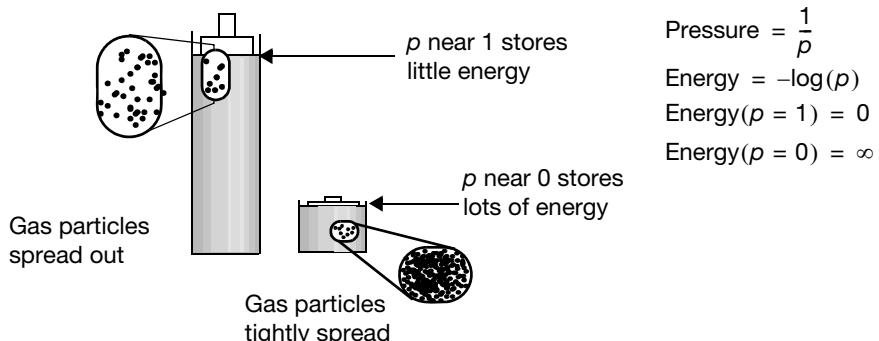
Using Boyle's law of gases (pressure times volume is constant), the pressure in a gas cylinder is proportional to the reciprocal of the distance from the bottom of the cylinder to the piston. The energy is the force integrated over the distance (starting from a distance,  $p$ , of 1), which turns out to be  $-\log(p)$ .

**Figure 19.12** Gas Pressure Cylinders Equate  $-\log(\text{probability})$  to Energy



Now that you know how pressure cylinders work, start thinking of the distance from the bottom of the cylinder to the piston as the probability that some statistical model attributes to some response. The height of 1 means no stored energy, no surprise, a probability of 1. The height of zero means infinite stored energy, an impossibility, a probability of zero.

When stretching springs, we measured energy by how much work it took to pull a spring, which turned out to be the square of the distance. Now we measure energy by how much work it takes to push a piston from distance 1 to distance  $p$ , which turns out to be  $-\log(p)$ , the logarithm of the probability. We used the logarithm of the probability before in categorical problems when we were doing maximum likelihood. The maximum likelihood method estimates the response probabilities so as to minimize the sum of the negative logarithms of the probability attributed to the responses that actually occurred. This is the same as minimizing the energy in gas pressure cylinders, as illustrated in **Figure 19.13**.

**Figure 19.13** Gas Pressure Cylinders Equate  $-\log(\text{probability})$  to Energy

## Estimating Probabilities

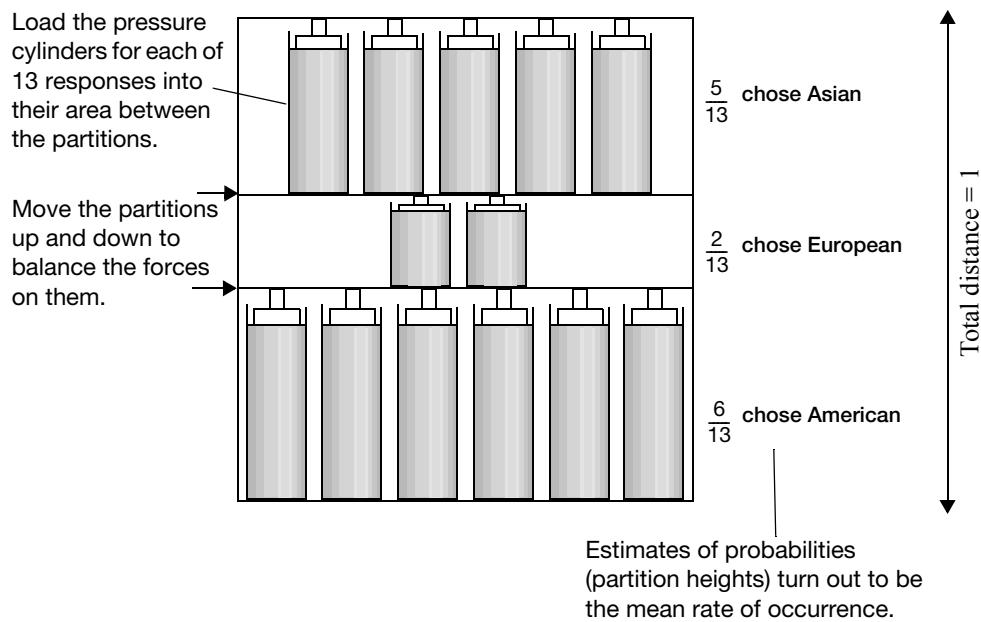
Now we want to estimate the probabilities by minimizing the energy stored in pressure cylinders. First, we need to build a partitioned frame with a compartment for each response category and add the constraint that the sum of the heights of the partitions is 1. We will move the partitions around so that the compartments for each response category can get bigger or smaller (see **Figure 19.14**).

For each observation on the categorical response, put a pressure cylinder into the compartment for that response. After you have all the pressure cylinders in the compartments, start moving the partitions around until the forces acting on the partitions balance out. This will be the solution to minimize the energy stored in the cylinders. It turns out that the solution for the minimization is to make the partition sizes proportional to the number of pressure cylinders in each compartment.

For example, suppose a survey asked 13 people what brand of car they preferred. Five people chose Asian, two chose European, and six chose American brands. Then you would stuff the pressure cylinders into the frame as in **Figure 19.14**, and the partition sizes that would balance the forces work out to  $5/13$ ,  $2/13$ , and  $6/13$ , which sum to 1.

To test that the true probabilities are some specific values, you move the partitions to those values and measure how much energy you had to add to the cylinders.

**Figure 19.14** Gas Pressure Cylinders Estimate Probabilities for a Categorical Response

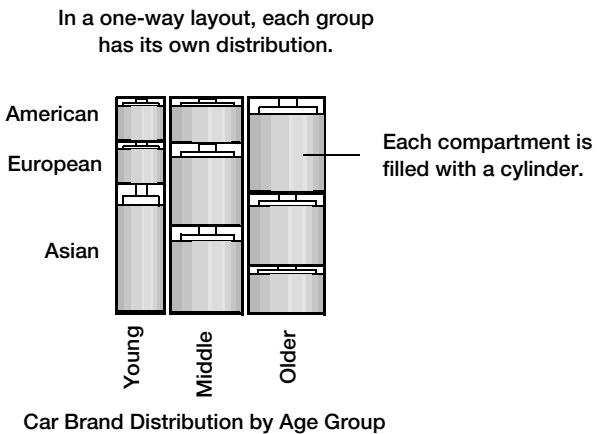


Minimize total energy in gas cylinders, that is, the sum of the negative logarithms of the probabilities associated with each.

## One-Way Layout for Categorical Data

If you have different groups, you can fit a different response probability to each group. The forces acting on the partitions balance independently for each group. The plot shown in **Figure 19.15** (which should remind you of a mosaic plot) helps maintain the visualization of pressure compartments. As an alternative to pressure cylinders, you can visualize with free gas in each cell.

**Figure 19.15** Gas Pressure Cylinder Estimate Probabilities for a Categorical Response

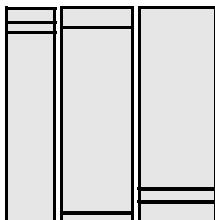


How do you test the hypothesis that the true rates are the same in each group and that the observed differences can be attributed to random variation? You just move the partitions so that they are in the position corresponding to the ungrouped population and measure how much more energy you had to add to the gas-pressure system to force the partitions to be in the same positions.

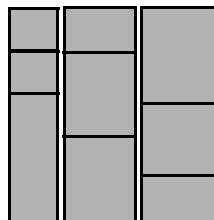
**Figure 19.16** shows the three kinds of results you can encounter, corresponding to perfect fit, significant difference, and nonsignificant difference. To the observer, the issue is whether knowing which group you are in will tell you which response level you will have. When the fit is near perfect, you know with near certainty. When the fit is intermediate, you have more information if you know the group you are in. When the fit is inconsequential, knowing which group you are in doesn't matter. To a statistician, though, what is interesting is how firmly the partitions are held by the gases, how much energy it would take to move the partitions, and what consequences would result from removing boundaries between samples and treating it as one big sample.

**Figure 19.16** Degrees of Fit

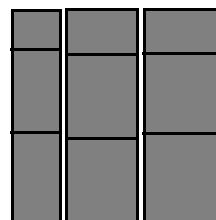
Three kinds of statistical results are possible.  
There are three response levels and three factor categories.

**Almost perfect fit**

Most probabilities of actual events are attributed to be near 1, lots of space for the gas.  
-log-likelihood is near zero. Huge differences with ungrouped case.  
LR test highly significant.

**Intermediate relationship**

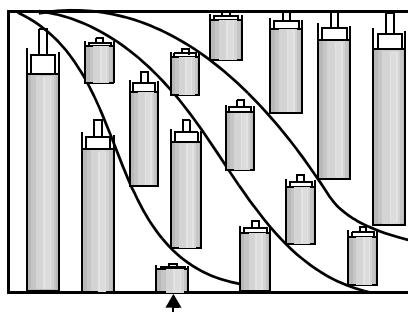
Probabilities of occurrence are not near 1, but they do differ among groups.  
-log-likelihood has intermediate value. Significant difference with ungrouped case by LR test.

**Homogeneous case**

Probabilities of occurrence are almost the same in each group.  
-log-likelihood has large value.  
Not much difference with ungrouped cases.  
Nonsignificant LR test.

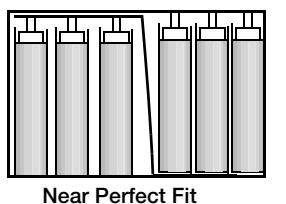
## Logistic Regression

Logistic regression is the fitting of probabilities over a categorical response to a continuous regressor. Logistic regression can also be visualized with pressure cylinders (see **Figure 19.17**). The difference with contingency tables is that the partitions change the probability as a continuous function of the  $x$ -axis. The distance between lines is the probability for one of the responses. The distances sum to a probability of 1. **Figure 19.18** shows what weak and strong relationships look like.

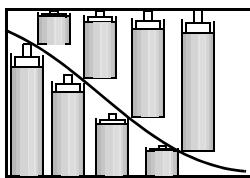
**Figure 19.17** Logistic Regression as the Balance of Cylinder Forces

The pressure cylinders push on the logistic curves that partition the responses, moving them so that the total energy in the cylinders is minimized. The more cylinders in an area, the more probability room it pushes for.

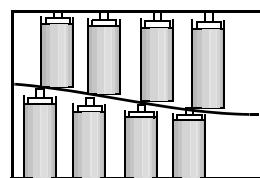
This is an outlier, a response that is holding a lot of energy, having been squeezed down to a small probability

**Figure 19.18** Strengths of Logistic Relationships

Near Perfect Fit



Strong Relationship



Weak Relationship

The probabilities are all near one. No cylinder is hot with energy.

Some cylinders must compete with nearby cylinders from a different response. More energy in some cylinders.

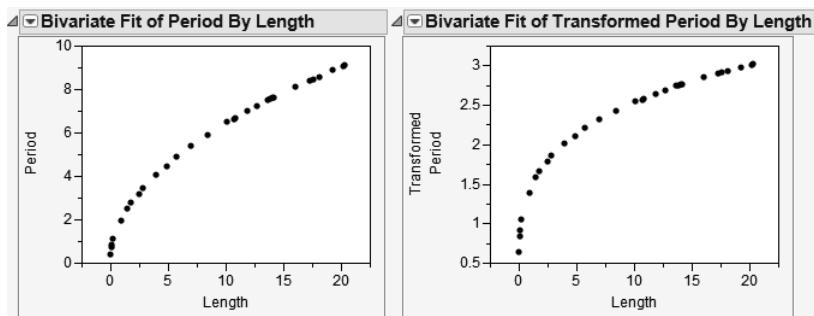
The probabilities are squeezed to near the horizontal case of homogeneity. All cylinders are warm with energy.



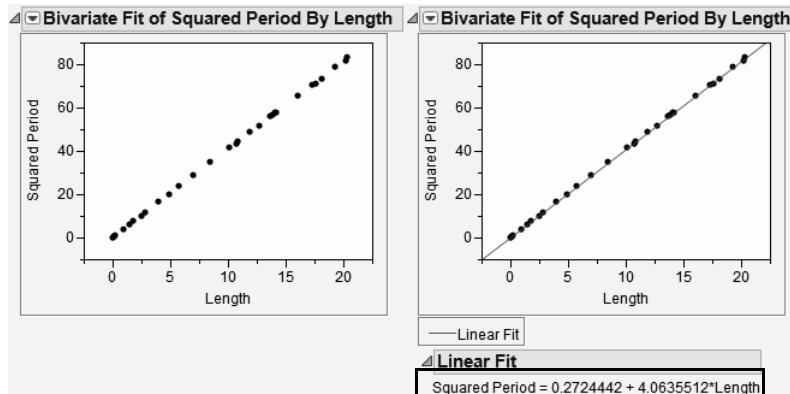
# Answers to Selected Exercises

## Chapter 4, “Formula Editor”

1a. (left) and 1b. (right)



1c. (left) - the Period<sup>2</sup> linearized the data, and 1d. (right), the line of best fit



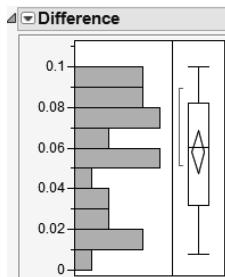
1e.  $\text{Period}^2 = 0.272 + 4.06 \times \text{Length}$ , so  $\text{Period} = \sqrt{0.272 + 4.06 \times \text{Length}}$

1f. Enter the formula below (left) into a new column. Then make another new column to find the difference between observed and theoretical period values (right).

$$\left( \frac{(2 \cdot \pi)}{\sqrt{9.8}} \right) \cdot \sqrt[2]{\text{Length}}$$

$\text{Period} - \text{Theoretical Period}$

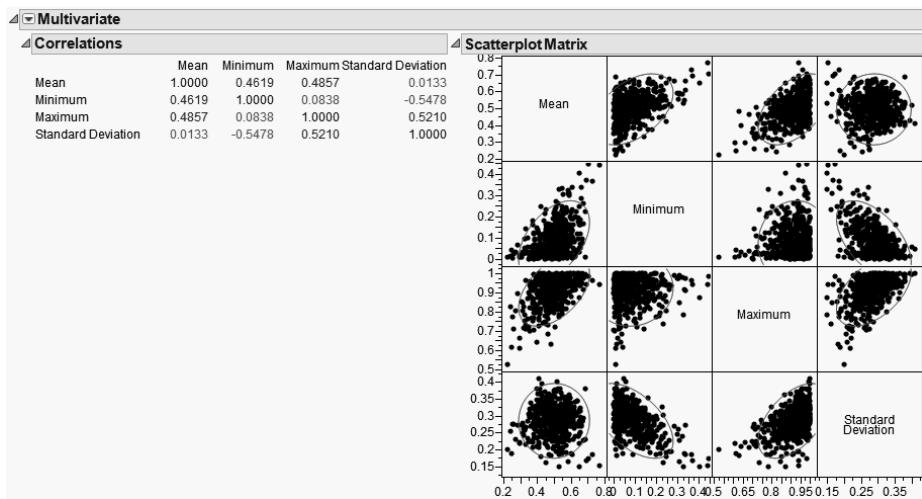
Use the Distribution platform to get the following histogram:



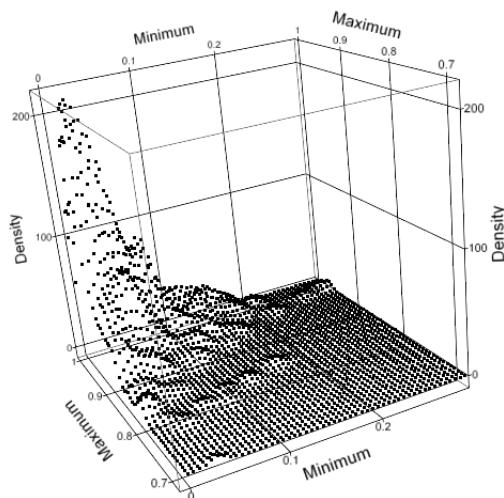
1g. This histogram reveals that the students' measurements are higher than the theoretically predicted values.

2b. For each column, select the statistic from the **Statistical Functions** group in the Formula Editor.

2c. The multivariate plot shows some correlation among the mean, minimum, and maximum, and among the standard deviation, minimum, and maximum.



2e. Minimum and Maximum yield the following using Scatterplot 3D.



3a. The value converges to  $\frac{\text{Fib}}{\text{Fib}_{\text{Row}-1}} \approx 1.618 \approx \frac{1 \pm \sqrt{5}}{2} = \phi$ , the golden ratio.

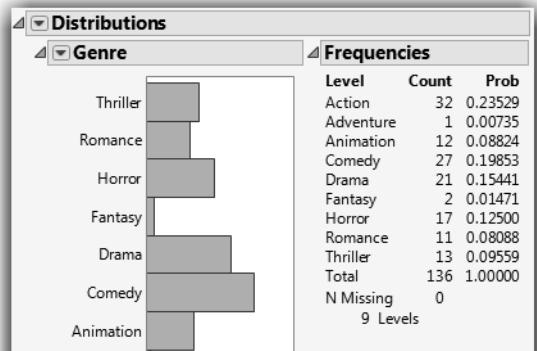
3c. The value converges to the same number.

3d. Again, the value converges to the same number.

3e. This time, the numbers converge to  $\frac{1}{4}$ .

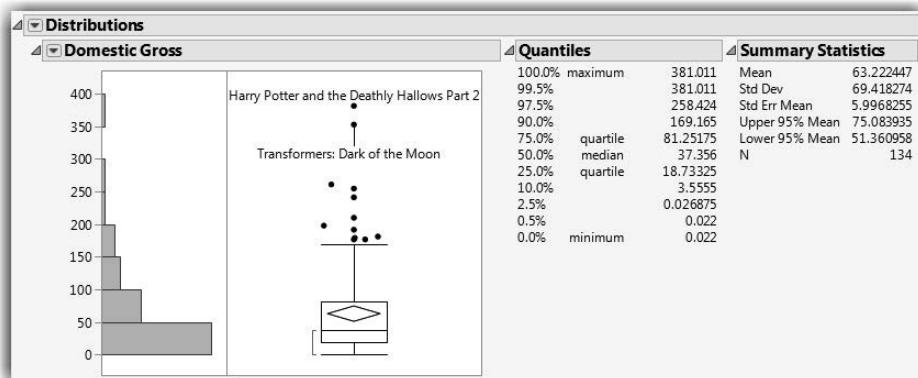
## Chapter 7, “Univariate Distributions: One Variable, One Sample”

1a. Levels and counts are shown in the Frequencies section of the report.



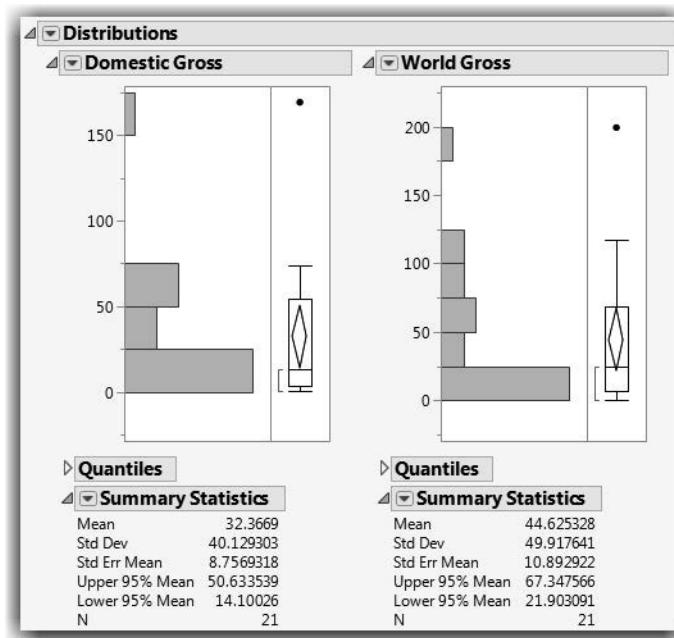
1b. The grosses range from \$22,000 to \$381 million with an average gross of \$63 million.

1c. Harry Potter and the Deathly Hallows, Part Two.

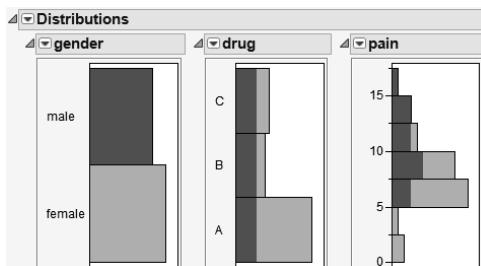


1d. To create the subset, use one of the following methods:

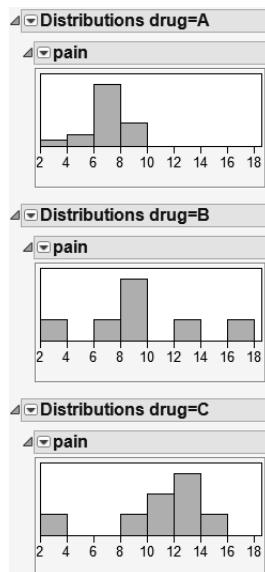
- To create the subset, use **Rows > Row Selection > Select Where** and complete the window to select where Genre equals Drama. Then, use **Tables > Subset** to create the data table.
- Return to the Distribution of Genre, and double-click on the bar for Drama.



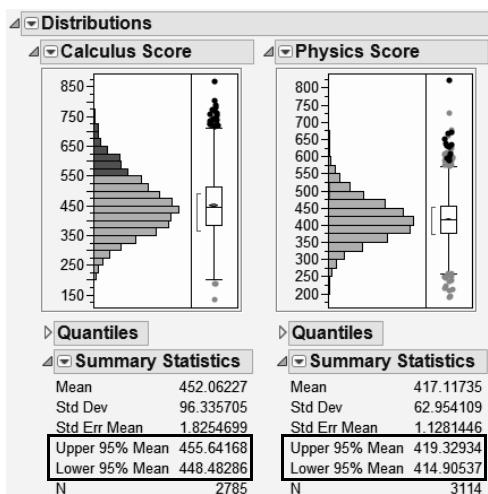
- For domestic gross, The Help appears to be an outlier. For foreign gross, Water for Elephants appears to be an outlier. However, the outlier rule applies only for symmetric distributions, and both of these distributions are skewed.
- 2a. The following picture has the males highlighted. There are far more females for drug A than males.



- 2b. To produce this report, select **Analyze > Distribution**, assign pain to **Y, Columns** and drug to **By**. Select **Uniform Scaling** from the top red triangle menu for each distribution. The means do not appear to be the same.



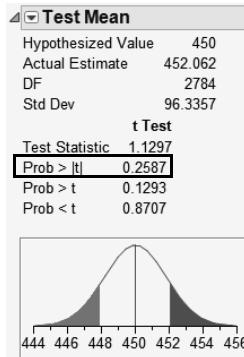
3a. High calculus scores do seem to correlate with high physics scores.



3b. To produce the relevant report, select Calculus Score as **Y, Columns** and Region as **By**. The means for the four regions are 467.54, 445.1, 464.9, and 441.27 respectively.

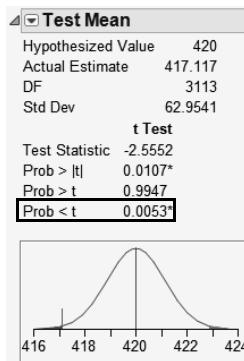
3c. The mean Physics scores for each of the four regions are 424.1, 404.8, 427.9, and 417.4 respectively.

3d. After requesting a distribution of the scores, use the **Test Mean** command from the platform menu to test that the mean is not 450. The following report appears, showing that there is not evidence that the mean is different from 450.



3e. The confidence interval is shown in the **Summary Statistics** section of the calculus score report in question 3a as Lower 95% Mean and Upper 95% Mean (448.48, 455.64).

3f. After requesting a distribution of the scores, use the **Test Mean** command from the platform menu. The resulting report shows that the mean appears to be less than 420.



3g. The confidence interval is shown in the **Moments** section of the physics score report in question 3a (414.9, 419.33).

4a. The mean is 1.44 g. Three cereals, 100% Natural Bran Oats & Honey, Banana Nut Crunch, and Cracklin' Oat Bran, appear to have unusually high amounts of fat.

4b. "All Bran with Extra Fiber" and "Fiber One"

4c. Cold cereals: (8.15g, 10.78g); Hot cereals (-2.46g, 5.13g). Note that there are only three hot cereals in the data set.

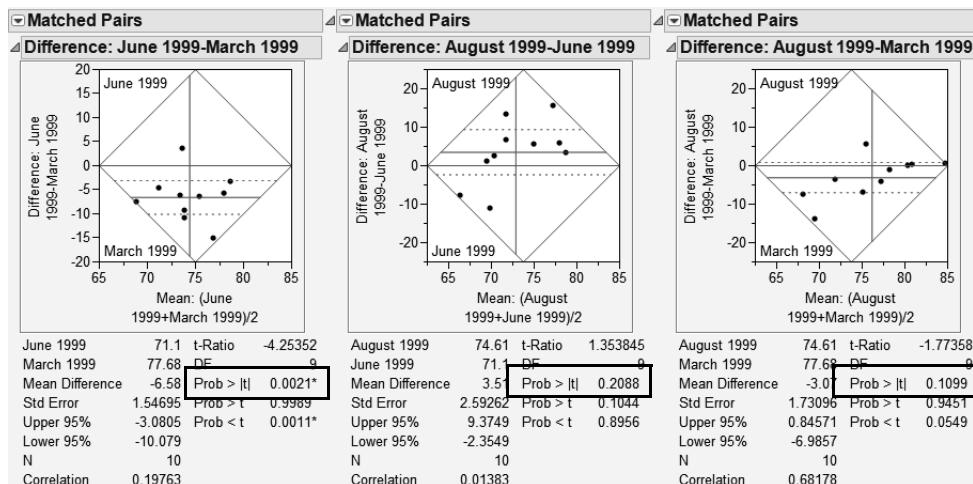
5a. Auto and Robbery seem to be skewed. The others have a more of a bell-shaped appearance.

5b. Nevada and New York

## Chapter 8, "The Difference Between Two Means"

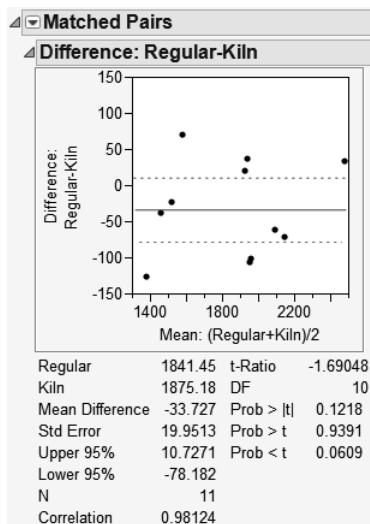
8a. A matched pairs approach is more appropriate, since these are repeated measures over time.

8b. The Matched Pairs platform yields the following report, showing a significant difference between the two months at a significance level of 0.05 (on the left, below).

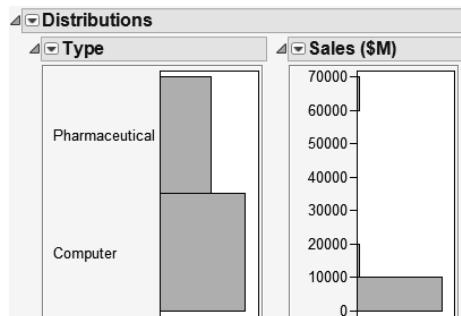


c. There is no evidence for a significant difference between August and June (middle, above). Similarly, there is no evidence for a difference between August and March (right, above).

2b. There does not appear to be strong evidence between the two ( $p$ -value is 0.1218). However, sample size is small. The result is marginal and deserves further investigation.

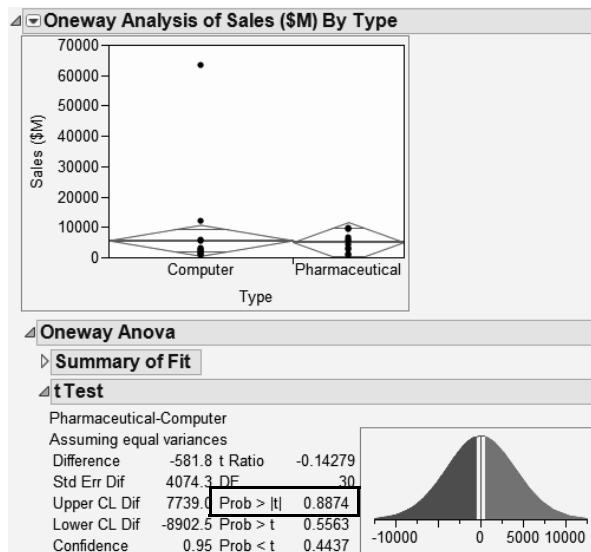


3a. The histograms are shown here. Note the outlier in the sales column and the skewed nature of the distribution.

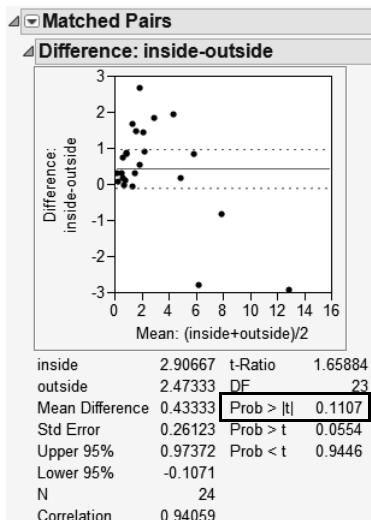


3b. Grouped means are appropriate in this situation.

3c. Using **Fit Y by X** with Sales as Y and Type as X allows for the pooled or unpooled two-sample t-Test. The **Means/Anova/ Pooled t** command, which produces the report shown below, does not show evidence of a difference. Similarly, using the unpooled **t Test** command also does not show evidence of a difference.

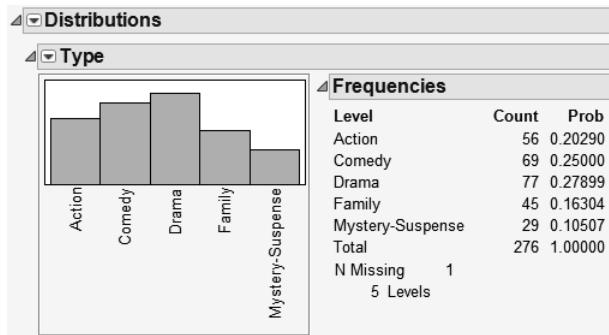


5b. The following report comes from the Matched Pairs platform. The  $p$ -value of 0.11 is a nonsignificant value.



## Chapter 9, “Comparing Many Means: One-Way Analysis of Variance”

1a. There are five levels. There is not an equal number of movies of each type (more Drama, fewer Mystery-Suspense).

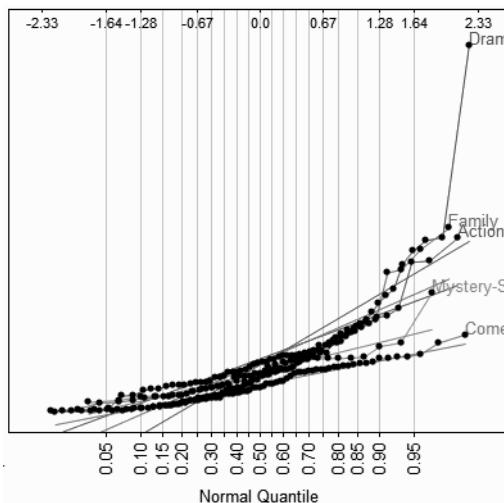


1b. The null hypothesis is that the average Worldwide \$ for different types of movies is the same. With a *p*-value of 0.0020, there is evidence for a difference between at least two movie types.

Analysis of Variance					
Source	DF	Sum of Squares		F Ratio	Prob > F
		Mean Square	F		
Type	4	591566.4	147892	4.3597	0.0020*
Error	271	9192954.0	33922		
C. Total	275	9784520.4			

1c. Action and Drama are not different from all other movie types—they are both significantly different from Comedy.

1d. Since the lines in the normal quantile plot appear to have very different slopes, a Welch ANOVA is not a bad idea.



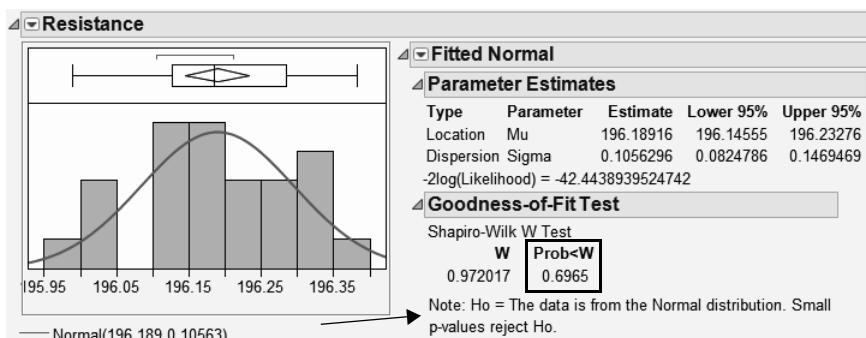
Test	F Ratio	DFNum	DDen	Prob > F
O'Brien[.5]	1.7421	4	271	0.1410
Brown-Forsythe	3.8784	4	271	0.0044*
Levene	5.6545	4	271	0.0002*
Bartlett	21.1239	4	.	<.0001*

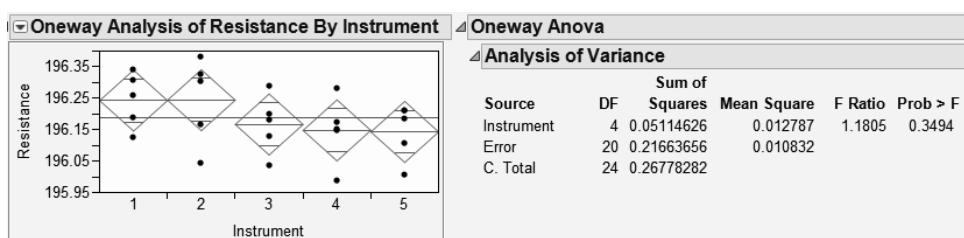
<b>Welch's Test</b>				
Welch Anova testing Means Equal, allowing Std Devs Not Equal				
F Ratio	DFNum	DDen	Prob > F	
9.2975	4	109.85	<.0001*	

The output from the UnEqual Variances command shows four tests that the variances are unequal. Three support the conclusion that they are not equal. The Welch ANOVA shows a similar conclusion to the parametric ANOVA: There is a difference among the movie types.

2a. An examination of the histograms shows that **Fit Distribution > Normal** overlays a normal curve. Select **Goodness of Fit** from the red triangle menu next to Fitted Normal. The null hypothesis is stated in the **Goodness-of-Fit Test** results. The data appear to be normal.



2b. The null hypothesis is that there is no difference in the mean resistance for the different instruments. Fit Y by X is used to generate the ANOVA shown here. There is no evidence that the instruments differ.

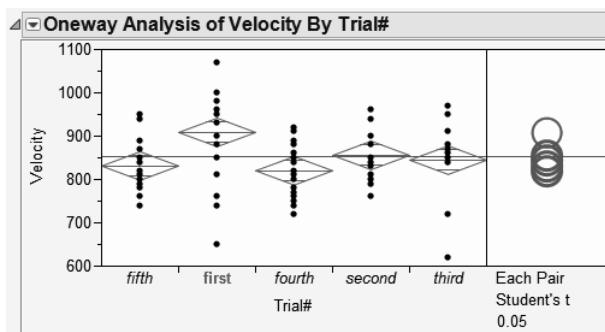


3a. 299,852.4 km/sec

3b. There is evidence that the trials differ.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Trial#	4	94514.00	23628.5	4.2878	0.0031*
Error	95	523510.00	5510.6		
C. Total	99	618024.00			

3c. The first set of observations is higher than the others on average.



3d. Excluding the first group and re-computing the mean result in a mean of 299838.25, which is closer to the true value of 299792.5 km/sec.

4b. All three give the same (significant) result.

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)						Van der Waerden Test (Normal Quantiles)					
Level	Count	Score Sum	Expected			Level	Count	Score Sum	Expected		
			Score	Score Mean	(Mean-Mean0)/Std0				Score	Score Mean	(Mean-Mean0)/Std0
1	22	687.000	1056.00	31.2273	-3.251	1	22	-10.591	0.000	-0.48139	-2.672
2	22	1128.50	1056.00	51.2955	0.635	2	22	2.206	0.000	0.10029	0.557
3	19	830.500	912.000	43.7105	-0.754	3	19	-3.154	0.000	-0.16599	-0.839
4	19	1081.00	912.000	56.8947	1.568	4	19	4.995	0.000	0.26292	1.329
5	13	833.000	624.000	64.0769	2.258	5	13	6.543	0.000	0.50327	2.026
1-way Test, ChiSquare Approximation						1-way Test, ChiSquare Approximation					
ChiSquare	DF	Prob>ChiSq				ChiSquare	DF	Prob>ChiSq			
15.3185	4	0.0041*				11.2432	4	0.0240*			
Median Test (Number of Points Above Median)											
Level	Count	Score Sum	Expected			Level	Count	Score Sum			
			Score	Score Mean	(Mean-Mean0)/Std0				Score	Score Mean	(Mean-Mean0)/Std0
1	22	4.000	10.884	0.181818	-3.331						
2	22	12.000	10.884	0.545455	0.540						
3	19	8.000	9.400	0.421053	-0.714						
4	19	13.000	9.400	0.684211	1.837						
5	13	10.000	6.432	0.769231	2.119						
1-way Test, ChiSquare Approximation											
ChiSquare	DF	Prob>ChiSq									
15.7366	4	0.0034*									

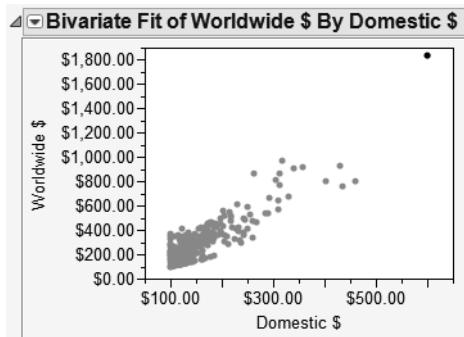
4c. An ANOVA shows a nonsignificant result.

5a. There is a difference among the regions for Calculus scores.

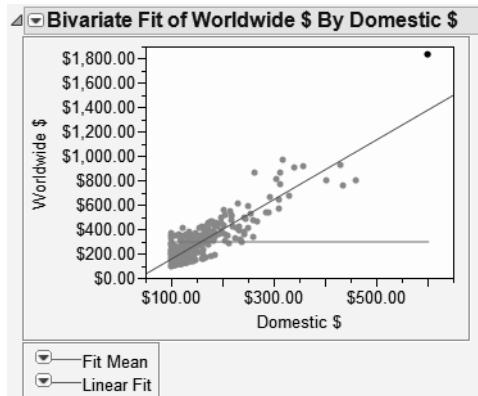
- 5b. There is a difference among the regions for Physics scores.  
 5c. Groups 1 and 3 appear to be quite different from groups 2 and 4.

## Chapter 10, “Fitting Curves through Points: Regression”

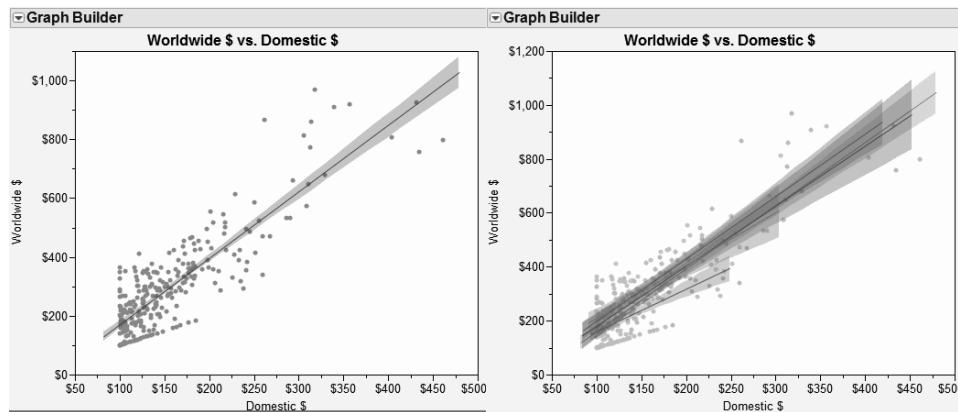
- 1a. One movie (Titanic) is definitely separated from the rest of the movies.



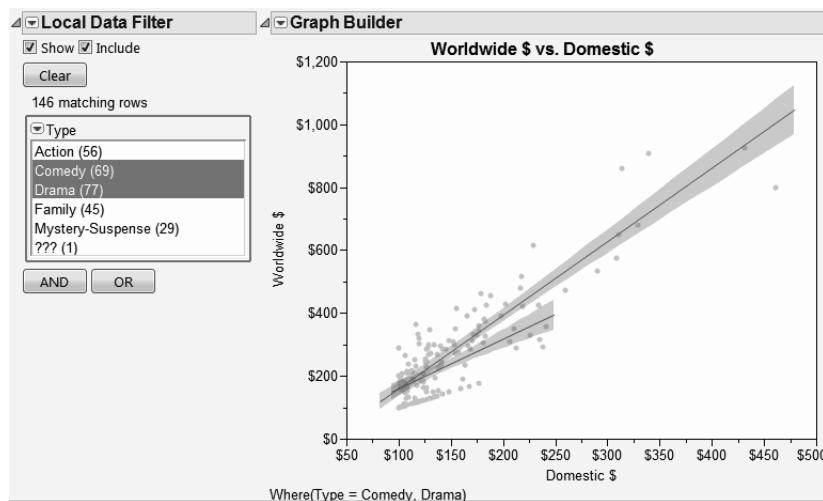
- 1b and 1c. The linear model does explain the data better than the simple mean model. (Why?)



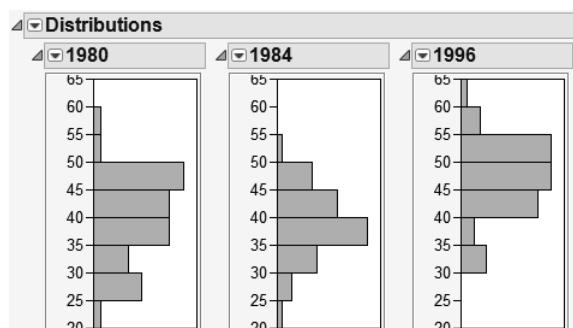
- 1d. The model with excluded outliers is probably a better summary of a typical movie's performance.  
 1e (left) and 1f (right). There are differences in gross dollars for the different types of movies.



1g. The regression lines for the two types of movies are different.

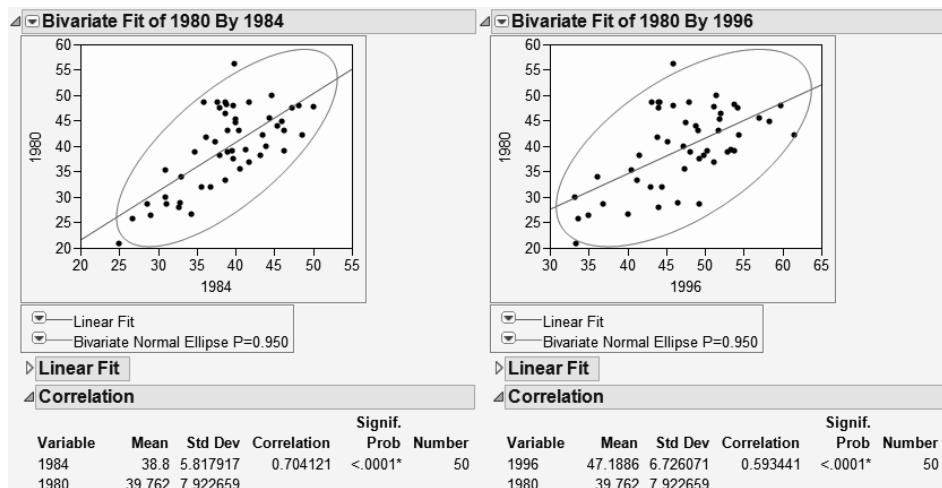


2a. The mean for 1996 is about 10 points higher than for the other two elections.



Perhaps 1996 (Clinton vs. Dole) was a more contested election than the others (Reagan vs. Carter, Reagan vs. Mondale).

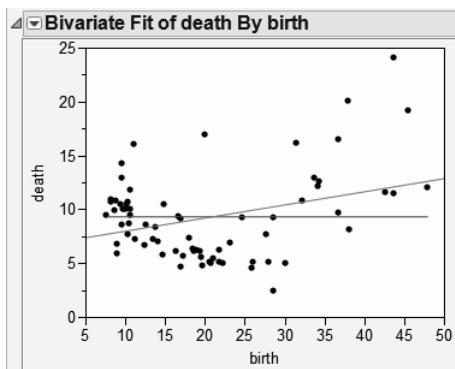
2b. 1984 vs. 1980: 0.704. 1996 vs 1980: 0.593. Correlations are stronger for years that are closer together.



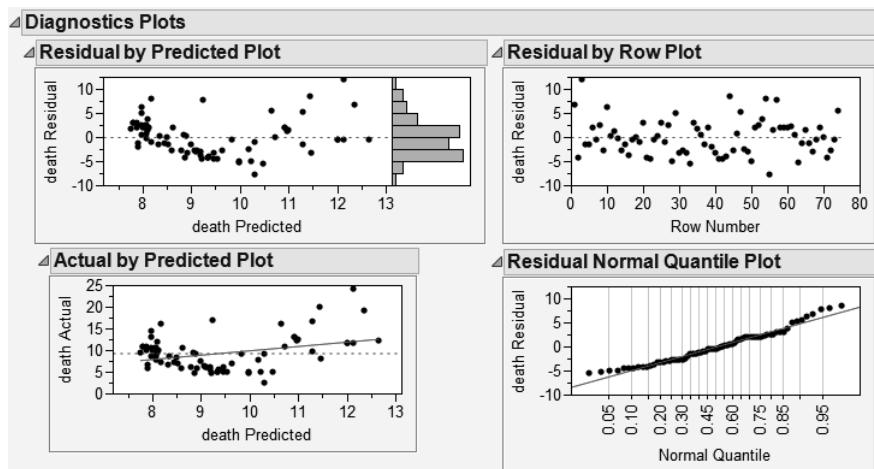
2c. Probably not. The next presidential election is far removed from 1996, 1984, and 1980.

3a. Afghanistan, Angola, and Mozambique are outliers for the death-rate variable.

3b. The line is a better predictor than the mean, but a line is not an appropriate model (see part d of this question).



3c. There is a u-shaped pattern to these residuals (in the first two, shown on the left below).



d. Because of the pattern of the residuals (and, actually, the pattern of the original data—clearly nonlinear), a line is not an appropriate model.

4a. 1-Octanol

4b. Right-click on the Parameter Estimates report and select **Columns > Lower 95%** and **Columns Upper 95%** to reveal the confidence interval we want.

Parameter Estimates						
Term	Estimate	Std Error	t Ratio	Prob> t	Lower 95%	Upper 95%
Intercept	0.600659	0.054128	11.10	<.0001*	0.4927051	0.7086129
Hexane	0.8638495	0.034004	25.40	<.0001*	0.7960313	0.9316677

## Chapter 11, “Categorical Distributions”

1a. The Test Probabilities command from the Distribution report gives the following results:

Test Probabilities			
Level	Estim	Prob	Hypoth Prob
Blue	0.21000	0.20000	
Brown	0.22000	0.30000	
Green	0.18000	0.20000	
Red	0.23000	0.20000	
Yellow	0.16000	0.10000	
Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	6.0786	4	0.1934
Pearson	6.4333	4	0.1690

Based on these numbers, there's no reason to dispute the company's claim.

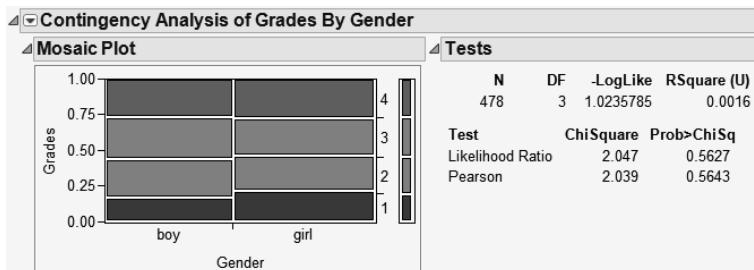
2. We entered the data into two columns of a data table, one representing the machines and the second representing the counts. Then, we selected **Analyze > Distribution** with Machine as **Y** and the counts as **Freq**. In the Distribution report, we used **Test Probabilities** to test that they were all the same.

<b>Test Probabilities</b>			
Level	Estim Prob	Hypoth Prob	
A	0.32554	0.33333	
B	0.41735	0.33333	
C	0.25711	0.33333	
<b>Test</b>	<b>ChiSquare</b>	<b>DF</b>	<b>Prob&gt;Chisq</b>
Likelihood Ratio	183.8488	2	<.0001*
Pearson	184.2160	2	<.0001*

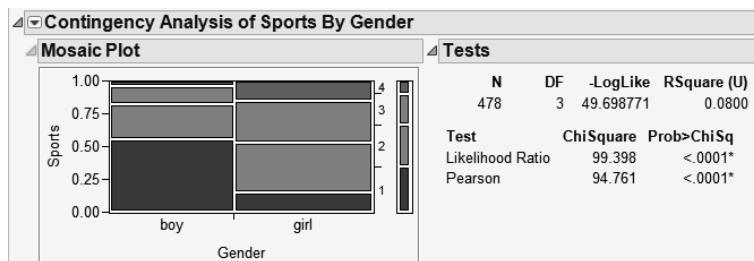
They clearly are different.

## Chapter 12, “Categorical Models”

1a. Use **Analyze > Fit y by X**, with Gender as **X** and Grades as **Y**. There is no significant difference.



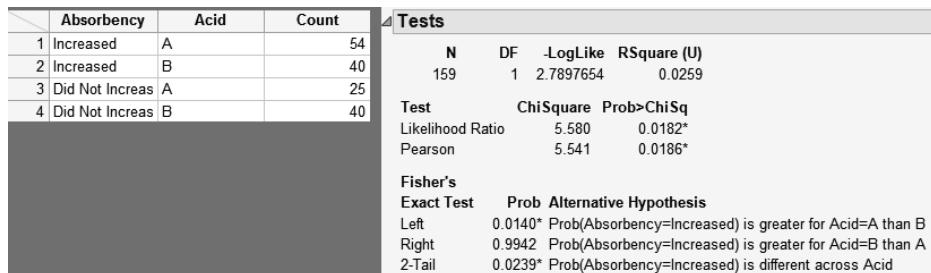
1b. Yes



1c. Looks: Yes; Money: No

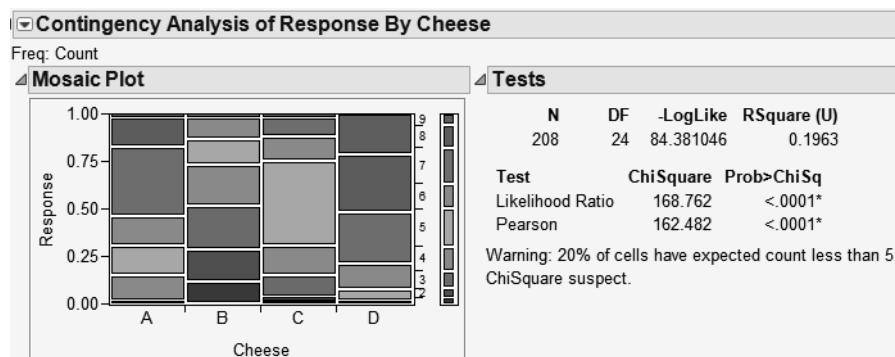
1d. Grades: No; Sports: No; Looks: No; Money: No

2. Create a data table, as shown below. Then, use **Analyze > Fit Y by X** with Absorbency as **Y**, Acid as **X**, and Count as **Freq**.

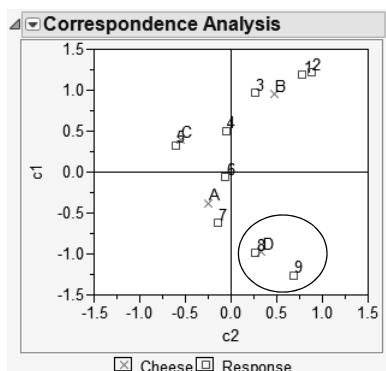


There is evidence of a difference in absorbency between the two acids.

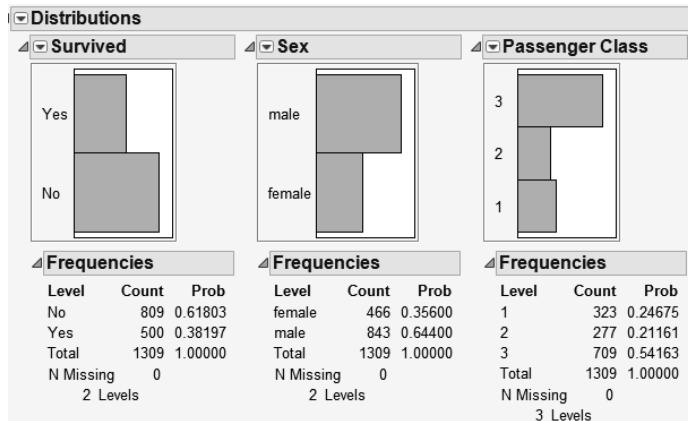
3a and 3b: The mosaic plot and the tests indicate that there is a difference in additives.



3c. Cheese D is associated with high tasting scores. Cheese A comes in second.



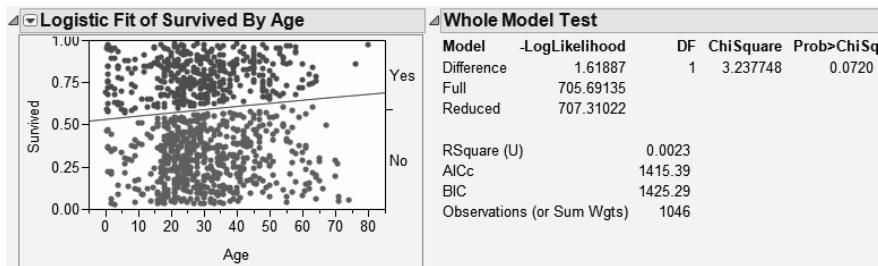
4a, 4b, and 4c. See the counts below.



4d (left) and 4e (right). Both passenger class and sex are significant. There are differences in the survival rate for different passenger classes and different sexes.

Tests				Tests			
	N	DF	-LogLike		N	DF	-LogLike
	1309	2	63.882734	RSquare (U)	1309	1	186.46067
Test				RSquare (U)			
Likelihood Ratio			127.765	<.0001*	Likelihood Ratio		372.921
Pearson			127.859	<.0001*	Pearson		365.887
				Prob>ChiSq			

4f. Age is marginally significant, with a *p*-value of 0.0720.



**Note:** For a better model, use Fit Model with multiple Xs.

## Chapter 13, “Multiple Regression”

1a. 0.533

1b. Yes, it increases to 0.893.

2b. The effect with the least significance is Shoulder, which has a *p*-value of 0.90.

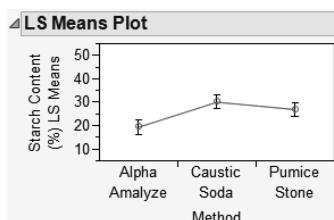
- 2c. After several repetitions, Fore, Waist, Height, and Thigh remain in the model.
- 2d. We reach an identical model.
- 3a. Dreaming and Non-Dreaming. Note that this model is degenerate since there is no variability in Dreaming that is not accounted for by the Non-Dreaming variable.
- 3b. Dreaming and Exposure.
- 3c. The  $R^2$  is 0.934.
- 3d. Forward stepwise gives a model with Dreaming, Gestation, and Danger.
- 3e. Mixed stepwise gives the same model as part d.

## Chapter 14, “Fitting Linear Models”

- 1a. The Summary of Fit and Analysis of Variance tables are the same in both platforms (below, Fit Y by X is shown left, and Fit Model is right).

Oneway Analysis of Starch Content (%) By Method					Response Starch Content (%)					
Oneway Anova					Whole Model					
Summary of Fit					Summary of Fit					
Rsquare	0.216702				RSquare	0.216702				
Adj Rsquare	0.200211				RSquare Adj	0.200211				
Root Mean Square Error	8.636243				Root Mean Square Error	8.636243				
Mean of Response	25.51663				Mean of Response	25.51663				
Observations (or Sum Wgts)	98				Observations (or Sum Wgts)	98				
Analysis of Variance					Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	Source	DF	Sum of Squares	Mean Square	F Ratio	
Method	2	1960.2363	980.118	13.1410	<.0001*	Model	2	1960.2363	980.118	13.1410
Error	95	7085.5450	74.585			Error	95	7085.5450	74.585	Prob > F
C. Total	97	9045.7813				C. Total	97	9045.7813		<.0001*

- 1b. This plot (like the comparison circles) tell us that Caustic Soda and Pumice Stone remove more starch than Alpha Amalyze.



- 1c. All effects are significant.

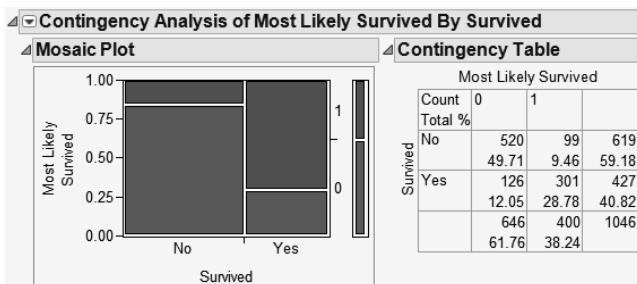
- 1d. The Method LS means for Method is identical to the one in part (b). The Sand Blasted plot indicates that sand blasting removes more starch than not.

1e. It is ( $F = 28.66, p < 0.001$ ).

1f. No interaction effects are significant.

Effect Tests					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Size of Load (lbs)	1	1	686.0201	10.6170	0.0016*
Sand blasted?	1	1	260.5629	4.0325	0.0477*
Method	2	2	1984.3526	15.3551	<.0001*
Method*Size of Load (lbs)	2	2	135.5183	1.0487	0.3547
Method*Sand blasted?	2	2	166.6581	1.2896	0.2805
Size of Load (lbs)*Sand blasted?	1	1	5.6768	0.0879	0.7676

2b. The model predicted correctly about 78% of the time -  $(520 + 301)/1046$ .

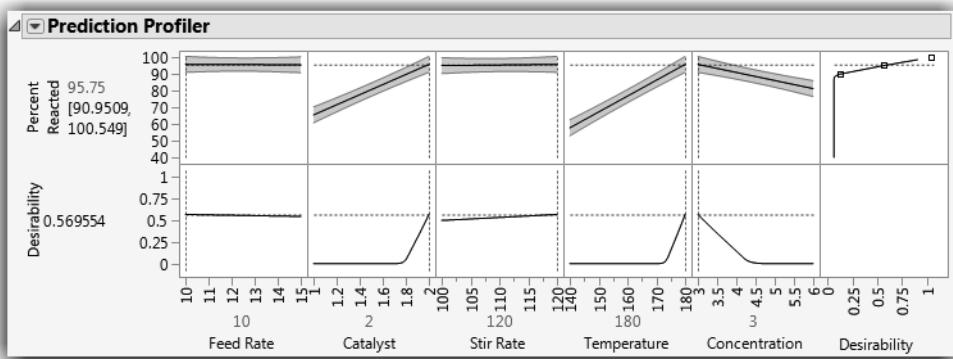


2c. All of the two-way interactions are significant.

## Chapter 15, “Design of Experiments”

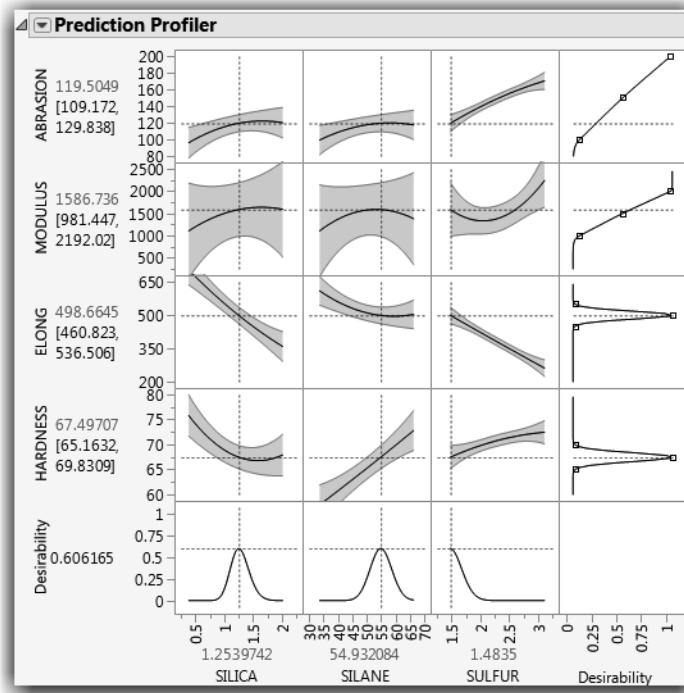
1b. Only main effects and two-way interactions are significant.

1d. The optimal settings are below. Results are the same as the 20-run reactor experiment.



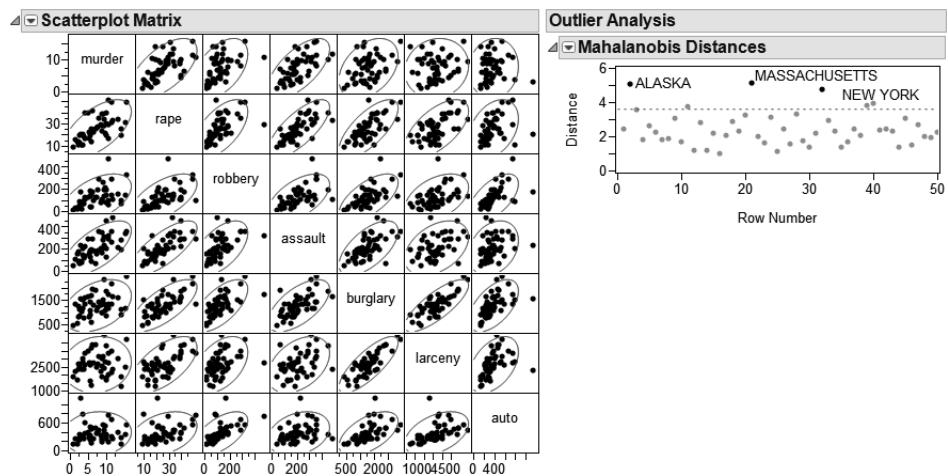
3a. 5 levels for each factor.

3d. Optimal settings are below. The goals were met for HARDNESS and ELONG. The model didn't do as well in maximizing ABRASION.



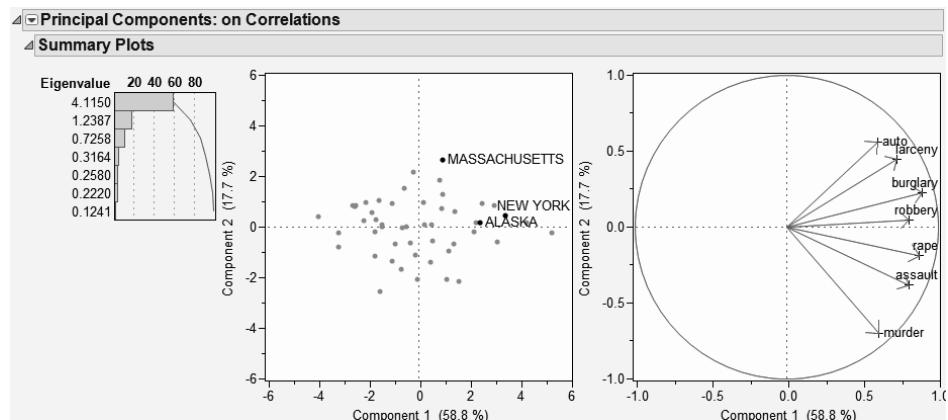
## Chapter 16, “Bivariate and Multivariate Relationships”

- Most of the variables are positively correlated. Alaska, Massachusetts, and New York are outliers.

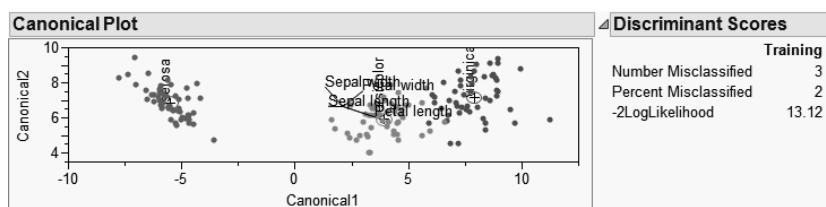


1c. Auto theft and robbery are close together. Murder is by itself.

1d. 2 or 3



3. JMP produces a model that misclassifies only three flowers.



4. The distance plot suggests around 4 clusters.

## Chapter 17, “Exploratory Modeling”

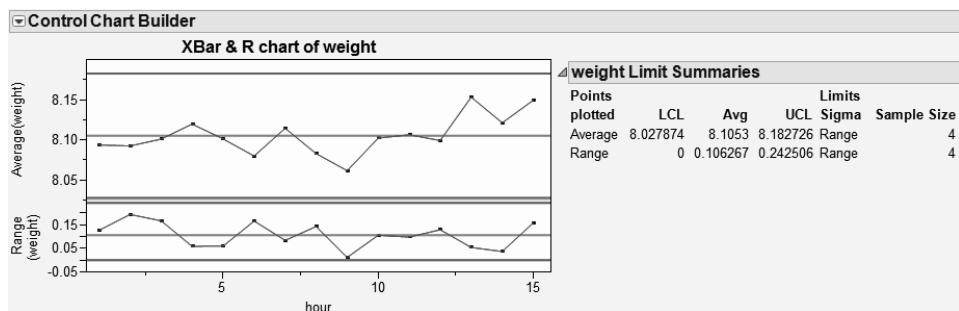
2a. Minimization of Time appears at Length=2.5, Freq=250, and Space = 0.721. This is from using the following model.

$$\begin{aligned}
 & -3.2452922536306 \\
 & +3.45254884159199 * \text{Squish} \left( \begin{array}{l} 1.23349639217962 \\ +4.0313426311121 * \text{Length} \\ +0.0079663084945 * \text{Freq} \\ +5.55404075455771 * \text{Space} \end{array} \right) \\
 & +2.2135073851837 * \text{Squish} \left( \begin{array}{l} 19.6355924059805 \\ +1.24984647690406 * \text{Length} \\ +0.0469271850159 * \text{Freq} \\ +17.262879335114 * \text{Space} \end{array} \right) \\
 & +2.76379447848082 * \text{Squish} \left( \begin{array}{l} -8.2031099033942 \\ +1.8298210665791 * \text{Length} \\ +0.00754281425927 * \text{Freq} \\ +13.2247711095988 * \text{Space} \end{array} \right) \\
 & *4.53631182790601
 \end{aligned}$$

## Chapter 18, “Control Charts and Capability”

1a. An XBar R or S is appropriate. There were multiple readings per hour.

1b and 1c. The mean is 8.1053. The process appears stable. No special causes signal.

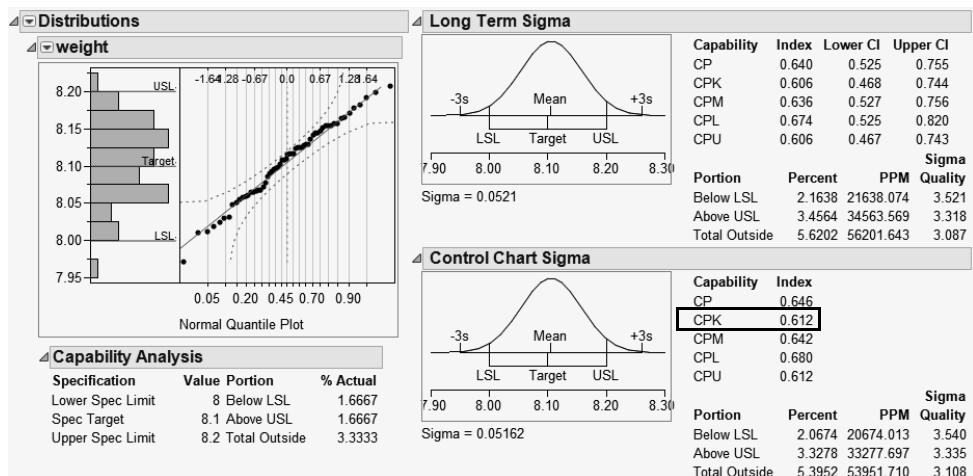


2a. In question 1, we saw that the process is stable and that the underlying distribution is normal.

2b and 2c. 3.33% fell outside the spec limits. 5.4% are predicted in the long term.

2d. The process is not capable. The  $C_{PK}$  is below 1.0.

2e. The process is on target. Engineers should focus on variability reduction.





## References and Data Sources

- Agresti, A. (1984), *Analysis of Ordinal Categorical Data*, New York: John Wiley and Sons, Inc.
- Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley and Sons, Inc.
- Allison, T., and Cicchetti, D. V. (1976), "Sleep in Mammals: Ecological and Constitutional Correlates." *Science*, November 12, 194, 732-734.
- American Society for Testing and Materials. *ASTM Manual on Presentation of Data and Control Chart Analysis*. STP No. 15-D. Philadelphia, 1976.
- Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, New York: John Wiley and Sons.
- Andrews, D.F. and A. M. Herzberg (1985), *Data: A Collection of Problems from Many Fields for the Student Worker*, New York: Springer-Verlag.
- Anscombe, F.J. (1973), *American Statistician*, 27, 17-21.
- Aviation Consumer Home Page*. U.S. Government, Department of Transportation.  
Retrieved from <https://www.transportation.gov/airconsumer>.
- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics*, New York: John Wiley and Sons.
- Berger, R. L., and Hsu, J. C. (1996), "Bioequivalence Trails, Intersection-Union Tests, and Equivalence Confidence Sets," *Statistical Science*, 11, 283-319.
- Box G.E.P., Jenkins G. M. and Reisnel, G.C. (1994), *Time Series Analysis: Forecasting and Control, Third Edition*. Prentice Hall: Englewood Cliffs, NJ.
- Box, G.E.P., Hunter, W.G., and Hunter, J.S. (2005), *Statistics for Experimenters*, New York: John Wiley and Sons, Inc.
- Chase, M. A., and Dummer, G. M. (1992), "The Role of Sports as a Social Determinant for Children," *Research Quarterly for Exercise and Sport*, 63, 418-424.

- Cochran, W.G. and Cox, G.M. (1957), *Experimental Designs, 2nd edition*, New York: John Wiley and Sons.
- Creighton, W. P. (2000), "Starch content's dependence on several manufacturing factors." Unpublished data.
- Cryer, J. and Wittmer, J. (1999) Notes on the Theory of Inference. NCSSM Summer Statistics Institute. Retrieved from [http://courses.ncssm.edu/math/Stat\\_Inst/Notes.htm](http://courses.ncssm.edu/math/Stat_Inst/Notes.htm).
- Daniel, C. (1959), "Use of Half-Normal Plots in Interpreting Factorial Two-level Experiments," *Technometrics*, 1, 311-314.
- Data and Story Library*, Retrieved from <http://lib.stat.cmu.edu/DASL/>.
- Ehrstein, James and Croarkin, M. Carroll. *Statistical Reference Datasets*. US Government, National Institute of Standards and Technology. Retrieved from <https://nist.gov/itl/products-services/statistical-reference-datasets>.
- Eppright, E.S., Fox, H.M., Fryer, B.A., Lamkin, G.H., Vivian, V.M., and Fuller, E.S. (1972), "Nutrition of Infants and Preschool Children in the North Central Region of the United States of America," *World Review of Nutrition and Dietetics*, 14.
- Eubank, R.L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker, Inc.
- Farebrother, R. W. (2002), *Visualizing Statistical Models and Concepts*, New York: Marcel Dekker, Inc.
- Fortune Magazine (1990), *The Fortune 500 List*, April 23, 1990.
- Gabriel, K.R. (1982), "Biplot," *Encyclopedia of Statistical Sciences, Volume 1*, eds. N.L.Johnson and S. Kotz, New York: John Wiley and Sons, Inc., 263-271.
- Gosset, W.S. (1908), "The Probable Error of a Mean," *Biometrika*, 6, pp 1-25.
- Hajek, J. (1969), *A Course in Nonparametric Statistics*, San Francisco: Holden-Day.
- Henderson, H. V. and Velleman, P. F. (1981), "Building Regression Models Interactively." *Biometrics*, 37, 391-411. Data originally collected from Consumer Reports.
- Hosmer, D.W. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley and Sons.
- Iman, R.L. (1995), *A Data-Based Approach to Statistics*, Belmont, CA: Duxbury Press.
- Iman, R.L. and Conover, W.J. (1979), "The Use of Rank Transform in Regression," *Technometrics*, 21, 499-509.
- Isaac, R. (1995), *The Pleasures of Probability*, New York: Springer-Verlag.
- John, P.M. (1971), *Statistical Design and Analysis of Experiments*, New York: Macmillan.

- Jones, B. and Nachtsheim, C.J. (2009), "Split-Plot Designs: What, Why, and How," *Journal of Quality Technology*, Vol 41, No. 4, pp 340-361.
- Kackar, R.N. and Harville, D.A. (1984), Approximations for standard errors of estimators of fixed and random effects in mixed linear models, *Journal of the American Statistical Association*, 79, 853-862
- Kaiser, H.F. (1958), "The varimax criterion for analytic rotation in factor analysis" *Psychometrika*, 23, 187-200.
- Kemp, A.W. and Kemp, C.D. (1991), "Weldon's dice data revisited," *The American Statistician*, 45 216-222.
- Klawiter, B. (2000), *An Investigation into the Potential for Geochemical / Geoarchaeological Provenance of Prairie du Chien Cherts*. Master's Thesis: University of Minnesota.
- Koehler, G. and Dunn, J.D. (1988), "The Relationship Between Chemical Structure and the Logarithm of the Partition," *Quantitative Structure Activity Relationships*, 7.
- Koopmans, L. (1987), *Introduction to Contemporary Statistical Methods*, Belmont, CA: Duxbury Press, p 86.
- Ladd, T. E.(1980 and 1984) and Carle, R. H. (1996), Clerks of the House of Representatives. *Statistics of the Presidential and Congressional Elections*. US Government. Retrieved from [http://clerk.house.gov/member\\_info/election.aspx](http://clerk.house.gov/member_info/election.aspx).
- Larner, M. (1996), Mass and its Relationship to Physical Measurements. MS305 Data Project, Department of Mathematics, University of Queensland.
- Lenth, R.V. (1989), "Quick and Easy Analysis of Unreplicated Fractional Factorials," *Technometrics*, 31, 469-473.
- Linnerud (see Rawlings (1988)).
- McCullagh, P. and Nelder, J. A. (1983), *Generalized Linear Models*. From *Monographs on Statistics and Applied Probability*, Cox, D. R. and Hinkley, D. V., eds. London: Chapman and Hall, Ltd.
- Miller, A.J. (1990), *Subset Selection in Regression*, New York: Chapman and Hall.
- Moore, D.S. and McCabe, G. P. (1989), *Introduction to the Practice of Statistics*, New York and London: W. H. Freeman and Company.
- Myers-Briggs Type Indicator (MBTI) Basics*. Retrieved from <http://www.myersbriggs.org/my-mbti-personality-type/mbti-basics>.
- Nelson, L. (1984), "The Shewhart Control Chart - Tests for Special Causes," *Journal of Quality Technology*, 15, 237-239.
- Nelson, L. (1985), "Interpreting Shewhart X Control Charts," *Journal of Quality Technology*, 17, 114-116.

- Patterson, H. D. and Thompson, R. (1974). Maximum likelihood estimation of components of variance, *Proc. Eighth International Biochem. Conf.*, 197-209.
- Perkiomäki, M. (1995) *Track and Field Statistics*. Retrieved from <http://mikap.iki.fi/sport/index.html>.
- Searle, S. R, Casella, G. and McCulloch, C. E. (1992) *Variance Components*, New York: John Wiley and Sons.
- Smyth, G. (2000), "Selling Price of Antique Grandfather Clocks." Retrieved from <http://www.statsci.org/data/general/auction.html>.
- Rasmussen, M. (1998), Activities of Dolphin Groups. University of Southern Denmark, Odense. Retrieved from <http://www.statsci.org/data/general/dolpacti.html>.
- Rawlings, J.O. (1988), *Applied Regression Analysis: A Research Tool*, Pacific Grove, CA: Wadsworth and Books/Cole.
- Sall, J.P. (1990), "Leverage Plots for General Linear Hypotheses," *American Statistician*, 44, (4), 303-315.
- SAS Institute Inc. (1996), *SAS/STAT Software, Changes and Enhancements through Version 6.11, The Mixed Procedure*, Cary, NC: SAS Institute Inc.
- SAS Institute (1986), *SAS/QC User's Guide, Version 5 Edition*, Cary, NC: SAS Institute Inc.
- SAS Institute (1987), *SAS/STAT Guide for Personal Computers, Version 6 Edition*, Cary, NC: SAS Institute Inc.
- SAS Institute (1988), *SAS/ETS User's Guide, Version 6 Edition*, Cary, NC: SAS Institute Inc.
- SAS Institute (1989), *SAS/Technical Report P-188: SAS/QC Software Examples, Version 6 Edition*. Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1989), *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC: SAS Institute Inc., 1165-1168.
- SAS Institute Inc. (2008), *SAS/STAT® 9.2 User's Guide*, Cary, NC: SAS Institute Inc.
- Satterthwaite, F.E., (1946), "An approximate distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110-114.
- Schiffman, A. (1982), *Journal of Counseling and Clinical Psychology*.
- Schuirmann, D.L. (1981), "On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval," *Biometrics* 37, 617.
- Snedecor, G.W. and Cochran, W.G. (1967), *Statistical Methods*, Ames, Iowa: Iowa State University Press.
- Simpson, E.H. (1951), The interpretation of interaction in contingency tables, *JRSSB* 13: 238-241.

- Stichler, R.D., Richey, G.G. and Mandel, J.(1953), "Measurement of Treadwear of Commercial Tires," *Rubber Age*, 73:2.
- Stigler, S.M. (1986), *The History of Statistics*, Cambridge: Belknap Press of Harvard Press.
- Stigler, S. M. (1977), Do Robust Estimators Work with Real Data? *The Annals of Statistics* 5:4, 1075.
- Swift, Jonathan (1735), *Gulliver's Travels*. Quote is from p. 44 of the *Norton Critical Edition*, (1961) Robert A. Greenwood, ed. New York: W.W. Norton & Co.
- Negiz, A. (1994) "Statistical Monitoring and Control of Multivariate Continuous Responses". NIST/SEMATECH e-Handbook of Statistical Methods. Retrieved from <http://www.itl.nist.gov/div898/handbook/pmc/section6/pmc621.htm>.
- Neter, J. and Wasserman, W. (1974), *Applied Linear Statistical Models*, Homewood, IL: Richard D Irwin, Inc.
- Theil and Fiebig, (1984), *Exploiting Continuity*, Cambridge, Mass: Ballinger Publishing Co.
- Third International Mathematics and Science Study*, (1995). US Government: National Center for Educational Statistics and International Education Association.
- Tukey, J. (1953), "A problem of multiple comparisons," Dittoed manuscript of 396 pages, Princeton University.
- Tversky and Gilovich (1989), "The Cold Facts About the Hot Hand in Basketball," *CHANCE*, 2, 16-21.
- Wardrop, Robert (1995), "Simpson's Paradox and the Hot Hand in Basketball", *American Statistician*, Feb 49:1, 24-28.
- Westlake, W.J. (1981), "Response to R.B.L. Kirdwood: bioequivalence testing--a need to rethink", *Biometrics* 37, 589-594.
- Wheeler, D.P. (2006), *EMP III: Using Imperfect Data*, Knoxville, TN., SPC Press and Statistical Process Controls, Inc.
- Winer, B.J., (1962), *Statistical Principles in Experimental Design, Second Edition*, New York: McGraw-Hill Publishing Company.
- Wolfinger, R., Tobias, R., and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM J. Sci. Comput.* 15, 6 (Nov. 1994), 1294-1310.
- Yule, G.U. (1903), "Notes on the theory of association of attributes in statistics," *Biometrika* 2: 121-134.



# Technology License Notices

The ImageMan DLL is used with permission of Data Techniques, Inc.

SAS INSTITUTE INC.'S LICENSORS MAKE NO WARRANTIES, EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, REGARDING THE SOFTWARE. SAS INSTITUTE INC.'S LICENSORS DO NOT WARRANT, GUARANTEE OR MAKE ANY REPRESENTATIONS REGARDING THE USE OR THE RESULTS OF THE USE OF THE SOFTWARE IN TERMS OF ITS CORRECTNESS, ACCURACY, RELIABILITY, CURRENTNESS OR OTHERWISE. THE ENTIRE RISK AS TO THE RESULTS AND PERFORMANCE OF THE SOFTWARE IS ASSUMED BY YOU. THE EXCLUSION OF IMPLIED WARRANTIES IS NOT PERMITTED BY SOME STATES. THE ABOVE EXCLUSION MAY NOT APPLY TO YOU.

IN NO EVENT WILL SAS INSTITUTE INC.'S LICENSORS AND THEIR DIRECTORS, OFFICERS, EMPLOYEES OR AGENTS (COLLECTIVELY SAS INSTITUTE INC.'S LICENSOR) BE LIABLE TO YOU FOR ANY CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES (INCLUDING DAMAGES FOR LOSS OF BUSINESS PROFITS, BUSINESS INTERRUPTION, LOSS OF BUSINESS INFORMATION, AND THE LIKE) ARISING OUT OF THE USE OR INABILITY TO USE THE SOFTWARE EVEN IF SAS INSTITUTE INC.'S LICENSOR'S HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. BECAUSE SOME STATES DO NOT ALLOW THE EXCLUSION OR LIMITATION OF LIABILITY FOR CONSEQUENTIAL OR INCIDENTAL DAMAGES, THE ABOVE LIMITATIONS MAY NOT APPLY TO YOU. SAS INSTITUTE INC.'S LICENSOR'S LIABILITY TO YOU FOR ACTUAL DAMAGES FOR ANY CAUSE WHATSOEVER, AND REGARDLESS OF THE FORM OF THE ACTION (WHETHER IN CONTRACT, TORT (INCLUDING NEGLIGENCE), PRODUCT LIABILITY OR OTHERWISE WILL BE LIMITED TO \$50.00.





# Index

## A

Abrasion.jmp 216  
Abstract Mathematics 100  
Add Multiple Columns Command 33  
Add Rows 112  
Add Statistics Column command 57  
Adjusted RSquare 225  
AdverseR.jmp 216  
Alcohol.jmp 314, 337, 338  
Alias Matrix 440, 455  
Alias Optimal Designs 455  
aliasing 439  
alpha level 106, 148  
alternative hypothesis 105, 147  
Analgesics.jmp 173, 244  
Analysis of Covariance 386, 390  
analysis of variance 182, 188  
Analysis of Variance table 354  
Analyze a Screening Model 433  
Analyze menu 22  
Animals.jmp 32, 407  
ANOVA 182, 217, 235  
ANOVA table 223  
Anscombe.jmp 266  
Assuming unequal variances 185  
Assumptions 102  
attributes charts 549  
Augmentation (DOE) 472  
Automess.jmp 51, 137  
average 141, 144

## B

BabySleep.jmp 205, 210  
Balanced (DOE) 472  
Balanced Data means comparisons 226  
balanced design 221, 432, 471  
Bartlett's test 237, 238  
beta level 106, 148

Big Class.jmp 60, 137  
binomial model 337  
biplot 503  
Birth Death.jmp 128, 165, 278  
bivariate density estimation 482  
bivariate distributions 480  
bivariate outliers 491  
Bivariate platform 208  
Blood Pressure by Time.jmp 205, 210  
Body Measurements.jmp 374  
Bonferroni's adjustment 104  
Box Corrosion Split-Plot Design.jmp 466  
box plots 139, 140, 180  
BP Study.jmp 34  
Braces.jmp 559  
Brown-Forsythe test 237, 238  
brush tool 18  
Building Formulas 90  
built-in script 155  
Business of Statistics 97

## C

C Total 224  
Candy.jmp 301  
Canonical Curvature table 459  
canonical plot 503  
capability analysis 564  
    CP, CPK, CPL and CPU 565  
    goal plot 569  
    indices 565  
    Long Term Sigma 573  
    PP, PPK 569  
    process capability 564  
    Short Term Sigma 573  
Capability.jmp 570  
Car Poll.jmp 306  
Cars.jmp 215  
Cassini.jmp 276

- Cassub.jmp 150
- categorical distributions, multivariate 303
- categorical responses 283
  - and Count Data 283
  - large sample sizes 289
  - Simulated 286
  - standard error of the mean 292
  - Variability in the Estimates 288
- Center points (DOE) 472
- centering and scaling 149
- Central Composite Design (CCD) 457, 463
- Central Limit Theorem 133, 165, 167, 170, 194, 293
- Central Limit Theorem.jmp 170
- Cereal.jmp 174
- ceteris paribus 103
- Chamber.jmp 211, 215
- Cheese Taste.jmp 53
- Cheese.jmp 343
- cherts 502
- cherts.jmp 502
- Children's Popularity.jmp 342
- chi-square 293, 303
  - formula for 315
  - test 281
  - univariate categorical 296
- Cities.jmp 375
- clipboard 46
- Clips1.jmp 562
- Cluster analysis 508
- clusters 484
- Coating.jmp 168
- Cochran-Mantel-Haenszel test 336
- Coding (DOE) 472
- Col Shuffle Command 86
- collinearity 358, 361, 365
  - exact singularity 362
  - example 364
  - longley data 364
- Color Clusters 510
- Color or Mark by Column 118
- columns
  - Selecting and Deselecting 30
- common causes 545
- Companies.jmp 55, 215
- Comparing the normal and Student's t distributions 153
- comparison circles 227, 228, 233
  - interpreting 229
- conditional distributions 136
- Conditional Expressions 72
- confidence interval 106, 120, 145, 146, 161, 183, 223
- fiducial 327
- for parameter estimates 356
- for slope 356
- Confidence.jsl 120
- Confounding 439
- Confounding Structure 439
- Constant Expressions 90
- Context Menu Commands 50
- contingency table 305, 308
- continuous values 21
- contour graph 482
- Contour Grid 461
- Contour Profiler 461
- control charts 112, 545, 548
  - C charts 557
  - common causes 551
  - for attributes 551, 553, 557
  - for variables 550, 552
  - in control 564
  - Levey-Jennings charts 561
  - NP charts 557
  - P charts 557
  - presummarize charts 560
  - stable 551, 564
  - Tailoring the Horizontal Axis 564
  - tests for special causes 552, 564
  - U Charts 557
  - unstable 551
  - UWMA charts 561
- coordinate exchange 440
- copy and paste 46
- Copy command 46
- Copy With Column Names 46
- correlation 487, 488, 489
  - coefficients 489
  - many variables 490
  - multivariate 486
- Correspondence Analysis 303, 318, 319
- Corrsim.jmp 487
- CP, CPK, CPL and CPU 565
- Create Web Report 544
- Creating a New JMP Table 32
- Crime.jmp 174, 515
- crossed effects 403
- crosshair tool 18
- crosstabs 54, 307
- cumulative distribution function 165
- Current Stock Averages.jmp 79
- Cursors
  - Arrow 30
  - Double arrow 31
  - I-beam 31
  - List check 31
  - Pointer 32
  - Selection 31

- curvature analysis 459  
 Custom Designer 427, 440, 452, 463  
 Customize JMP 19  
 Cutting, Dragging, and Pasting Formulas 90
- D**
- Data Mining 103  
 Exploratory Modeling 517  
 Neural Nets 531  
 Recursive Partitioning 519  
 with stepwise regression 369
- Database > Open 45  
 Decathlon.jmp 420  
 Definitive Screening Designs 473  
 degrees of freedom 107, 152, 185, 220, 238  
 delimiters 41  
 demoKernel.jsl 172  
 demoLeastSquares.jsl 248  
 dendrogram 510, 512  
 Denim.jmp 419  
 density contours 483  
 Density Ellipse command 360  
 density estimation 481  
 density function 130, 131  
 Design Evaluation 440  
 Design of Experiments (DOE) 421  
   Classical designs 425  
   Factorial Designs 425  
   factors 424  
   Model reduction 452  
   response 424  
   Response Surface Designs 425  
   Screening Design 425
- Dice Rolls.jmp 111  
 Diet.jmp 159  
 disclosure icons 16, 128  
 discriminant analysis 326, 327, 485, 502  
   biplot 503  
   canonical plot 503  
   discriminant scores 504  
   Stepwise variable selection 507
- distribution  
 Bernoulli 284  
 categorical 129  
 continuous 129  
 double exponential 140  
 exponential 140, 164  
 $F$  189  
 Gaussian 133  
 multinomial 284  
 normal 134  
 Poisson 284  
 probability 130  
 simulation 117
- uniform 163  
 unimodal 130
- Distribution command 128  
 Distribution platform 125, 128, 134, 135, 139  
 DOE > Sample Size and Power 243
- DOE Glossary**
- Augmentation 472
  - Balanced 472
  - Coding 472
  - Definitive Screening Designs 473
  - Effect Precision 473
  - Effect Size 473
  - Folding 473
  - Fractional Factorial 473
  - Optimality 473
  - Orthogonal 474
  - Plackett-Burman 474
  - Resolution 474
  - Response Surface 475
  - Robustness 474
  - Screening Designs 475
  - Space Filling 475
  - Split Plot 475
  - Taguchi 476
- DOE platform 421  
 Dolphins.jmp 343  
 Doped Wafers.jmp 242  
 Drawing Marbles 114  
 Drug.jmp 221, 227, 231, 232, 240, 382, 393  
 dummy variables 381
- E**
- Earth's Ecliptic 150, 276  
 Effect Leverage Plots 356  
 Effect Precision (DOE) 473  
 Effect Screening 434  
 Effect Size (DOE) 473  
 effect sparsity 474  
 Effect Test table 204  
 eigenvalues 459  
 eigenvectors 459  
 EMP (Evaluating the Measurement Process) 574  
 Entering Count Data 312  
 Entering data 35  
 EWMA 563  
 exact nonparametric tests 239  
 Excel Files 47  
 expected value 313  
 Exponentially Weighted Moving Average (EWMA)  
   chart 563  
 extreme values 143

**F**

F 255  
 Faces of Statistics 98  
 Factor Analysis 501  
 Factorial 473  
*F*-distribution 220  
 fiducial confidence intervals 327  
 File > Export 49  
 first principal component 497  
 Fit Line command 251  
 Fit Mean command 251  
 Fit Model launch window 350  
 Fit Polynomial command 260  
 Fit Special command 264  
 Fit Y by X command 181  
 Fit Y by X Platform 182  
 Flipping Coins 114  
*Flrpaste.jmp* 433  
 Folding (DOE) 473  
*Football.jmp* 175  
 Formulation 473  
 Fractional Factorial (DOE) 473  
 F-ratio 255  
 F-ratio 188, 189, 204, 225  
 Frequencies 129  
 frequency counts 130  
 F-statistic 220, 254  
*F*-test 188, 189, 193, 217, 219, 354  
 Full Factorial Design 442

**G**

$G^2$  and  $X^2$   
 testing with 305  
 $G^2$  Likelihood Ratio Chi-Square 294, 296  
 Galton, Francis 245, 269, 272  
*Galton.jmp* 269  
 Gauge R&R 574  
 general linear models 377, 379, 382  
 Effect Tests 391  
 Interaction effects 400  
 Parameters and Means 385  
 prediction equation 389  
 Regressor Construction 384  
 Separate Slopes 396  
 generalized linear models 337  
 Geometric Moving Average (GMA) chart 563  
 GLIM 337  
 GMA 563  
 goal plot 569  
 Goodness of Fit 165  
 Gosset, William 214  
*Gosset's Corn.jmp* 214  
 grabber tool 18

grand mean 188

*Grandfather Clocks.jmp* 373  
 Graph menu 22  
 group means 177, 189, 223  
 grouped *t*-test 196  
*Growth.jmp* 250

**H**

hand tool 18  
 hierarchical clustering 508  
 high-leverage points 391  
 histogram 23, 128, 129, 135, 136, 137  
 Home Window (JMP on Windows) 10  
 Honestly Significant Difference (HSD) 232  
*Hot Dogs.jmp* 175  
 Hotelling-Lawley trace 418  
*Hothand.jmp* 334  
 HSD 232, 233  
 HTML, interactive 49  
*Htwt12.jmp* 179  
*Htwt15.jmp* 192, 213

**I**

Importing Data 39  
 Importing Microsoft Excel Files 44  
 Importing Text Files 41  
 Independence  
 expected values 313  
 testing for 309, 314  
 Independent Groups 177, 179  
 indicator variables 381  
 Individual Measurement charts 552  
 inference 144  
 Initialize Data 86  
 interaction plot 449  
 interactions 379  
 Interactive Teaching Modules 123  
 interquartile range 139, 140, 235  
 inverse prediction 327, 328  
*Iris.jmp* 481, 516  
 iterative proportional fitting 285

**J**

jackknifed distances 496  
 JMP Data Tables 27, 29, 47  
 JMP Main Menu 10  
 JMP Starter 10  
 Join command 52  
 JSON 48  
 Juggling Data Tables 51

**K**

Kendall's tau 24

Kernel Density Estimates 172  
 kernel smoothers 482  
 Kolmogorov-Smirnov test 165, 239  
     KSL 165  
     Lilliefors 165  
 Kruskal-Wallis test 240  
 KSL test 165  
 kurtosis 143

**L**  
 L1 (median least absolute values estimator) 142  
 lack-of-fit test 394, 401  
 lasso tool 18  
 Launch an Analysis Platform 14  
 law of large numbers 113  
 LD50 327, 328  
 Leaf Report 527  
 Least Significant Difference (LSD) 230, 233  
 least squares 141, 247, 345, 347  
     demonstration 248  
     least squares means 393  
 Lenth's PSE 436  
 Levels of Uncertainty 101  
 Levene's test 237, 238  
 leverage plots 355, 361, 364, 390, 391, 392  
     effect 356  
     schematic 392  
 Levi Strauss Run-Up.jmp 243  
 likelihood 294  
 likelihood ratio chi-square 295, 298, 309, 310  
 Likelihood Ratio Tests 295  
 Lilliefors test 165  
 linear dependency 362  
 linear models  
     coding scheme 381  
     kinds of effects 380  
 linear regression  
     models 347  
 Linnerud.jmp 349, 359  
 Linnrand.jmp 369  
 Lipid Data.jmp 521, 541  
 loading plot 500  
 local variables 87, 114  
 log odds-ratio 322  
 logistic regression 284, 285, 321, 324, 502  
     Degrees of Fit 325  
 logit 322  
 log-likelihood 294, 309  
 log-linear models 284  
 Longley.jmp 364  
 LSD (Least Significant Difference) 230  
 LSMeans Plot command 405

**M**  
 MAD (median minimum absolute deviation estimator) 142  
 magnifier tool 18  
 Mahalanobis distance 493  
 Main Effects Only 452  
 Make Alias Optimal Design 455  
 Make into Data Table command 50  
 Make into Matrix 50  
 Mann-Whitney *U*-test 24, 213  
 marginal homogeneity 309  
 Mark Clusters 510, 512  
 matched pairs 177, 196, 200, 209  
 Matched Pairs platform 24, 208, 269  
 maximum likelihood 102, 309  
 maximum likelihood estimator 141, 294  
 Mb-dist.jmp 297  
 Mbtd.jmp 318, 319  
 mean 130, 132, 141, 142, 143  
 mean square 220, 225  
     error 188  
 Means and StdDev command 236  
 Means Comparisons 230  
     unbalanced data 227  
 means diamonds 183, 222, 228, 236  
     overlap marks 226  
 Means/Anova command 222  
 Means/Anova/Pooled t command 23, 183, 185  
 Mechanics of Fit 577  
     Categorical Responses 586  
 median 132, 142  
 Median rank scores 239  
 Median test 211  
 Mesh Plot command 484  
 Method of Moments 418  
 Michelson.jmp 243  
 Minimum-Aberration 473  
 Mixture Design 473  
 Mixtures, Modes, and Clusters 484  
 modeling types 21, 23, 129  
 modes 484  
 moments 129, 130, 143, 146, 151  
 Monte Carlo 109, 289, 291, 292  
 mosaic plots 54, 307, 308  
 movies.jmp 173, 242, 277  
 moving average 79  
     using the summation function 79  
 moving average charts 560  
 Moving data out of JMP 47  
 moving range charts 553  
 multinomial distribution 284  
 multiple comparisons 217  
     adjusting for 232  
 multiple regression 345

Effect Tests 356  
example 348  
Hidden Leverage Point 366  
predicted values 351  
prediction formula 352  
stepwise 369

**N**

Navigating Platforms 22  
nested effects 379, 406, 407, 408  
Neural Nets  
hidden nodes 532  
Overfitting 534  
Validation 533  
Neural platform 531  
New Column Command 36  
nominal values 21  
non-central t-distribution 159  
nonparametric methods 217, 239  
nonparametric multiple comparison 240  
nonparametric statistics 24  
Nonparametric tests 155, 211  
Nonparametric-Wilcoxon command 213  
normal density 136, 162  
normal distribution 85, 133, 143, 145, 149, 164  
elliptical contours 485  
Normal Plot 435  
normal plot 436  
normal quantile plot 162, 164, 165, 167, 194, 236  
and normality 194  
Normal Quantile Plot command 195  
normality 166  
null hypothesis 105, 147, 166  
Number of Center Points 472  
Number of Replicate Runs 472

**O**

O'Brien's test 237, 238  
one-way layout 219  
On-Time Arrivals.jmp 214  
Open command 39  
Open Database 45  
Opening a JMP Data Table 12  
Optimality (DOE) 473  
Ordinal Logistic Regression 331  
ordinal values 21  
Orthogonal (DOE) 474  
Outlier Analysis command 493  
outliers 140, 142, 182, 198, 492, 496  
bivariate 491

**P**

paired *t*-test 196, 200, 203, 207

interpretation rule 203  
Partition Analysis  
Candidates 524  
Crossvalidation 528  
growing trees 521  
LogWorth 523  
Partition Platform 519  
Paste command 46  
Paste With Column Names 46  
Peanuts.jmp 532, 542  
Pearson chi-square 293, 298, 310  
Pendulum.jmp 91  
Pillai's trace 418  
Plackett-Burman (DOE) 474  
Plot Residuals command 258, 262  
Plotting Data 37  
point estimate 144  
Polycity.jmp 275  
Polynomial Models 260, 261  
Pitfalls 275  
polytomous responses 330  
pooled *t*-test 185  
Popcorn.jmp 400  
power 106, 148, 157  
PowerPoint 49  
practical difference 167  
Prediction Profiler 437, 438, 447, 450, 460  
Desirability 450  
Maximize Desirability 450  
Response Goal 451  
Preferences 19  
Presidential Elections.jmp 278  
Pressure Cylinders 587  
principal components 497, 499, 515  
Probability and Randomness 102  
probability distribution 102, 147  
process capability 564  
Process Screening 557  
Product function 78  
proof by contradiction 147  
pure error 395, 396  
*p*-value 106, 148, 151, 156  
Animation 155

**Q**

quantile box plot 146, 147, 235  
Quantile function 78  
quantiles 129, 130, 132, 140, 142, 145, 235  
quartiles 139  
Question Mark tool 19

**R**

raking 285

Randdist.jmp 134, 135  
 Random 86  
 random effects 406, 409  
     Correlated Measurements-Multivariate Model 416  
     Mixed Model 409  
     Reduction to the Experimental Unit 414  
     Varieties of Analysis 418  
 Random Normal function 85  
 Random Number Functions 82  
 Random Uniform function 83, 111, 291  
 Randomize 472  
 rank ordering 211  
 rank scores 239  
*r*-charts 552  
 Reactor 20 Custom.jmp 444  
 Reactor 32 Runs.jmp 452  
 regression 245, 247, 269, 270  
     clusters of points 275  
     Confidence Intervals on the Estimates 256  
     line 247  
     Parameter Estimates 255  
     properties 247  
     residuals 258, 262  
     statistical tables 252  
     switching variables 272  
     Testing the Slope 250, 252  
     three-dimensional view 257  
     to the mean 245  
     Why It's Called Regression 269  
 regression definitions  
     coefficients, parameters 347  
     error, residual 348  
     intercept term 348  
     regressors, X's 347  
     response, Y 347  
 REML report 469  
 repeated measures 406, 409, 410, 416  
 Replicates (DOE) 472  
 resampling methods 104  
 residual error 355  
 residual plot 351, 367  
 residual variance 188  
 residuals 167, 188, 219, 247, 258, 261, 262, 351, 352, 367  
 resizing graphs 18  
 Resolution (DOE) 474  
 response categories 283  
 response probabilities 305  
 response surface designs (DOE) 452, 456, 475  
 Response Surface report 459  
 Ro.jmp 366  
 Robustness (DOE) 474  
 robustness of the median 142

Rolling Dice 111  
 root mean square error 188, 225  
 rows  
     adding 34  
     excluding 258  
     Highlighting 15  
     Selecting and Deselecting 30  
 Roy's maximum root 418  
 RSquare 225  
     Adjusted 225

**S**

S charts 552  
 sample mean 132, 144  
 Sample Mean versus True Mean 107  
 sample median 132  
 sample quantile 132  
 sample size 160  
 Sample Size and Power 190  
 sample variance 132, 141  
 Sampling Candy 114  
 SAS Transport Files 48  
 SAS V7 Data Set 48  
 saturated design 435  
 saturated model 394  
 Save As command 47  
 Save Residuals command 262  
 Save Selection As Command 49  
 scatterplot 269  
 Scatterplot 3D 494  
 scatterplots 24  
 Score Summaries report 506  
 Scores.jmp 173, 244  
 scree plot 510  
 Screening Designs (DOE) 475  
     Examples 452  
 Screening for Interactions 442  
 second principal component 497  
 Select Misclassified Rows 505  
 Selecting Expressions 91  
 selection tool 19  
 Shapiro-Wilk W test 165, 212  
 Shewhart control chart 545  
 Shewhart, Walter 545  
 shortest half 140  
 signed-rank test 211  
 Significance 475  
 significance level 106, 148  
 significant difference 159  
 Simprob.jmp 292  
 Simpson's paradox 334, 336  
 SimulatedClusters.jmp 509  
 Singularity Details report 363  
 skewness 143

- Sleeping Animals.jmp 374  
 Slope.jmp 275  
 smooth curve 481  
 Socioeconomic.jmp 515  
 Solubility.jmp 279, 490  
 Sort by Column command 50  
 Space Filling (DOE) 475  
 sparse table 306  
 Spearman's correlation 24  
 special causes 545  
 Special Tools 18  
 Specifying Response Surface Effects Manually 462  
 Spline Fit 265  
 Split command 416  
 Split Plot Designs  
     whole plots 464  
 Split Plot designs (DOE) 475  
     easy to change factors 465  
     hard to change factors 465  
     subplot, split plot 465  
     whole plots 465  
 split plots 410  
 spread 132  
 Spring.jmp 321, 326, 331  
 Springs for Continuous Responses 579  
 Stack command 52, 53  
 stacked data 203  
 standard deviation 107, 130, 141, 143, 144, 163  
 standard error 107, 184, 223, 289  
 standard error of the mean 132, 144  
 statistical inference 144, 177  
 Statistical Process Control 545  
 Statistical Quality Control 545  
 Statistical Significance 105, 160  
 Statistical Terms 104  
 statistical thinking 95  
 statistics 95, 131, 132, 134, 545  
     Biased, Unbiased 107  
 stem-and-leaf plot 137  
 stepwise regression 369, 370  
 Student's *t*-distribution 145, 152  
 Student's *t*-statistic 152  
 Student's *t*-test 152, 184, 199  
     one-sided 187  
 Subgroup button 58  
 Subset command 51  
 Sum function 78  
 sum of squares 190  
 Summarize Down Columns or Across Rows 77  
 Summary command 55  
 Summary of Fit table 223  
 Summary Statistics 55  
 Summation Function 78  
 sum-to-zero coding 382  
 Surface Effects 461  
 Surface Plot 540  
 Surface Plot command 541  
 Surface Profiler command 540  
 Survey Data 306
- T**
- Table Styles 50  
 Table Variables 87  
 Taguchi (DOE) 476  
 Teeth.jmp 511  
 Test Mean command 23, 199, 200  
 testing for normality 166  
 Testing Hypotheses 147  
 Text Export Files 47  
 Therm.jmp 197  
 time series 545  
 Tip of the Day 9  
 Titanic.jmp 343, 420  
 TOST method 168  
 Transformed Fits 263  
 Triangle Probability.jsl 115  
 true mean 107  
*t*-statistic 154  
*t*-test 23, 160, 168, 185  
     grouped 196  
     paired 196  
 Tukey-Kramer Honestly Significant Difference 232  
 Two-Sided versus One-Sided 105  
 Two-Tailed versus One-Tailed 105  
 Two-way Analysis of Variance 400  
 two-way tables 303, 312  
     entering 314  
 Type I error 148, 157, 232  
 Type II error 148, 157, 159  
 Typing.jmp 30
- U**
- U charts 559  
 unbiased estimator 107  
 unequal variance  
     testing means 238  
     tests 238  
 UnEqual Variance command 237  
 Uniformly Weighted Moving Average (UWMA)  
     Charts 561  
 UWMA 562
- V**
- validate experimental results 452  
 validating statistical assumptions 177  
 Van der Waerden rank scores 239  
 van der Waerden rank scores 211

van der Waerden score 162  
variability 132  
variables charts 549  
variance 107, 131, 132, 141, 143  
variance components 469  
Variance Components Analysis 574

**W**

*W* statistic 165  
Washers.jmp 558  
Western Electric Rules 552  
Westgard Rules 552  
whiskers 139  
whole-model test 390  
Wilcoxon rank scores 211, 239  
Wilcoxon Rank Sum test 24, 213, 214, 239  
Wilcoxon signed-rank test 155, 200, 211  
Wilk's lambda 418  
Working with Graphs and Reports 48  
World Demographics.jmp 516

**X**

XBar charts 552

**Z**

*z*-statistic 149  
*z*-test 149–151



# **Do You Like What You Just Read?**

To learn about our authors and their books, download free sample chapters, access example code and data, and more, go to the SAS Press Author pages.

To find additional books that are just right for you, browse the full SAS Publishing catalog.

Email us: [sasbooks@sas.com](mailto:sasbooks@sas.com)

Call: 1-800-727-3228





